LETTER **Communicated by Nakul Verma**

Safe Triplet Screening for Distance Metric Learning

Tomoki Yoshida
yoshida.t.mllab.nit@gmail.com
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555, Japan

Ichiro Takeuchi
takeuchi.ichiro@nitech.ac.jp
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555, Japan; National Institute for Materials Science, Sengen, Tsukuba, Ibaraki, 305-0047, Japan; and RIKEN Center for Advanced Intelligence Project, Nihonbashi, Chuo-ku, Tokyo, 103-0012, Japan

Masayuki Karasuyama
karasuyama@nitech.ac.jp
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555, Japan; National Institute for Materials Science, Sengen, Tsukuba, Ibaraki, 305-0047, Japan; and Japan Science and Technology Agency, Honcho, Kawaguchi-shi, Saitama, 332-0012, Japan

Distance metric learning has been widely used to obtain the optimal distance function based on the given training data. We focus on a triplet-based loss function, which imposes a penalty such that a pair of instances in the same class is closer than a pair in different classes. However, the number of possible triplets can be quite large even for a small data set, and this considerably increases the computational cost for metric optimization. In this letter, we propose safe triplet screening that identifies triplets that can be safely removed from the optimization problem without losing the optimality. In comparison with existing safe screening studies, triplet screening is particularly significant because of the huge number of possible triplets and the semidefinite constraint in the optimization problem. We demonstrate and verify the effectiveness of our screening rules by using several benchmark data sets.

1 Introduction

Using an appropriate distance function is essential for various machine learning tasks. For example, the performance of a k-nearest neighbor (k-NN) classifier, one of the most standard classification methods, depends

Neural Computation **31, 2432–2491** (2019)
https://doi.org/10.1162/neco_a_01240

crucially on the distance between different input instances. The simple Euclidean distance is usually employed, but it is not necessarily optimal for a given data set and task. Thus, the adaptive optimization of the distance metric based on supervised information is expected to improve the performance of machine learning methods including k-NN.

Distance metric learning (Weinberger & Saul, 2009; Schultz & Joachims, 2004; Davis, Kulis, Jain, Sra, & Dhillon, 2007; Kulis, 2013) is a widely accepted technique for acquiring the optimal metric from observed data. The 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 semidefinite matrix. This approach has been applied to tasks such as classification (Weinberger & Saul, 2009), clustering (Xing, Jordan, Russell, & Ng, 2003), and ranking (McFee & Lanckriet, 2010). These studies show that the optimized distance metric improves the prediction performance of each task. Metric optimization has also attracted wide interest, even from researchers engaged in recent deep network studies (Schroff, Kalenichenko, & Philbin, 2015; Hoffer & Ailon, 2015).

The seminal work of distance metric learning (Weinberger & Saul, 2009) presents a triplet-based formulation. A triplet (i, j, l) is defined by the pair $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which have the same label (same class), and $\boldsymbol{x}_l$, which has a different label (different class). For a triplet (i, j, l), the 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 pair in the same class is closer than the pair in different classes. For each of the triplets, Weinberger and Saul (2009) define a loss function that penalizes violations of this constraint,

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

where $\mathcal{T}$ is a set of triplets and $\ell : \mathbb{R} \to \mathbb{R}$ is some loss function (e.g., the standard hinge loss function). In addition to the triplet loss, other approaches, such as pairwise- and virtual point–based loss functions have been studied. In the pairwise approach, the number of pairs can be much smaller than the triplets; Davis et al. (2007) used only $20c^2$ pairs, where c is the number of classes. The virtual point approach (Perrot & Habrard, 2015) converts the metric learning problem into a least squares problem, which minimizes a loss function for n virtual points. We particularly focus on the triplet approach because the relative evaluation $d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) < d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_l)$ would be more appropriate for many metric learning applications such as nearest-neighbor classification (Weinberger & Saul, 2009), and similarity search (Jain, Kulis, Dhillon, & Grauman, 2009), in which relative comparison among objects plays an essential role. In fact, a recent comprehensive

survey (Li & Tian, 2018) showed that many current state-of-the-art methods are based on triplet loss, which they referred to as relative loss. Note that although the quadruplets approach (Law, Thome, & Cord, 2013) can also incorporate higher-order relations, we mainly focus on the triplet approach because it is much more popular in the community, although our framework can also accommodate the quadruplet case, as we explain in section 6.3.

However, the set of triplets $\mathcal{T}$ is quite large even for a small data set. For example, in a two-class problem with 100 instances in each class, the number of possible triplets is 1,980,000. Because processing a huge number of triplets is computationally prohibitive, a small subset is often used in practice (Weinberger & Saul, 2009; Shi, Bellet, & Sha, 2014; Capitaine, 2016). Typically, a subset of triplets is selected by using the neighbors of each training instance. For n training instances, Shi et al. (2014) selected only $30n$ triplets, and Weinberger and Saul (2009) selected at most $O(kn^2)$ triplets, where k is a prespecified constant. However, the effect on the final accuracy of these heuristic selections is difficult to know beforehand. Jain, Mason, and Nowak (2017) theoretically analyzed a probabilistic generalization error bound for a random subsampling strategy of triplets. Their analysis revealed the sample complexity of metric learning, but the tightness of the bound is not clear and they did not demonstrate the practical use of determining the required number of triplets. For ordinal data embedding, Jamieson and Nowak (2011) showed a lower bound of required triplets $\Omega(dn \log n)$ to determine the embedding, but the tightness of this bound is also not known. Further, the applicability of the analysis to metric learning was not clarified.

Our safe triplet screening enables the identification of triplets that can be safely removed from the optimization problem without losing the optimality of the resulting metric. This means that our approach can accelerate the optimization of time-consuming metric learning with the guarantee of optimality. Figure 1 shows a schematic illustration of safe triplet screening.

Our approach is inspired by the safe feature screening of Lasso (Ghaoui, Viallon, & Rabbani, 2010), in which unnecessary features are identified by the following procedure:

Step 1: Construct a bounded region in which the optimal dual solution is guaranteed to exist.
Step 2: Given the bound created by step 1, remove features that cannot be selected by Lasso.

This procedure is useful to mitigate the optimization difficulty of Lasso for high-dimensional problems; thus, many papers propose a variety of approaches to create bounded regions for obtaining a tighter bound that increases screening performance (Wang, Zhou, Wonka, & Ye, 2013; Liu,

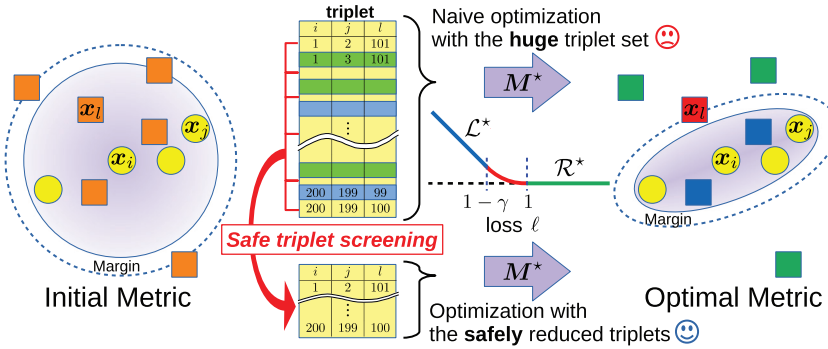


Figure 1: Metric learning with safe triplet screening. The naive optimization needs to minimize the sum of the loss function values for a huge number of triplets (i, j, l). Safe triplet screening identifies a subset of $\mathcal{L}^\star$ (blue points in the illustration on the right) and $\mathcal{R}^\star$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

Zhao, Wang, & Ye, 2014; Fercoq, Gramfort, & Salmon, 2015; Xiang, Wang, & Ramadge, 2017). As another direction of research, the screening idea was applied to other learning methods, including support vector machine non-support vector screening (Ogawa, Suzuki, & Takeuchi, 2013), nuclear norm regularization subspace screening (Zhou & Zhao, 2015), and group Lasso group screening (Ndiaye, Fercoq, Gramfort, & Salmon, 2016).

Based on the safe feature screening techniques, we build the procedure of our safe triplet screening as follows:

Step 1: Construct a bounded region in which the optimal solution $\boldsymbol{M}^\star$ is guaranteed to exist.
Step 2: For each triplet $(i, j, l) \in \mathcal{T}$, verify the possible loss function value under the condition created by step 1.

We show that as a result of step 2, we can reduce the size of the metric learning optimization problem, by which the computational cost of the optimization can be drastically reduced. Although a variety of extensions of safe screening have been studied in the machine learning community (Lee & Xing, 2014; Wang, Wonka, & Ye, 2014; Zimmert, de Witt, Kerg, & Kloft, 2015; Zhang et al., 2016; Ogawa et al., 2013; Okumura, Suzuki, & Takeuchi, 2015; Shibagaki, Karasuyama, Hatano, & Takeuchi, 2016; Shibagaki, Suzuki, Karasuyama, & Takeuchi, 2015; Nakagawa, Suzumura, Karasuyama, Tsuda, & Takeuchi, 2016; Takada, Hanada, Yamada, Sakuma, & Takeuchi, 2016; Hanada, Shibagaki, Sakuma, & Takeuchi, 2018), to the best

of our knowledge, no studies have considered screening for metric learning. Compared with existing studies, our safe triplet screening is particularly significant due to the huge number of possible triplets and the semidefinite constraint. Our technical contributions are summarized as follows:

- We derive six spherical regions in which the optimal $\boldsymbol{M}^\star$ must lie and analyze their relationships.
- We derive three types of screening rules, each of which employs a different approach to the semidefinite constraint.
- We derive efficient rule evaluation for a special case when $\boldsymbol{M}$ is a diagonal matrix.
- We build an extension for the regularization path calculation.

We further demonstrate the effectiveness of our approach based on several benchmark data sets with a huge number of triplets.

This letter is organized as follows. In section 2, we define the optimization problem of large-margin metric learning. In section 3, we first derive six bounds containing optimal $\boldsymbol{M}^\star$ for the subsequent screening procedure. Section 4 derives the rules and constructs our safe triplet screening. The computational cost for the rule evaluation is analyzed in section 5. Extensions are discussed in section 6, in which an algorithm specifically designed for the regularization path calculation, and a special case, in which $\boldsymbol{M}$ is a diagonal matrix, are considered. In section 7, we present the evaluation of our approach through numerical experiments. Section 8 concludes.

1.1 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 semidefinite matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ is denoted by $\boldsymbol{M} \succeq \boldsymbol{O}$ or $\boldsymbol{M} \in \mathbb{R}_+^{d \times d}$. By using the 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 a projection of $\boldsymbol{M}$ onto the semidefinite cone—$\boldsymbol{M}_+ = \mathrm{argmin}_{\boldsymbol{A} \succeq \boldsymbol{O}} \|\boldsymbol{A} - \boldsymbol{M}\|_F^2$.

2 Preliminary

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,

$$d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) := \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j)}, \tag{2.1}$$

where $\boldsymbol{M} \in \mathbb{R}_+^{d \times d}$ is a positive semidefinite matrix that parameterizes distance. As a general form of the metric learning problem, we consider a regularized triplet loss minimization (RTLM) problem. Our formulation is mainly based on a model originally proposed by Weinberger and Saul (2009), which is reduced to a convex optimization problem with the semidefinite constraint. 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 letter.

2.1 Triplet-Based Loss Function. We define a triplet of instances as

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

where $\mathcal{S} = \left\{(i, j) \mid y_i = y_j, i \neq j, (i, j) \in [n] \times [n]\right\}$. 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 class that differs from that of i and j. We refer to the following loss as the 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. By substituting equation 2.1 into the triplet loss, this can be written as

$$\ell\left(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl} \rangle\right),$$

where $\boldsymbol{H}_{ijl} := (\boldsymbol{x}_i - \boldsymbol{x}_l)(\boldsymbol{x}_i - \boldsymbol{x}_l)^\top - (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$. For the triplet loss, we consider the hinge function,

$$\ell(x) = \max\{0, 1 - x\}, \tag{2.2}$$

or the smoothed hinge function,

$$\ell(x) = \begin{cases} 0, & x > 1, \\ \frac{1}{2\gamma}(1 - x)^2, & 1 - \gamma \leq x \leq 1, \\ 1 - x - \frac{\gamma}{2}, & x < 1 - \gamma, \end{cases} \tag{2.3}$$

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 imposes 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 different classes. Both of the two loss functions contain a region in which no penalty is imposed. We refer to this as the zero region.

The two loss functions also contain a region in which the penalty increases linearly, which we refer to as the linear region.

2.2 Primal and Dual Formulation of Triplet-Based Distance Metric Learning. Using the standard squared regularization, we consider the following RTLM as a general form of metric learning:

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

where $\sum_{ijl}$ denotes $\sum_{(i,j,l)\in\mathcal{T}}$, and $\lambda > 0$ is a regularization parameter. In section 6.3, we discuss the relation of RTLM to existing metric learning methods.

The dual problem is written as

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

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

$$\boldsymbol{M}_\lambda(\boldsymbol{\alpha},\boldsymbol{\Gamma}) := \frac{1}{\lambda}\left[\sum_{ijl}\alpha_{ijl}\boldsymbol{H}_{ijl} + \boldsymbol{\Gamma}\right]. \tag{2.4}$$

A derivation of this dual problem is presented in appendix A. Because the last term $\max_{\boldsymbol{\Gamma}\succeq\boldsymbol{O}} -\frac{1}{2}\left\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha},\boldsymbol{\Gamma})\right\|_F^2$ is equivalent to the projection onto a semidefinite cone (Boyd & Xiao, 2005; Malick, 2004), the above problem, Dual1, can be simplified as

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

where

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

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

$$\mathcal{R}^\star := \{(i,j,l)\in\mathcal{T} \mid \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}^\star\rangle > 1\},$$

$$\mathcal{C}^\star := \{(i, j, l) \in \mathcal{T} \mid 1 - \gamma \leq \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}^\star \rangle \leq 1\},$$
$$\mathcal{L}^\star := \{(i, j, l) \in \mathcal{T} \mid \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}^\star \rangle < 1 - \gamma\}. \tag{2.5}$$

This indicates that the triplets in $\mathcal{R}^\star$ and those in $\mathcal{L}^\star$ are the zero region and linear region of the loss function, respectively. The well-known KKT conditions provide the following relation between the optimal dual variable and the derivative of the loss function (see appendix A for details):

$$\alpha^\star_{ijl} = -\nabla \ell(\langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle). \tag{2.6}$$

In the case of hinge loss, the derivative is written as

$$\nabla \ell(\langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle) = \begin{cases} 0, & \langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle > 1, \\ -c, & \langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle = 1, \\ -1, & \langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle < 1, \end{cases}$$

where $\forall c \in [0, 1]$. In the case of smoothed hinge loss, the derivative is

$$\nabla \ell(\langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle) = \begin{cases} 0, & \langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle > 1, \\ -\frac{1}{\gamma}(1 - \langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle), & 1 - \gamma \leq \langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle \leq 1, \\ -1, & \langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle < 1 - \gamma. \end{cases}$$

Both cases can be represented as

$$\nabla \ell(\langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl} \rangle) \begin{cases} = 0, & (i, j, l) \in \mathcal{R}^\star, \\ \in -[0, 1], & (i, j, l) \in \mathcal{C}^\star, \\ = -1, & (i, j, l) \in \mathcal{L}^\star. \end{cases} \tag{2.7}$$

From equations 2.7 and 2.6, we obtain the following rules for the optimal dual variable:

$$(i, j, l) \in \mathcal{R}^\star \Rightarrow \alpha^\star_{ijl} = 0,$$
$$(i, j, l) \in \mathcal{C}^\star \Rightarrow \alpha^\star_{ijl} \in [0, 1],$$
$$(i, j, l) \in \mathcal{L}^\star \Rightarrow \alpha^\star_{ijl} = 1. \tag{2.8}$$

The nonlinear semidefinite programming problem of RTLM can be solved by gradient methods including the primal-based (Weinberger & Saul, 2009) and dual-based approaches (Shen, Kim, Liu, Wang, & Van Den Hengel, 2014). However, the amount of computation may be prohibitive because of the large number of triplets. The naive calculation of the objective

function requires $O(d^2|\mathcal{T}|)$ computations for both the primal and the dual cases.

2.3 Reduced-Size Optimization Problem. Assuming that we have a subset of triplets $(i, j, l) \in \mathcal{L}^\star \cup \mathcal{R}^\star$ before solving the optimization problem. Let $\hat{\mathcal{L}} \subseteq \mathcal{L}^\star$ and $\hat{\mathcal{R}} \subseteq \mathcal{R}^\star$ be the subsets of $\mathcal{L}^\star$ and $\mathcal{R}^\star$ we identify. Then, based on this prior knowledge, the optimization problem, Primal, can be transformed into the following reduced-size problem:

$$\tilde{P}_\lambda(\boldsymbol{M}) = \sum_{(i,j,l)\in\tilde{\mathcal{T}}} \ell(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle) + \frac{\lambda}{2}\|\boldsymbol{M}\|_F^2 + \left(1 - \frac{\gamma}{2}\right)|\hat{\mathcal{L}}| - \langle \boldsymbol{M}, \sum_{(i,j,l)\in\hat{\mathcal{L}}} \boldsymbol{H}_{ijl}\rangle, \tag{2.9}$$

where $\tilde{\mathcal{T}} := \mathcal{T} - \hat{\mathcal{L}} - \hat{\mathcal{R}}$. This problem differs from the original, Primal, as follows

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

The dual problem of this reduced-size problem can be written as

$$\begin{aligned} &\min_{\mathbf{0}\le\boldsymbol{\alpha}\le\mathbf{1}} \tilde{D}_\lambda(\boldsymbol{\alpha}) := -\frac{\gamma}{2}\|\boldsymbol{\alpha}\|_2^2 + \boldsymbol{\alpha}^\top\mathbf{1} - \frac{\lambda}{2}\left\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha})\right\|_F^2, \\ &\text{s.t. } \boldsymbol{\alpha}_{\hat{\mathcal{L}}} = \mathbf{1},\ \boldsymbol{\alpha}_{\hat{\mathcal{R}}} = \mathbf{0}. \end{aligned} \tag{2.10}$$

which is the same optimization problem as Dual2 except that $\boldsymbol{\alpha}_{\hat{\mathcal{L}}}$ and $\boldsymbol{\alpha}_{\hat{\mathcal{R}}}$ are fixed. Because of this constraint, the number of free variables in this dual problem is $|\tilde{\mathcal{T}}|$. An important property of a reduced-size problem is that it retains the same optimal solution as the original problem:

Lemma 1. *The primal-dual problem pair, equations 2.9 and 2.10, and the original problem pair, Primal and Dual2, have the same optimal primal and dual solutions.*

The proof of this lemma is shown in appendix B, along with the derivation of the reduced-size dual, equation 2.10. Therefore, if a large number of $\hat{\mathcal{L}}$ and $\hat{\mathcal{R}}$ could be detected beforehand (i.e., $|\tilde{\mathcal{T}}| \ll |\mathcal{T}|$), the metric learning optimization would be accelerated dramatically.

3 Spherical Bound

As we will see, our safe triplet screening is derived by using a spherical region that contains the optimal $\boldsymbol{M}^\star$. In this section, we show that six variants

of the regions are created by three types of different approaches. Note that the proofs for all the theorems appear in the appendixes.

3.1 Gradient Bound. We first introduce a hypersphere, which we name gradient bound (GB), because the center and radius of the hypersphere are represented by the subgradient of the objective function:

Theorem 1 *(GB). Given any feasible solution $\boldsymbol{M} \succeq \boldsymbol{O}$, the optimal solution $\boldsymbol{M}^\star$ for λ exists in the following hypersphere:*

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

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

The proof is in appendix C. This theorem is an extension of the sphere for SVM (Shibagaki et al., 2015), which can be treated as a simple unconstrained problem.

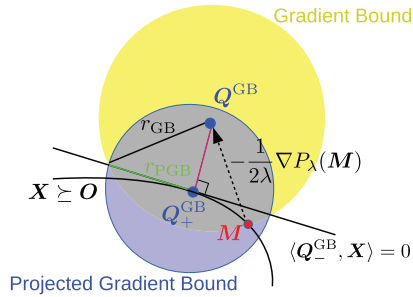
3.2 Projected Gradient Bound. Even when we substitute the optimal $\boldsymbol{M}^\star$ into the reference solution $\boldsymbol{M}$, the radius of the GB is not guaranteed to be 0. By projecting the center of GB onto the feasible region (i.e., a semidefinite cone), another GB-based hypersphere can be derived, which has a radius converging to 0 at the optimal. We refer to this extension as projected gradient bound (PGB); a schematic illustration is shown as Figure 2a. In Figure 2a, the center of the GB $\boldsymbol{Q}^{\mathrm{GB}}$ (the abbreviation of $\boldsymbol{Q}^{\mathrm{GB}}(\boldsymbol{M})$) is projected onto the semidefinite cone, which becomes the center of PGB $\boldsymbol{Q}_+^{\mathrm{GB}}$. The sphere of PGB can be written as

Theorem 2 *(PGB). Given any feasible solution $\boldsymbol{M} \succeq \boldsymbol{O}$, the optimal solution $\boldsymbol{M}^\star$ for λ exists in the following hypersphere:*

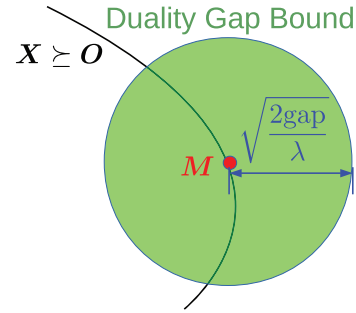
$$\left\|\boldsymbol{M}^\star - \left[\boldsymbol{Q}^{\mathrm{GB}}(\boldsymbol{M})\right]_+\right\|_F^2 \leq \left(\frac{1}{2\lambda}\left\|\nabla P_\lambda(\boldsymbol{M})\right\|_F\right)^2 - \left\|\left[\boldsymbol{Q}^{\mathrm{GB}}(\boldsymbol{M})\right]_-\right\|_F^2.$$

The proof is in appendix D. PGB contains the projections onto the positive and the negative semidefinite 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, only needs to be performed once to evaluate the screening rules of all the triplets. In the standard optimization procedures of RTLM, including Weinberger and Saul (2009), the eigenvalue decomposition of the $d \times d$ matrix is calculated in every iterative cycle, and thus, the computational complexity is not increased by PGB.

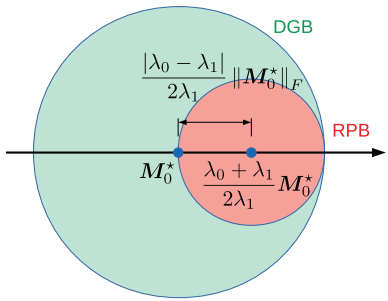
The following theorem shows a superior convergence property of PGB compared to GB:



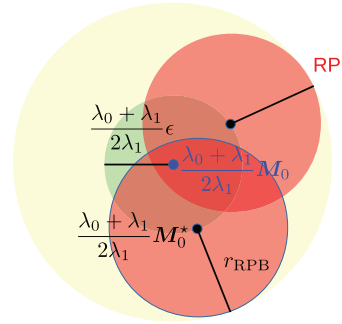
(a) Projected Gradient Bound (PGB). The center of GB Q^{GB} is projected onto the semidefinite cone. The radius is calculated as $r_{\mathrm{PGB}}^2 = r_{\mathrm{GB}}^2 - \|Q^{\mathrm{GB}} - Q_+^{\mathrm{GB}}\|_F^2$.



(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 8).



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

Figure 2: Illustrations of spherical bounds.

Theorem 3. *There exists a subgradient $\nabla P_\lambda(M^\star)$ 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 that set the radius equal to 0. This theorem is an immediate consequence of the proof in appendix I, which is the proof for the relation between PGB and the other bound derived in section 3.4.

From Figure 2a, we see that the half space $\langle -Q_-^{\mathrm{GB}}, X\rangle \geq 0$, where $Q_-^{\mathrm{GB}} = Q^{\mathrm{GB}} - Q_+^{\mathrm{GB}}$, can be used as a linear relaxation of the semidefinite constraint for the linear constraint rule in section 4.3. Interestingly, the GB with this linear constraint is tighter than the PGB. This is proved in appendix D, which gives the proof of the PGB.

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

Theorem 4 *(DGB). Let M be a feasible solution of the primal problem and α and Γ be feasible solutions of the dual problem. Then the optimal solution of the primal problem $M^\star$ exists in the following hypersphere:*

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

The proof is in appendix E. Because the radius is proportional to the square root of the duality gap, DGB obviously converges to 0 at the optimal solution (see Figure 2b). The DGB, unlike the previous bounds, requires a dual feasible solution. This means that when a primal-based optimization algorithm is employed, we need to create a dual feasible solution from the primal feasible solution. A simple way to create a dual feasible solution is to substitute the current M into $M^\star$ of equation 2.6. When a dual-based optimization algorithm is employed, a primal feasible solution can be created by equation 2.4.

For the DGB, we can derive a tighter bound, the constrained duality gap bound (CDGB), with an additional constraint. However, except for a special case (dynamic screening with a dual solver), additional transformation of the reference solution is necessary, which can deteriorate the duality gap. See appendix F for further details.

3.4 Regularization Path Bound. In Wang et al. (2014), a hypersphere is proposed specifically for the regularization path, in which the optimization problem should be solved for a sequence of λs. Suppose that λ_0 has already been optimized and it is necessary to optimize λ_1. Then the same approach as Wang et al. (2014) is applicable to our RTLM, which derives a bound depending on the optimal solution for λ_0 as a reference solution:

Theorem 5 *(RPB). Let $M_0^\star$ be the optimal solution for λ_0. Then the optimal solution $M_1^\star$ for λ_1 exists in the following hypersphere:*

$$\left\|M_1^\star - \frac{\lambda_0 + \lambda_1}{2\lambda_1} M_0^\star\right\|_F^2 \leq \left(\frac{\lambda_0 - \lambda_1}{2\lambda_1}\left\|M_0^\star\right\|_F\right)^2.$$

The proof is in appendix G. We refer to this bound as the regularization path bound (RPB).

The RPB requires the theoretically optimal solution $\boldsymbol{M}_0^\star$, which is numerically impossible. Furthermore, because the reference solution is fixed on $\boldsymbol{M}_0^\star$, the 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.

3.5 Relaxed Regularization Path Bound. To use the RPB in practice, we modify this bound in such a way that the approximate solution can be used as a reference solution. Assume that $\boldsymbol{M}_0$ should satisfy

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

where $\epsilon \geq 0$ is a constant. Given $\boldsymbol{M}_0$, which satisfies the above condition, we obtain the relaxed regularization path bound (RRPB):

Theorem 6 *(RRPB). Let $\boldsymbol{M}_0$ be an approximate solution for λ_0, which satisfies $\|\boldsymbol{M}_0^\star - \boldsymbol{M}_0\|_F \leq \epsilon$. The optimal solution $\boldsymbol{M}_1^\star$ for λ_1 exists in the following hypersphere:*

$$\left\| \boldsymbol{M}_1^\star - \frac{\lambda_0 + \lambda_1}{2\lambda_1} \boldsymbol{M}_0 \right\|_F^2 \leq \left(\frac{|\lambda_0 - \lambda_1|}{2\lambda_1} \|\boldsymbol{M}_0\|_F + \frac{|\lambda_0 - \lambda_1| + \lambda_0 + \lambda_1}{2\lambda_1} \epsilon \right)^2. \tag{3.1}$$

The proof is in appendix H. The intuition behind the RRPB is shown in Figure 2d, in which the approximation error for the center of the RPB is depicted. In the theorem, the RRPB also considers the error in the radius, although 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 4 (DGB) as follows:

$$\epsilon = \sqrt{2(P_{\lambda_0}(\boldsymbol{M}_0) - D_{\lambda_0}(\boldsymbol{\alpha}_0, \boldsymbol{\Gamma}_0))/\lambda_0}. \tag{3.2}$$

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 the 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 $\lambda_1 = \lambda_0$, the RRPB can be applied to perform the screening of λ_1 using any approximate solution having $\|\boldsymbol{M}_1^\star - \boldsymbol{M}\|_F \leq \epsilon$, and then the RRPB is equivalent to the DGB.

3.6 Analytical Relation between Bounds. The following theorem describes the relation between PGB and RPB:

Table 1: Comparison of Sphere Bounds.

	Radius Convergence	Dynamic Screening	Reference Solution	Exact Optimality of Reference
GB	Can be > 0	Applicable	Primal	Not necessary
PGB	$= 0^a$	Applicable	Primal	Not necessary
DGB	$= 0$	Applicable	Primal/dual	Not necessary
CDGB	$= 0$	Applicable	Primal/dual	Not necessary
RPB	NA	Not applicable	Primal	Necessary
RRPB (RPB + DGB)	$= 0$	Applicable	Primal/dual	Not necessary

Note: The radius convergence indicates a radius when the reference solution is the optimal solution.
[a]For the hinge loss ($\gamma = 0$) case, a subgradient is required to be selected appropriately for achieving this convergence.

Theorem 7 *(Relation 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 exists a subgradient $\nabla P_{\lambda_1}(\boldsymbol{M}_0^\star)$ by which the PGB and RPB provide the same center and radius for $\boldsymbol{M}_1^\star$.*

The proof is presented in appendix I. The following theorem describes the relation between the DGB and RPB:

Theorem 8 *(Relation between DGB and RPB). Suppose that the optimal solutions $\boldsymbol{M}_0^\star, \boldsymbol{\alpha}_0^\star$, and $\boldsymbol{\Gamma}_0^\star$ for λ_0 are substituted into the reference solutions $\boldsymbol{M}, \boldsymbol{\alpha}$, and $\boldsymbol{\Gamma}$ of the 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.*

The proof is in appendix J. Figure 2c illustrates the relation between the DGB and RPB, which shows the theoretical advantage of the RPB for the regularization path setting.

Using the analytical results obtained thus far, we summarize relative relations between the bounds as follows. First, we consider the case in which the reference solution is optimal for λ_0 in the regularization path calculation. We obviously see $r_{\mathrm{GB}} \geq r_{\mathrm{PGB}}$ from Figure 2a, and from theorems 7 and 8, we see DGB $\supseteq$ PGB $=$ RPB $=$ RRPB. When the reference solution is an approximate solution in the regularization path calculation, we see only $r_{\mathrm{GB}} \geq r_{\mathrm{PGB}}$. For dynamic screening in which the reference solution is always an approximate solution, we see $r_{\mathrm{GB}} \geq r_{\mathrm{PGB}}$, and we also see RRPB $=$ DGB when ϵ is determined by DGB as written in equation 3.2.

Other properties of the bounds are summarized in Table 1. Although DGB and RRPB (RPB + DGB) have the same properties, our empirical evaluation in section 7.2 shows that RRPB often outperforms DGB in the regularization path calculation. (Note that although CDGB also has the same

properties as the above two methods, we omit it in the empirical evaluation because of its practical limitation, as we see in section 3.3.)

4 Safe Rules for Triplets

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 based on the following procedure:

Step 1: Identify the spherical region in which the optimal solution $\boldsymbol{M}^\star$ lies, based on the current feasible solution we refer to as the reference solution.
Step 2: For each triplet $(i, j, l) \in \mathcal{T}$, verify the 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, we showed that there exist a variety of approaches to creating the spherical region for step 1. In this section, we describe the procedure of step 2 given the sphere region.

Letting $\mathcal{B}$ be a region that contains $\boldsymbol{M}^\star$, the following screening rule can be derived from equation 2.5:

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

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

Based on these rules, $\hat{\mathcal{L}} \subseteq \mathcal{L}^\star$ and $\hat{\mathcal{R}} \subseteq \mathcal{R}^\star$ are constructed as

$$\hat{\mathcal{L}} = \left\{ (i, j, l) \;\middle|\; \max_{\boldsymbol{X}\in\mathcal{B}} \langle \boldsymbol{X}, \boldsymbol{H}_{ijl}\rangle < 1-\gamma \right\},$$

$$\hat{\mathcal{R}} = \left\{ (i, j, l) \;\middle|\; \min_{\boldsymbol{X}\in\mathcal{B}} \langle \boldsymbol{X}, \boldsymbol{H}_{ijl}\rangle > 1 \right\}.$$

We present an efficient approach to evaluating these rules. Because equation R1 can be evaluated in the same way as R2, we are concerned only with equation R2 henceforth.

4.1 Spherical 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 equation R2, we consider the following minimization problem, equation P1:

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

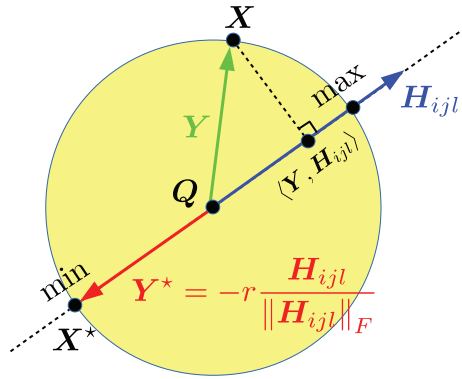


Figure 3: Spherical rule defined by equation P1. The yellow sphere indicates the region in which the optimal $M^\star$ must exist. The terms "max" and "min" indicate the points at which the maximum and minimum values of the inner product $\langle X, H_{ijl}\rangle$ are attained. If $\langle X^\star, H_{ijl}\rangle > 1$ holds, condition R2 is guaranteed to be satisfied.

Letting $Y := X - Q$, this problem is transformed into

$$\min_Y \langle Y, H_{ijl}\rangle + \langle Q, H_{ijl}\rangle \text{ s.t. } \|Y\|_F^2 \le r^2.$$

Because $\langle Q, H_{ijl}\rangle$ is a constant, this optimization problem entails minimizing the inner product $\langle Y, H_{ijl}\rangle$ under the norm constraint. The optimal $Y^\star$ of this optimization problem is easily derived as

$$Y^\star = -rH_{ijl}/\|H_{ijl}\|_F,$$

and then the minimum value of P1 is $\langle H_{ijl}, Q\rangle - r\|H_{ijl}\|_F$. Figure 3 shows a schematic illustration. This derives the following spherical rule:

$$\langle H_{ijl}, Q\rangle - r\|H_{ijl}\|_F > 1 \Rightarrow (i, j, l) \in \mathcal{R}^\star. \tag{4.1}$$

This condition can be easily evaluated for a given Q and r.

4.2 Spherical Rule with a Semidefinite Constraint. The spherical rule does not utilize the positive semidefiniteness of $M^\star$; therefore, a stronger rule can be constructed by incorporating a semidefinite constraint into equation P1:

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

Although the analytical solution is not available, equation P2 can be solved efficiently by transforming it into the semidefinite least squares (SDLS) problem (Malick, 2004).

Let $\mathcal{B}_{\mathrm{PSD}} := \{\boldsymbol{X} \mid \|\boldsymbol{X} - \boldsymbol{Q}\|_F^2 \leq r^2, \boldsymbol{X} \succeq \boldsymbol{O}\}$ be the feasible region of the optimization problem P2. To present the connection between SDLS and equation P2, we first assume that there exists a feasible solution $\boldsymbol{X}_0$ for equation P2 that satisfies $\langle \boldsymbol{X}_0, \boldsymbol{H}_{ijl} \rangle > 1$:

$$\exists \boldsymbol{X}_0 \text{ such that } \langle \boldsymbol{X}_0, \boldsymbol{H}_{ijl} \rangle > 1 \text{ and } \boldsymbol{X}_0 \in \mathcal{B}_{\mathrm{PSD}}. \tag{4.2}$$

Instead of equation P2, we consider the following SDLS problem:

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

If the optimal value of this problem is greater than r^2 (i.e., $\|\boldsymbol{X}^\star - \boldsymbol{Q}\|_F^2 > r^2$), there is no intersection between $\mathcal{B}_{\mathrm{PSD}}$ and the subspace defined by $\langle \boldsymbol{X}, \boldsymbol{H}_{ijl} \rangle = 1$:

$$\left\{\boldsymbol{X} \mid \langle \boldsymbol{X}, \boldsymbol{H}_{ijl} \rangle = 1,\ \boldsymbol{X} \in \mathcal{B}_{\mathrm{PSD}}\right\} = \emptyset. \tag{4.3}$$

From assumption 4.2, we have

$$\left\{\boldsymbol{X} \mid \langle \boldsymbol{X}, \boldsymbol{H}_{ijl} \rangle > 1,\ \boldsymbol{X} \in \mathcal{B}_{\mathrm{PSD}}\right\} \neq \emptyset. \tag{4.4}$$

As $\mathcal{B}_{\mathrm{PSD}}$ is a convex set, based on the two conditions 4.3 and 4.4, we derive

$$\left\{\boldsymbol{X} \mid \langle \boldsymbol{X}, \boldsymbol{H}_{ijl} \rangle \leq 1,\ \boldsymbol{X} \in \mathcal{B}_{\mathrm{PSD}}\right\} = \emptyset,$$

which indicates

$$\min_{\boldsymbol{X} \in \mathcal{B}_{\mathrm{PSD}}} \langle \boldsymbol{X}, \boldsymbol{H}_{ijl} \rangle > 1,$$

Thus, the condition of equation R2 is satisfied.

Based on the connection shown above, the rule evaluation, equation R2, with the semidefinite constraint is summarized as follows:

1. Select an arbitrary feasible solution $\boldsymbol{X}_0 \in \mathcal{B}_{\mathrm{PSD}}$. If $\langle \boldsymbol{X}_0, \boldsymbol{H}_{ijl} \rangle \leq 1$, we immediately see that the condition of equation R2 is not satisfied for the triplet (i, j, l). Otherwise, go to the next step. Note that in this case, assumption 4.2 is confirmed because $\langle \boldsymbol{X}_0, \boldsymbol{H}_{ijl} \rangle > 1$).
2. Solve SDLS. If the optimal value satisfies $\|\boldsymbol{X}^\star - \boldsymbol{Q}\|_F^2 > r^2$, the triplet (i, j, l) is guaranteed to be in $\mathcal{R}^\star$.

For calculating the second step, we derive the following dual problem of equation SDLS based on Malick (2004):

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

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

Although the computation of $[\boldsymbol{Q} + y\boldsymbol{H}_{ijl}]_+$ requires an eigenvalue decomposition, this computational requirement can be alleviated when the center $\boldsymbol{Q}$ of the hypersphere is positive semidefinite. The definition determines that $\boldsymbol{H}_{ijl}$ has at most one negative eigenvalue, and then $\boldsymbol{Q} + y\boldsymbol{H}_{ijl}$ also has at most one negative eigenvalue. Let $\lambda_{\min}$ be the negative (minimum) eigenvalue of $\boldsymbol{Q} + y\boldsymbol{H}_{ijl}$, and $\boldsymbol{q}_{\min}$ be the corresponding eigenvector. The projection $[\boldsymbol{Q} + y\boldsymbol{H}_{ijl}]_+$ can be expressed as $[\boldsymbol{Q} + y\boldsymbol{H}_{ijl}]_+ = (\boldsymbol{Q} + y\boldsymbol{H}_{ijl}) - \lambda_{\min}\boldsymbol{q}_{\min}\boldsymbol{q}_{\min}^\top$. Computation of the minimum eigenvalue and eigenvector is much easier than the full eigenvalue decomposition (Lehoucq & Sorensen, 1996).

As a special case, when $\boldsymbol{M}$ is a diagonal matrix, the semidefinite constraint is reduced to the nonnegative constraint, and analytical calculation of rule P2 is possible (see section 6.2).

4.3 Spherical Rule with Linear Constraint. Here, we reduce the computational complexity by considering the relaxation of the semidefinite constraint into a linear constraint. Suppose that a region defined by the linear inequality $\{\boldsymbol{X} \in \mathbb{R}^{d\times d} \mid \langle \boldsymbol{P}, \boldsymbol{X}\rangle \geq 0\}$ contains a semidefinite cone, $\mathbb{R}_+^{d\times d} \subseteq \{\boldsymbol{X} \in \mathbb{R}^{d\times d} \mid \langle \boldsymbol{P}, \boldsymbol{X}\rangle \geq 0\}$, for which we describe the determination of $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ later. Using this relaxed constraint, condition R2 is

$$\min_{X} \langle \boldsymbol{X}, \boldsymbol{H}_{ijl}\rangle \text{ s.t. } \|\boldsymbol{X} - \boldsymbol{Q}\|_F^2 \leq r^2, \; \langle \boldsymbol{P}, \boldsymbol{X}\rangle \geq 0. \tag{P3}$$

This problem can be solved analytically by considering the KKT conditions as follows (see appendix K).

Theorem 9 *(Analytical Solution of Equation P3). The optimal solution of equation P3 is as follows:*

$$\langle \boldsymbol{H}_{ijl}, \boldsymbol{X}^\star\rangle = \begin{cases} 0, & \text{if } \boldsymbol{H}_{ijl} = a\boldsymbol{P}, \\ \langle \boldsymbol{H}_{ijl}, \boldsymbol{Q}\rangle - r\|\boldsymbol{H}_{ijl}\|_F, & \text{if } \left\langle \boldsymbol{P}, \boldsymbol{Q} - r\frac{\boldsymbol{H}_{ijl}}{\|\boldsymbol{H}_{ijl}\|_F}\right\rangle \geq 0, \\ \left\langle \boldsymbol{H}_{ijl}, \frac{\beta\boldsymbol{P} - \boldsymbol{H}_{ijl}}{\alpha} + \boldsymbol{Q}\right\rangle, & \text{otherwise,} \end{cases}$$

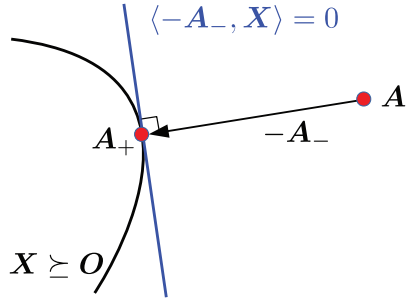


Figure 4: Linear relaxation of semidefinite constraint. From the projection of A to A_+, the supporting hyperplane $\langle -A_-, X \rangle = 0$ is constructed, and the half-space $\{X \mid \langle -A_-, X \rangle \geq 0\}$ contains the semidefinite cone $X \succeq O$.

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 semidefinite cone. Let $A \in \mathbb{R}^{d \times d}$ be a matrix external to the semidefinite cone as illustrated in Figure 4. In the figure, A_+ is the projection of A onto the semidefinite cone. For example, when the projected gradient for the primal problem (Weinberger & Saul, 2009) is used as an optimizer, A can be an update of the gradient descent $A = M - \eta \nabla P_\lambda(M)$ with some step size $\eta > 0$. Because $M - \eta \nabla P_\lambda(M)$ is projected onto the semidefinite cone at every iterative step of the optimization, no additional calculation is required to obtain A and A_+. Defining $A_- := A - A_+$, for any $X \succeq O$, we obtain

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

The inequality on the left has its origins in the property of a supporting hyperplane (Boyd & Vandenberghe, 2004), and for the inequality on the right, we use $\langle A_+, A_- \rangle = 0$. By setting $P = -A_-$, we obtain a linear approximation of the semidefinite constraint, which is a superset of the original semidefinite cone.

A necessary condition for performing our screening is that a loss function needs to have at least one linear region or a zero region. For example, the logistic loss cannot be used for screening because it has neither a linear nor a zero region.

Algorithm 1: Gradient Descent (Fixed Step Size η) with Safe Triplet Screening for RTLM.

function RTLM-STS(λ, $\boldsymbol{M}$, $\mathcal{T}$, freq, MaxIter, eps)

$\hat{\mathcal{R}} = \emptyset, \hat{\mathcal{L}} = \emptyset$

for $\mathrm{iter} = 0, 1, \ldots, \mathrm{MaxIter}$ **do**

$\tilde{\mathcal{T}} = \mathcal{T} - \hat{\mathcal{L}} - \hat{\mathcal{R}}$

if $\frac{\tilde{P}_\lambda(\boldsymbol{M}) - \tilde{D}_\lambda(\boldsymbol{\alpha})}{|\tilde{P}_\lambda(\boldsymbol{M})|} \leq \mathrm{eps}$ **then** ▷ convergence

return $\boldsymbol{M}$

end if

if $\mathrm{mod}(\mathrm{iter}, \mathrm{freq}) = 0$ **then** ▷ perform safe triplet screening

Compute a bound $\mathcal{B}$ based on the reference solution $\boldsymbol{M}$ ▷ section 3

for $(i, j, l) \in \tilde{\mathcal{T}}$ **do**

if $\max_{\boldsymbol{X} \in \mathcal{B}} \langle \boldsymbol{H}_{ijl}, \boldsymbol{X} \rangle < 1 - \gamma$ **then** ▷ evaluate safe rule (section 4)

$\hat{\mathcal{L}} = \hat{\mathcal{L}} \cup \{(i, j, l)\}$

end if

if $\min_{\boldsymbol{X} \in \mathcal{B}} \langle \boldsymbol{H}_{ijl}, \boldsymbol{X} \rangle > 1$ **then** ▷ evaluate safe rule (section 4)

$\hat{\mathcal{R}} = \hat{\mathcal{R}} \cup \{(i, j, l)\}$

end if

end for

end if

$\boldsymbol{M} = \left[\boldsymbol{M} - \eta \nabla \tilde{P}_\lambda(\boldsymbol{M})\right]_+$ ▷ update $\boldsymbol{M}$ (η: step size)

end for

return $\boldsymbol{M}$

end function

5 Computations

Algorithm 1 shows the detailed procedure of our safe screening with simple fixed step-size gradient descent. (Note that any other optimization

algorithm can be combined with our screening procedure.) In the algorithm, for every `freq` iteration of the gradient descent, the screening rules are evaluated by using the current solution $\boldsymbol{M}$ as the reference solution. As the quality of the approximate solution $\boldsymbol{M}$ improves, the larger the number of triplets that can be removed from $\mathcal{T}$. Thus, the quality of the initial solution affects the efficiency. In the case of the regularization path calculation, in which RTLM is solved for a sequence of λs, a reasonable initial solution is the approximate solution to the previous λ. We discuss a further extension specific to the regularization path calculation in section 6.1.

Considering the computational cost of the screening procedure of algorithm 1, the rule evaluation (step 2) described in section 4 is often dominant, because the rule needs to be evaluated for each one of the triplets. The sphere, constructed in step 1, can be fixed during the screening procedure as long as the reference solution is fixed.

To evaluate the spherical rule, equation 4.1, given the center $\boldsymbol{Q}$ and the radius r, the inner product $\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q} \rangle$ and the norm $\|\boldsymbol{H}_{ijl}\|_F$ need to be evaluated. The inner product $\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q} \rangle$ can be calculated in $O(d^2)$ operations because it is expanded as a sum of quadratic forms: $\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q} \rangle = (\boldsymbol{x}_i - \boldsymbol{x}_l)^\top \boldsymbol{Q}(\boldsymbol{x}_i - \boldsymbol{x}_l) - (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{Q}(\boldsymbol{x}_i - \boldsymbol{x}_j)$. Further, we can reuse this term from the objective function $P_\lambda(\boldsymbol{M})$ calculation in the case of the DGB, RPB, and RRPB. The norm $\|\boldsymbol{H}_{ijl}\|_F$ can be calculated in $O(d)$ operations, and this is constant throughout the optimization process. Thus, for the DGB, RPB, or RRPB, it is possible to reduce the additional computational cost of the spherical rule for (i, j, l) to $O(1)$ by calculating $\|\boldsymbol{H}_{ijl}\|_F$ beforehand. The computational cost of the spherical rule with the semidefinite constraint (see section 4.2) is that of the SDLS algorithm. The SDLS algorithm needs $O(d^3)$ because of the eigenvalue decomposition in every iterative cycle, which may considerably increase the computational cost. The computational cost of the spherical rule with the linear constraint (see section 4.3) is $O(d^2)$.

6 Extensions

6.1 Range-Based Extension of Triplet Screening. The screening rules presented in section 4 relate to the problem of a fixed λ. In this section, we regard a screening rule as a function of λ to derive a range of λs in which the screening rule is guaranteed to be satisfied. This is particularly useful for calculating the regularization path 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 the screening rules.

6.1.1 Deriving the Range. Let

$$\boldsymbol{Q} = \boldsymbol{A} + \boldsymbol{B}\frac{1}{\lambda} \tag{6.1}$$

be the general form of the center of a hypersphere for some constant matrices $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ and

$$r^2 = a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2} \tag{6.2}$$

be the general form of the radius for some constants $a \in \mathbb{R}$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$. The GB, DGB, RPB, and RRPB can be in this form (details are provided in appendix L, section L.1). Note that in the RRPB, equation 3.1, λ_1 is regarded as λ in the general form and λ_0 is a constant. The condition of the spherical rule $\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q}\rangle - r\left\|\boldsymbol{H}_{ijl}\right\|_F > 1$ in equation 4.1 can be rewritten as

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

with the assumption

$$\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q}\rangle - 1 > 0.$$

Because $\langle \boldsymbol{H}_{ijl}, \boldsymbol{Q}\rangle = \langle \boldsymbol{H}_{ijl}, \boldsymbol{A}\rangle + \langle \boldsymbol{H}_{ijl}, \boldsymbol{B}\rangle\frac{1}{\lambda}$, these two inequalities can be transformed into quadratic and linear functions of λ, respectively. The range of λ that satisfies the two inequalities simultaneously represents the range of λ in which a triplet (i, j, l) must be in $\mathcal{R}^*$. 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 10 *(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* $\|\boldsymbol{M}_0^\star - \boldsymbol{M}_0\|_F \le \epsilon$*, 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

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

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

Refer to section L.2 for the proof. The computational procedure for range-based screening is shown in algorithm 2.

6.1.2 Consideration for Range Extension with Other Bounds. As shown in equation 3.1, the RRPB is based on the optimality ϵ for the current λ_0, and does not depend on the optimality for λ_1, which is regarded as λ in the general form of equations 6.1 and 6.2. Because of this property, the RRPB is

Algorithm 2: Regularization Path Calculation with Range-Based Safe Triplet Screening.

Require: $\lambda_0 > \lambda_1 > \cdots > \lambda_T > 0, \mathcal{T}, \mathrm{freq}, \mathrm{MaxIter}, \mathrm{eps}$

$\mathrm{Range}^{ijl}_{\mathcal{R}} = \emptyset, \mathrm{Range}^{ijl}_{\mathcal{L}} = \emptyset$ for all $(i,j,l) \in \mathcal{T}$ ▷ initialize range of λ

$\boldsymbol{M}_{-1} = \boldsymbol{O}$ ▷ initialize $\boldsymbol{M}$

for $t = 0, 1, \ldots, T$ **do** ▷ iterate for each λ

$\hat{\mathcal{R}} = \emptyset, \hat{\mathcal{L}} = \emptyset$

if $t \neq 0$ **then**

for $(i,j,l) \in \mathcal{T}$ **do**

if $\lambda_t \in \mathrm{Range}^{ijl}_{\mathcal{L}}$ **then** ▷ range-based safe screening for $\mathcal{L}$

$\hat{\mathcal{L}} = \hat{\mathcal{L}} \cup \{(i,j,l)\}$

else if $\lambda_t \in \mathrm{Range}^{ijl}_{\mathcal{R}}$ **then** ▷ range-based safe screening for $\mathcal{R}$

$\hat{\mathcal{R}} = \hat{\mathcal{R}} \cup \{(i,j,l)\}$

else

Update $\mathrm{Range}^{ijl}_{\mathcal{R}}, \mathrm{Range}^{ijl}_{\mathcal{L}}$ based on $\boldsymbol{M}_{t-1}$ ▷ Theorem 10

end if

end for

end if

$\tilde{\mathcal{T}} = \mathcal{T} - \hat{\mathcal{L}} - \hat{\mathcal{R}}$

$\boldsymbol{M}_t = \text{RTLM-STS}(\lambda_t, \boldsymbol{M}_{t-1}, \tilde{\mathcal{T}}, \mathrm{freq}, \mathrm{MaxIter}, \mathrm{eps})$ ▷ solve reduced-size problem

end for

particularly suitable to range-based screening among the spheres we derived thus far. To calculate ϵ, equation 3.2 for the RRPB, the duality gap $P_{\lambda_0}(\boldsymbol{M}_0) - D_{\lambda_0}(\boldsymbol{\alpha}_0, \boldsymbol{\Gamma}_0)$ is required. Instead of the original $P_{\lambda_0}(\boldsymbol{M}_0) - D_{\lambda_0}(\boldsymbol{\alpha}_0, \boldsymbol{\Gamma}_0)$, we can use problems with a reduced size, $\tilde{P}_{\lambda_0}(\boldsymbol{M}_0) - \tilde{D}_{\lambda_0}(\boldsymbol{\alpha}_0, \boldsymbol{\Gamma}_0)$, for efficient computation, where $\tilde{D}_{\lambda_0}$ is the dual objective in which $\alpha_i = 0$ for $i \in \hat{\mathcal{R}}$ and $\alpha_i = 1$ for $i \in \hat{\mathcal{L}}$ are fixed. Because the reduced-size problem shares

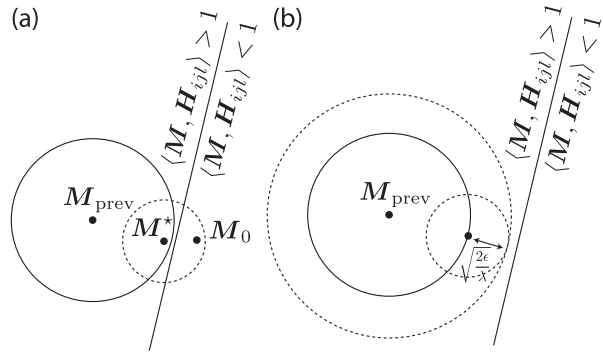


Figure 5: (a) Suppose that $M^\star$ is the optimal solution for λ_0, and some iterative optimization algorithm obtains M_{prev} in the middle of the optimization process. The circle around M_{prev} represents the DGB, which contains $M^\star$. Then the screening rule can eliminate the triplet (i, j, l) because $\langle M, H_{ijl}\rangle > 1$ holds for any points in the circle. Now suppose that M_0 is the approximate solution we obtain after the optimization algorithm terminates with some small tolerance of the duality gap. The circle with the dashed line represents the region in which the duality gap is less than the tolerance. Although M_0 satisfies the terminate condition, the inequality $\langle M, H_{ijl}\rangle > 1$ does not hold for M_0. In this case, we cannot ignore this triplet (i, j, l) to evaluate the duality gap for different $\lambda \neq \lambda_0$ because it causes a nonzero penalty. (b) An enlarged bound. Because of the inequality of DGB $\|M^\star - M_0\|_F \leq \sqrt{2\epsilon/\lambda}$, this enlarged region contains any approximate solutions with the duality gap $\leq \epsilon$.

exactly the same optimal solution with the original problems, this gap also provides a valid bound. As a result, we can avoid computing the sum over all triplets in $\mathcal{T}$ (e.g., to calculate the loss term in the original primal) to evaluate a bound.

In the other bounds, the loss term in the primal objective needs to be carefully considered. Suppose that we have an approximate solution M_0 for λ_0 as a reference solution. To regard a bound as a function of λ in the GB and PGB, it is necessary to consider the gradient for λ (i.e., $\nabla P_\lambda(M_0)$), and the DGB requires the objective value for λ (i.e., $P_\lambda(M_0)$). These two terms may not be correctly calculated if we replace them with the reduced-size primal created for λ_0. Figure 5a illustrates an example of this problem in the case of DGB. To safely replace $P_\lambda(M_0)$ with the reduced-size primal $\tilde{P}_\lambda(M_0)$ for these cases, the following conditions need to hold:

$$\langle M_0, H_{ijl}\rangle < 1 - \gamma, \text{ for } \forall (i, j, l) \in \hat{\mathcal{L}},$$
$$\langle M_0, H_{ijl}\rangle > 1, \text{ for } \forall (i, j, l) \in \hat{\mathcal{R}}.$$

If the reference solution $\boldsymbol{M}_0$ is exactly optimal for λ_0, these conditions hold. However, in practice, this cannot be true because of numerical errors, and furthermore, the optimization algorithm is usually terminated with some tolerance.

This problem can be avoided by enlarging the radius of spherical bounds such that the bound contains the approximate solution. Assuming that $\boldsymbol{M}_0$ is an approximate solution with the duality gap ϵ, then from the DGB, we see that the distance between $\boldsymbol{M}_0$ and the optimal solution $\boldsymbol{M}^\star$ satisfies

$$\|\boldsymbol{M}^\star - \boldsymbol{M}_0\|_F \leq \sqrt{\frac{2\epsilon}{\lambda}}.$$

This inequality indicates that by enlarging the radius of the hypersphere by $\sqrt{\frac{2\epsilon}{\lambda}}$, we can guarantee that the bound includes any approximate solutions (Figure 5b shows an illustration). Using the radius r introduced in section 6.1, we obtain the enlarged radius R as follows:

$$R = \underbrace{\sqrt{a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2}}}_{r} + \sqrt{\frac{2\epsilon}{\lambda}}. \tag{6.3}$$

The reduced-size problems created by this enlarged radius can be safely used to evaluate the duality gap for any λ. However, this enlarged radius no longer has the general form of the radius 6.2 we assumed to derive the range. Although it is possible to derive a range even for the enlarged radius R, the calculation becomes quite complicated, and thus we do not pursue this direction in this study. (Appendix M shows the computational procedure.) Further, an increase in the radius may decrease the screening rate.

6.2 Screening with Diagonal Constraint. When the matrix $\boldsymbol{M}$ is constrained to be a diagonal matrix, metric learning is reduced to feature weighting in which the Mahalanobis distance, equation 2.1, simply adapts a weight of each feature without combining different dimensions. Although correlation in different dimensions is not considered, this simpler formulation is useful to avoid a large computational cost for high-dimensional data mainly because of the following two reasons:

- The number of variables in the optimization decreases from d^2 to d.
- The semidefinite constraint for a diagonal matrix is reduced to the nonnegative constraint of diagonal elements.

Both properties are also beneficial for efficient screening rule evaluation; in particular, the second property makes the screening rule with the semidefinite constraint easier to evaluate.

The minimization problem of the spherical rule with the semi-definite constraint, equation P2, is simplified as

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \boldsymbol{x}^\top \boldsymbol{h}_{ijl} \text{ s.t. } \left\|\boldsymbol{x}-\boldsymbol{q}\right\|_2^2 \le r^2,\ \boldsymbol{x}\ge \boldsymbol{0}, \tag{P4}$$

where $\boldsymbol{h}_{ijl} := \mathrm{diag}(\boldsymbol{H}_{ijl})$. Let

$$L(\boldsymbol{x},\alpha,\boldsymbol{\beta}) := \boldsymbol{x}^\top \boldsymbol{h}_{ijl} - \alpha(r^2 - \left\|\boldsymbol{x}-\boldsymbol{q}\right\|_2^2) - \boldsymbol{\beta}^\top \boldsymbol{x},$$

be the Lagrange function of equation P4, where $\alpha \ge 0$ and $\boldsymbol{\beta}\ge \boldsymbol{0}$ are dual variables. The KKT conditions are written as

$$\partial L/\partial \boldsymbol{x} = \boldsymbol{h}_{ijl} + 2\alpha(\boldsymbol{x}-\boldsymbol{q}) - \boldsymbol{\beta} = \boldsymbol{0}, \tag{6.4a}$$

$$\alpha(r^2 - \left\|\boldsymbol{x}-\boldsymbol{q}\right\|_2^2) = 0,\ \beta_k x_k = 0, \tag{6.4b}$$

$$\alpha \ge 0,\ r^2 - \left\|\boldsymbol{x}-\boldsymbol{q}\right\|_2^2 \ge 0,\ \boldsymbol{\beta}\ge\boldsymbol{0},\ \boldsymbol{x}\ge\boldsymbol{0}. \tag{6.4c}$$

We derive the analytical representation of the optimal solution for cases of $\alpha > 0$ and $\alpha = 0$, respectively. For $\alpha > 0$, the following theorem is obtained.

Theorem 11. *If the optimal dual variable satisfies $\alpha > 0$, the optimal $\boldsymbol{x}$ and $\boldsymbol{\beta}$ of equation P4 can be written as*

$$x_k = \begin{cases} q_k - h_{ijl,k}/2\alpha, & \text{if } h_{ijl,k} - 2\alpha q_k \le 0, \\ 0, & \text{otherwise,} \end{cases} \tag{6.5}$$

and

$$\boldsymbol{\beta} = \boldsymbol{h}_{ijl} + 2\alpha(\boldsymbol{x}-\boldsymbol{q}). \tag{6.6}$$

Then $\boldsymbol{x}$ also satisfies

$$\|\boldsymbol{x}-\boldsymbol{q}\|_2^2 = r^2. \tag{6.7}$$

For $\alpha = 0$, the following theorem is obtained.

Theorem 12. *If the optimal dual variable satisfies $\alpha = 0$, the optimal $\boldsymbol{x}$ and $\boldsymbol{\beta}$ of equation P4 can be written as*

$$x_k = \begin{cases} 0, & \text{if } h_{ijl,k} > 0 \\ max\ \{q_k, 0\}, & \text{otherwise} \end{cases}, \tag{6.8}$$

and

$$\boldsymbol{\beta} = \boldsymbol{h}_{ijl}. \tag{6.9}$$

The proofs for theorems 11 and 12 are in sections N.1 and N.2, respectively.

Based on the theorems, the optimal solution of equation P4 can be calculated analytically. The detail of the procedure is shown in section N.3, which requires $O(d^2)$ computations. Although this procedure obtains the solution by using the fixed steps of analytical calculations, for larger values of d, iterative optimization algorithms can be faster. For example, we can apply the SDLS dual ascent to problem P4 in which each iterative step takes $O(d)$.

6.3 Applicability to More General Formulation. Throughout the letter, we analyze screening theorems based on the optimization problem defined by Primal. RTML is the Frobenius norm-regularized triplet loss minimization, which has been shown to be an effective formulation of metric learning (Schultz & Joachims, 2004; Shen et al., 2014). Further, with slight modifications, our screening framework can accommodate a wider range of metric learning methods. Here we redefine the optimization problem as follows:

$$\min_{\boldsymbol{M} \succeq \boldsymbol{O}} \sum_i \ell\left(\langle \boldsymbol{M}, \boldsymbol{C}_i \rangle\right) + \frac{\lambda}{2} \|\boldsymbol{M}\|_F^2 , \tag{6.10}$$

where $\boldsymbol{C}_i \in \mathbb{R}^{d \times d}$ is a constant matrix. All our sphere bounds (GB, PGB, DGB, RPB, and RRPB) still hold for this general representation if the loss function $\ell : \mathbb{R} \to \mathbb{R}$ is convex subdifferentiable. The rules (spherical rule, sphere with semidefinite constraint, and sphere with linear constraint) can also be constructed if the loss function has the form of the hinge type loss function, equations 2.2 and 2.3, by which standard hinge-, smoothed hinge–, and squared hinge–loss functions are included.

We can incorporate an additional linear term into the objective function 6.10. Defining a pseudo-loss function $\tilde{\ell}(x) = -x$, we write the primal problem with a linear term as

$$\min_{\boldsymbol{M} \succeq \boldsymbol{O}} P_\lambda(\boldsymbol{M}) := \sum_{ijl} \ell\left(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl} \rangle\right) + \tilde{\ell}(\langle \boldsymbol{M}, \boldsymbol{C} \rangle) + \frac{\lambda}{2} \|\boldsymbol{M}\|_F^2 , \tag{6.11}$$

which can be seen as a special case of equation 6.10 because $\tilde{\ell}$ is convex subdifferentiable. Suppose that $\eta_{ij} \in \{0, 1\}$ indicates whether $\boldsymbol{x}_j$ is a target neighbor of $\boldsymbol{x}_i$, which is a neighbor of $\boldsymbol{x}_i$ having the same label. When we define $\boldsymbol{C} := -\sum_{ij} \eta_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$ and employ the hinge loss, equation 2.2, this formulation is the well-known LMNN (Weinberger & Saul, 2009) with the additional Frobenius norm regularization. Another interpretation of this linear term is the trace norm regularization (Kulis, 2013). For the pseudo-loss term $\tilde{\ell}$, the derivative is $\nabla \tilde{\ell}(x) = -1$, and the conjugate is $\tilde{\ell}^*(-a) = 0$ if $-a = -1$; otherwise, ∞, where a is a dual variable. Then, by

using the derivation of the dual in appendix A, the dual problem is modified as

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

where

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

Because equation 6.11 is a special case of equation 6.10, all spheres can be derived, and we can construct the same screening rules for α_{ijl} for $(i,j,l)\in\mathcal{T}$. The only difference is that the dual variable a is not associated with any screening rule because it is fixed to 1 by the dual constraint.

About the loss term, pairwise- and quadruplet-loss functions can also be incorporated into our framework. The pairwise approach considers a set of pairs in the same class $\mathcal{S}$ and a set of pairs in the different classes $\mathcal{D}$. Davis et al. (2007) considered constraints with threshold parameters U and L: $d^2_{\boldsymbol{M}}(\boldsymbol{x}_i,\boldsymbol{x}_j)\le U$ for $(i,j)\in\mathcal{S}$ and $d^2_{\boldsymbol{M}}(\boldsymbol{x}_i,\boldsymbol{x}_l)\ge L$ for $(i,l)\in\mathcal{D}$. Let $\ell_t(x) = [t-x]_+$ be a hinge loss function with threshold t. By using ℓ_t, the above two constraints can be relaxed to soft constraints that result in

$$\min_{\boldsymbol{M}\succeq O}\sum_{(i,j)\in\mathcal{S}}\ell_{-U}\left(\langle\boldsymbol{M},-\boldsymbol{C}_{ij}\rangle\right) + \sum_{(i,l)\in\mathcal{D}}\ell_L\left(\langle\boldsymbol{M},\boldsymbol{C}_{il}\rangle\right) + \frac{\lambda}{2}\|\boldsymbol{M}\|_F^2.$$

Because of the threshold parameters, the second term of the dual problem, Dual2, changes from $\boldsymbol{\alpha}^\top\mathbf{1}$ to $\boldsymbol{\alpha}^\top\boldsymbol{t}$, where $\boldsymbol{t} := [L,\cdots,L,-U,\cdots,-U]^\top\in\mathbb{R}^{|\mathcal{D}|+|\mathcal{S}|}$. Our bounds still hold because ℓ_t is convex subdifferentiable, and screening rules are formulated as evaluating whether the inner product $\langle\boldsymbol{M},\boldsymbol{C}_{il}\rangle$ (or $\langle\boldsymbol{M},-\boldsymbol{C}_{ij}\rangle$) is larger or smaller than the threshold t.

Law et al. (2013) proposed a loss function based on a quadruplet of instances. The basic idea is to compare pairs of dissimilarity $d^2_{\boldsymbol{M}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ and $d^2_{\boldsymbol{M}}(\boldsymbol{x}_k,\boldsymbol{x}_l)$. For example, when (k,l) should be more dissimilar than (i,j), the loss is defined as $\ell(d^2_{\boldsymbol{M}}(\boldsymbol{x}_k,\boldsymbol{x}_l) - d^2_{\boldsymbol{M}}(\boldsymbol{x}_i,\boldsymbol{x}_j))$. They define the following optimization problem,

$$\min_{\boldsymbol{M}\succeq O}\sum_{(i,j,k,l)\in\mathcal{Q}}\ell\left(\langle\boldsymbol{M},\boldsymbol{C}_{ijkl}\rangle\right) + \frac{\lambda}{2}\|\boldsymbol{M}\|_F^2,$$

where $\boldsymbol{C}_{ijkl} = (\boldsymbol{x}_k-\boldsymbol{x}_l)(\boldsymbol{x}_k-\boldsymbol{x}_l)^\top - (\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top$ and $\mathcal{Q}$ is a set of quadruplets. This is also a special case of equation 6.10.

Table 2: Summary of Data Sets.

	#dimension	#sample	#classes	k	#triplet	$\lambda_{\max}$	$\lambda_{\min}$
Iris	4	150	3	∞	546,668	1.3e+7	2.3e+1
Wine	13	178	3	∞	910,224	2.0e+7	5.1e+1
Segment	19	2310	7	20	832,000	2.5e+6	4.2e+0
Satimage	36	4435	6	15	898,200	1.0e+7	8.8e+0
Phishing	68	11,055	2	7	487,550	5.0e+3	2.0e−1
SensIT Vehicle	100	78,823	3	3	638,469	1.0e+4	2.9e+0
a9a	16[a]	32,561	2	5	732,625	1.2e+5	3.1e+2
Mnist	32[a]	60,000	10	5	1,350,025	7.0e+3	9.6e−1
Cifar10	200[a]	50,000	10	2	180,004	2.0e+3	3.3e+1
Rcv1.multiclass	200[b]	15,564	53	3	126,018	3.0e+2	6.0e−4

Note: #triplet and $\lambda_{\min}$ are the average value for subsampled random trials.
[a]The dimension was reduced by AutoEncoder.
[b]The dimension was reduced by PCA.

Note that pairwise-, triplet- and quadruplet-loss functions can be used simultaneously, and safe screening can be applied to remove any of those loss terms.

7 Experiment

We evaluate the performance of safe triplet screening using the benchmark data sets listed in Table 2, which are from LIBSVM (Chang & Lin, 2011) and the Keras data set (Chollet et al., 2015). We create a set of triplets by following the approach by Shen et al. (2014), in which k neighborhoods in the same class $\boldsymbol{x}_j$ and k neighborhoods in a 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 $\lambda_0, \lambda_1, \ldots, \lambda_T$. To determine $\lambda_0 = \lambda_{\max}$, from a sufficiently large λ in which $\mathcal{R}$ is empty, we gradually reduced λ by multiplying it by 0.9 and started the regularization path calculation from λ in which $\mathcal{R}$ is not empty. 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:

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

where $\text{loss}(\lambda_t)$ is the loss function value at λ_t. We randomly selected 90% of the instances of each data set five 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 became less than 10^{-6}. For the loss function ℓ, we used the smoothed hinge loss of $\gamma = 0.05$. (We also provide results for the hinge loss

in section 7.4.1). We performed safe triplet screening after every 10 iterative cycles of the gradient descent. We refer to the first screening for a specific λ_t, in which the solution of the previous λ_{t-1} is used as the reference solution for regularization path screening. The screening performed during the optimization process (after regularization path screening) is termed dynamic screening. We performed both of these screening procedures for all experiments. As a baseline, we refer to the RTLM optimization without screening as naive optimization. We initialized with $\boldsymbol{M} = \boldsymbol{O}$ at λ_0 because $\boldsymbol{M} = \boldsymbol{O}$ is the optimal solution when $\lambda \to \infty$. 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

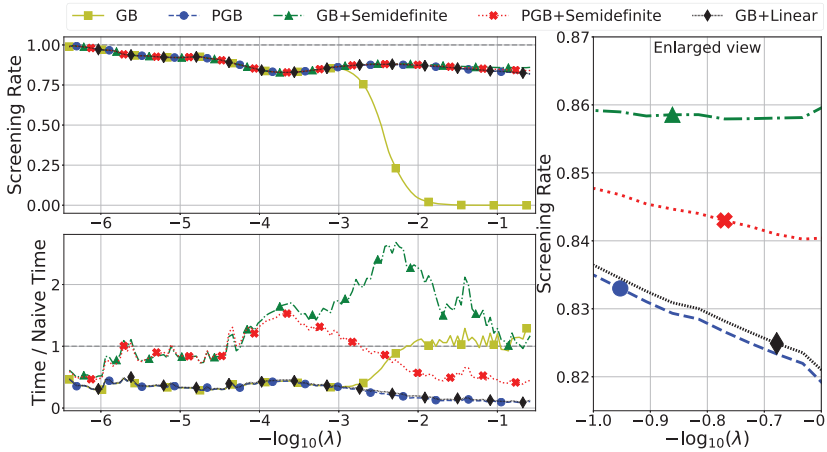
$$\frac{1}{2}\left|\frac{\langle \Delta \boldsymbol{M}, \Delta \boldsymbol{G}\rangle}{\langle \Delta \boldsymbol{G}, \Delta \boldsymbol{G}\rangle} + \frac{\langle \Delta \boldsymbol{M}, \Delta \boldsymbol{M}\rangle}{\langle \Delta \boldsymbol{M}, \Delta \boldsymbol{G}\rangle}\right|,$$

where $\Delta \boldsymbol{M} = \boldsymbol{M}_t - \boldsymbol{M}_{t-1}$, $\Delta \boldsymbol{G} = \nabla P_\lambda(\boldsymbol{M}_t) - \nabla P_\lambda(\boldsymbol{M}_{t-1})$ (Barzilai & Borwein, 1988). In SDLS dual ascent, we used the conjugate gradient method (Yang, 1993) to find the minimum eigenvalue.

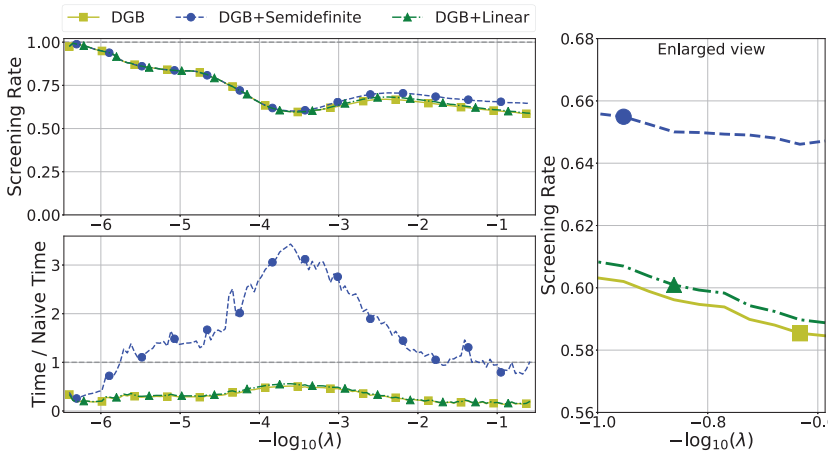
7.1 Comparing Rules. We first validate the screening performance (screening rate and CPU time) of each screening rule introduced in section 4 by using algorithm 2 without the range-based screening process. Here, the screening rate is defined by #screened triplets/$|\{(i, j, l) \in \mathcal{T} \mid \langle \boldsymbol{H}_{ijl}, \hat{\boldsymbol{M}}\rangle > 1 \text{ or } \langle \boldsymbol{H}_{ijl}, \hat{\boldsymbol{M}}\rangle < 1 - \gamma\}|$ where $\hat{\boldsymbol{M}}$ is the solution after convergence.

7.1.1 GB-Based Rules. Here we use the GB and PGB as spheres, and we observe the effect of the semidefinite constraint in the rules. As a representative result, Figure 6a compares the performance of the rules by using segment data.

First, except for the GB, the rules maintain a high screening rate for the entire regularization path, as shown in the top left plot. Note that this rate is only for regularization path screening, meaning that dynamic screening can further increase the screening rate during the optimization, as discussed in the section 7.1.2. The bottom left plot of the same figure shows that PGB and GB+Linear are the most efficient and achieved CPU times approximately 2 to 10 times faster than the naive optimization. The screening rate of the GB was severely reduced along the latter half of the regularization path. As illustrated in Figure 2a, the center of the GB can be external to the semidefinite 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 with their center inside the semidefinite cone. This causes performance deterioration particularly for smaller values of λ, because the minimum of the loss term is usually outside the semidefinite cone.



(a) GB-based screening rules.



(b) DGB-based screening rules.

Figure 6: Comparison of screening rules on the segment data set. For both panels a and b, the plots are aligned as follows. (Top left) Performance of regularization path screening. (Bottom left) Ratio of CPU time compared with the naive optimization for each λ. (Right) Enlargement of the upper left plot for the range $-\log_{10}(\lambda) \in [-1, -0.6]$.

The screening rates of GB+Linear and GB+Semidefinite are slightly higher than that of the PGB (the plot on the right), which can be seen from their geometrical relation illustrated in Figure 2a. GB+Semidefinite achieved the highest screening rate, but eigenvalue decomposition is

necessary to repeatedly perform the calculation in SDLS, which resulted in the CPU time increasing along the latter half of the path. Although PGB+Semidefinite is also tighter than PGB, the CPU time increased from approximately $-\log_{10}(\lambda) \approx -4$ to -3. Because the center of PGB is positive semidefinite, only the minimum eigenvalue is required (see section 4.2), but it can still increase the CPU time.

Among the screening methods compared here, our empirical analysis suggests that the use of the spherical rule with PGB, in which a semidefinite constraint is implicitly incorporated in the projection process, is the most cost-effective. We did not observe that the other approach to considering the semidefinite (or relaxed linear) constraint in the rule substantially outperforms PGB in terms of CPU time despite its high screening rate. We observed the same tendency for DGB. The screening rate did not change markedly even if the semidefinite constraint was explicitly considered.

7.1.2 DGB-Based Rules. Next, by using the DGB, we compared the performance of the three rules presented in section 4. Figure 6b shows the results, which are similar to those obtained for the GB, shown in Figure 6a. The semidefinite and the linear constraint slightly improve the rate. However, the large computational cost for screening with the semidefinite constraint caused the overall CPU time to increase. Therefore, although the linear constraint is much easier to evaluate, the CPU time was almost the same as that required for the DGB because of the slight improvement in the screening rate.

7.2 Comparing Bounds. Here we compare the screening performance (screening rate and CPU time) of each bound introduced in section 3 by using algorithm 2 without the range-based screening process. We do not use RPB because it needs the strictly optimal previous solution.

Based on the results in the previous section, we employed the spherical rule. The result obtained for the phishing data set is shown in Figure 7. The screening rate of the GB (top right) again decreased from the middle of the horizontal axis compared with the other spheres. The other spheres also have lower screening rates for small values of λs. As mentioned in section 6.1, the radii 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 $r \to \infty$. In the case of the PGB, although the dependency on λ cannot be written explicitly, the same tendency was observed. We see that the PGB and RRPB have similar results as suggested by theorem 7, and the screening rate of the DGB is lower than that of the RRPB, as suggested by theorem 8. A comparison of the PGB and RRPB indicated that the former achieved a higher screening rate, but the latter is more efficient with less CPU time, as shown in the plot at the bottom right, because the PGB requires a matrix inner product calculation for each triplet. Bounds other than the GB are more than twice as fast as the naive calculation for most values of λ.

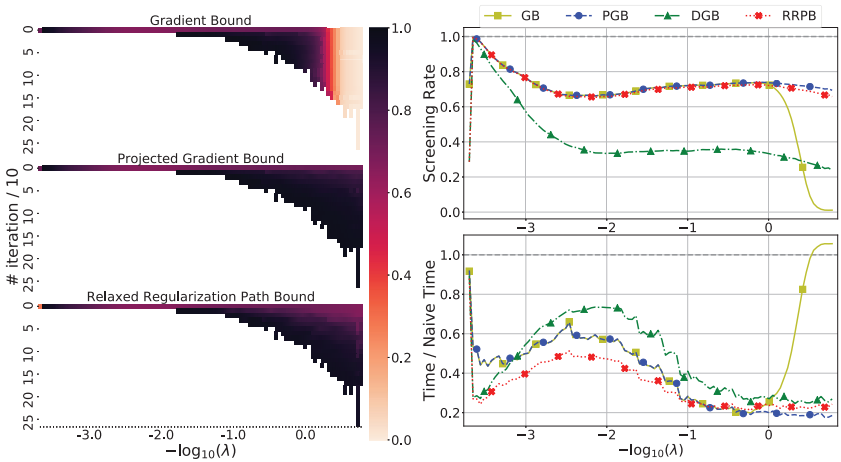


Figure 7: Comparison of spherical bounds on the phishing data set. The heat maps on the left show the dynamic screening rate. The vertical axes of these heat maps represent the number of iterative steps required for the optimization divided by 10 to perform the screening. (Top right) Rate of regularization path screening. (Bottom right) Ratio of CPU time compared with naıve optimization.

A comparison of the dynamic screening rate (the three plots on the left in Figure 7) of PGB and RRPB shows that the rate of PGB is higher. In terms of the regularization path screening (top right), RRPB and PGB have similar screening rates, but PGB has a higher dynamic screening rate. Along the latter half of the regularization path, the number of gradient descent iterations increases; consequently, the dynamic screening significantly affects the CPU time, and the PGB becomes faster despite the additional computation it requires to compute the inner product.

We further evaluate the performance of the range-based extension described in section 6.1. Figure 8 shows the rate of the range-based screening for the segment data set. The figure shows that a wide range of λ can be screened, particularly for small values of λ; although the range is smaller for large values of λ, than for the small values, a high screening rate is observed when λ approaches λ_0. A significant advantage of this approach is that for those triplets screened by using the specified range, we no longer need to evaluate the screening rule as long as λ is within the range.

The total CPU time for the regularization path is shown in Figure 9. In addition to GB, PGB, DGB, and RRPB, we further evaluate the performance when PGB and RRPB are used simultaneously. The use of two rules can improve the screening rate; however, additional computations are required to evaluate the rule. In the figure, for four out of six data sets, the PGB+RRPB combination requires the least CPU time.

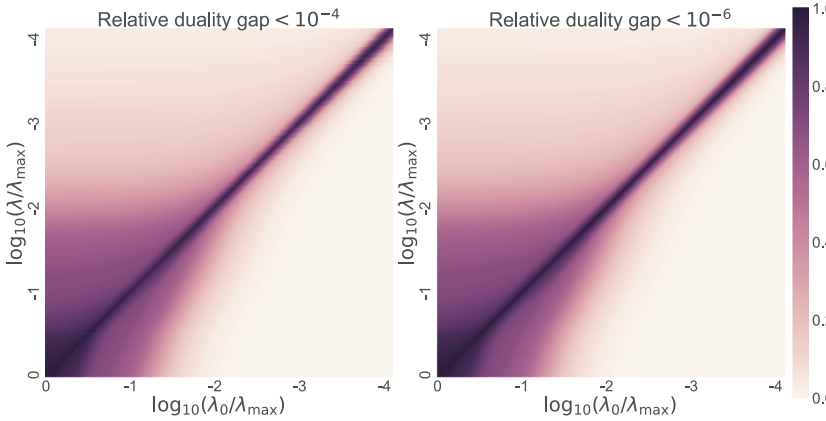


Figure 8: Screening rate of range-based screening on the segment data set. The color indicates the screening rate for λ on the vertical axis based on the reference solution using λ_0 on the horizontal axis. The accuracy of the reference solution is 10^{-4} and 10^{-6} for the plots on the left and right, respectively.

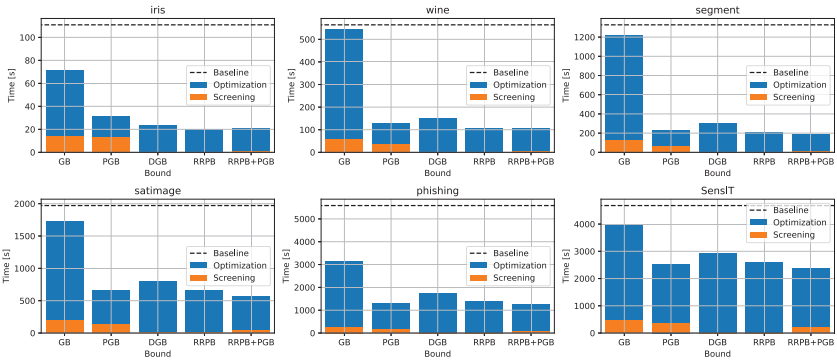


Figure 9: Total CPU time of regularization path (seconds). The term RRPB+PGB indicates that the spherical rules are performed with these two spheres. "Baseline" indicates the computational time without screening. The bars show the total time and the proportion of the optimization algorithm and the screening calculation.

7.3 Evaluating the Practical Efficiency. We next considered a computationally more expensive setting to evaluate the effectiveness of the safe screening approach in a practical situation. To investigate the regularization path more precisely, we set a finer grid of regularization parameters defined as $\lambda_t = 0.99\lambda_{t-1}$. We also incorporated the well-known active set heuristics to conduct our experiments on larger data sets. Note that because of the

Table 3: Evaluation of the Total CPU Time (Seconds) with the Active Set Method.

Method\Data Set	phishing	SensIT	a9a	mnist	cifar10	rcv
ActiveSet	7989.5	16,352.1	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**

Note: Results in bold indicate the fastest method.

above differences, the computational time shown here cannot be directly compared with the results in sections 7.1 and 7.2. The active set method uses only a subset of triplets of which the loss is greater than 0 as the active set. The gradient is calculated by using only the active set, and the overall optimality is confirmed when the iteration converges. We employed the active set update strategy shown by Weinberger and Saul (2009), in which the active set is updated once every ten iterative cycles.

Table 3 compares the CPU time for the entire regularization path. Based on the results in the previous section, we employed RRPB and RRPB+PGB (evaluating rules based on both spheres) for triplet screening. Further, the range-based screening described in section 6.1 is also performed using RRPB, for which we evaluate the range at the beginning of the optimization for each λ, as shown in algorithm 2. Our safe triplet screening accelerates the optimization process by up to 10 times compared to the simple active set method. The results for higher-dimensional data sets with a diagonal $\boldsymbol{M}$ are presented in section 7.4.1.

7.4 Empirical Evaluation of Three Special Cases. Here we evaluate three special cases of our formulation: nonsmoothed hinge loss, the Mahalanobis distance with a diagonal matrix, and dynamic screening for a certain value of λ.

7.4.1 Nonsmoothed Hinge Loss. In previous experiments, we used the smoothed hinge loss function $\gamma = 0.05$. However, the hinge loss function $\gamma = 0$ is also widely used. Figure 10 shows the screening result of the PGB spherical rule for the segment data. Here, the loss function of RTLM is the hinge loss function, and the other settings are the same as those of the experiments in the main text. The results show that PGB achieved a high screening rate and that the CPU time substantially improved.

7.4.2 Learning with Higher-Dimensional Data Using Diagonal Matrix. Here we evaluate the screening performance when the matrix $\boldsymbol{M}$ is confined to being a diagonal matrix. Based on the same setting as section 7.3, comparison with the ActiveSet method is shown in Table 4. We used RRPB and RRPB+PGB, both of which largely reduced the CPU time. Attempts to

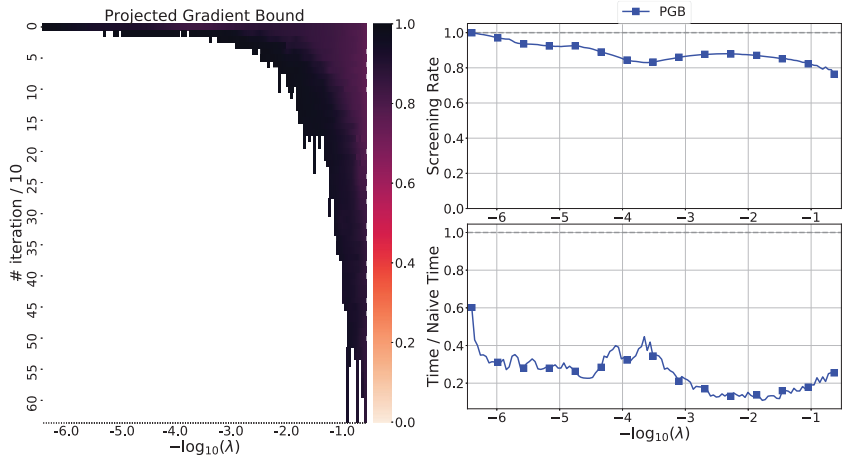


Figure 10: Performance evaluation of PGB for the hinge loss setting. The heat map on the left shows the dynamic screening rate with the vertical axis showing the number of iterative cycles for optimization divided by 10 at which screening is performed. (Top right) Rate of regularization path screening. (Bottom right) Ratio of CPU time compared with the naive optimization.

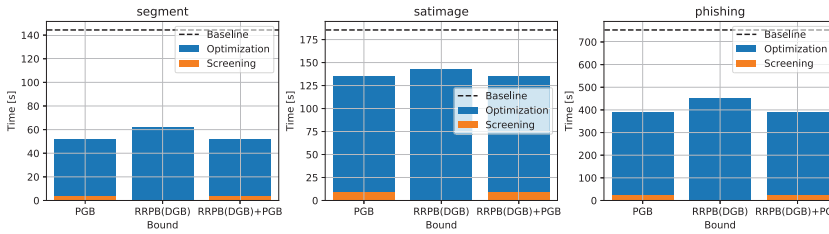
Table 4: Total Time (Seconds) of the Regularization Path for Diagonal $\boldsymbol{M}$.

Method\Data Set	USPS	Madelon	Colon-Cancer	Gisette
ActiveSet	2485.5	7005.8	3149.8	–
ActiveSet+RRPB	**326.7**	593.4	632.2	133,870.0
ActiveSet+RRPB+PGB	336.6	**562.4**	**628.2**	**127,123.8**
#dimension	256	500	2000	5000
#samples	7291	2000	62	6000
#triplet	656,200	720,400	38,696	1,215,225
k	10	20	∞	15
$\lambda_{\max}$	1.0e+7	2.0e+14	5.0e+7	4.5e+8
$\lambda_{\min}$	1.9e+3	4.7e+11	7.0e+3	2.1e+3

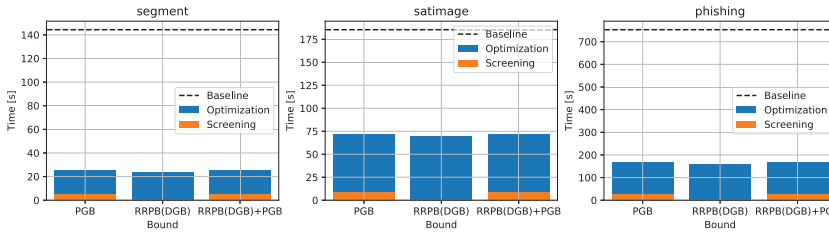
Notes: The results in bold indicate the fastest method. The Gisette data set did not produce results by ActiveSet because of the time limitation.

process the Gisette data set, which has the largest dimension, 5,000, with the active set method were unsuccessful and the method did not terminate even after 250,000 s.

7.4.3 Dynamic Screening for Fixed λ. Here, we evaluate the performance of dynamic screening for a fixed λ. For λ, we used $\lambda_{\min}$ in Table 2 for which the screening rate was relatively low in our results thus far (e.g., see



(a) Time comparison of naïve and dynamic screening



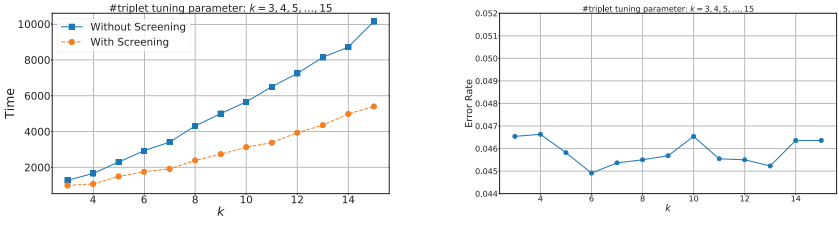
(b) Time comparison of naïve and active set + dynamic screening

Figure 11: Evaluation of the computational time for dynamic screening. The computational time required for (a) dynamic screening (a) without the active set and (b) with the active set. "Baseline" indicates the results obtained for the naive method without screening and the active set strategy. The bars show the total time and the proportion of the optimization algorithm and the screening calculation.

Figure 6a). Figure 11 compares the computational time of the naive approach without screening and with the dynamic screening shown in algorithm 1. The plots in Figure 11a show that dynamic screening accelerates the learning process. The plots in Figure 11b show the performance of the active set strategy, indicating that the combination of dynamic screening and the active set strategy is effective for further acceleration.

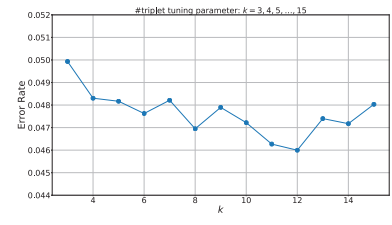
7.5 Effect of Number of Triplets on Prediction Accuracy. Finally, we examine the relation between the number of triplets contained in $\mathcal{T}$ and the prediction accuracy of the classification. We employed the nearest-neighbor (NN) classifier to measure the prediction performance of the learned metric. The data set was randomly divided into training data (60%), validation data (20%), and test data (20%). The regularization parameter λ changed from 10^5 to 0.1 and was chosen by minimizing the validation error. The experiment was performed 10 times by randomly partitioning the data set in different ways.

The results are shown in Figure 12, which summarizes the CPU time and test error rate for different settings of the number of triplets. The horizontal

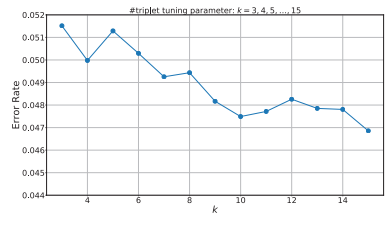


(a) Learning time.

(b) 10-NN test error rate.



(c) 20-NN test error rate.



(d) 30-NN test error rate.

Figure 12: CPU time (seconds) and test error rate on the phishing data set.

axes in all four plots, a to d, represent the number of neighbors k used to define the original triplet set $\mathcal{T}$ as described at the beginning of section 7. Figure 12a shows the CPU time to calculate the entire regularization path with and without screening. Here "Without Screening" indicates the ActiveSet approach, and "With Screening" indicates the ActiveSet+RRPB approach. These results show that the learning time increases as k increases, and safe triplet screening shows larger decreases in the CPU time for larger values of k. Figures 12b to 12d show the test error rates, each calculated by 10 NN, 20 NN, and 30 NN classifiers, respectively. In Figure 12b, the 10 NN test error is minimized at $k = 6$, with screening requiring less than approximately 2,000 seconds, whereas the naive approach (Without Screening) can calculate only approximately $k = 4$ in the same computational time. In Figure 12c, the 20 NN test error is minimized at $k = 12$, with screening requiring approximately 4000 seconds, whereas the naıve approach can calculate only approximately $k = 8$. In Figure 12d, the 30 NN test error is minimized at $k = 15$, with screening requiring approximately 5000 seconds, whereas the naïve approach can calculate only approximately $k = 9$. These results indicate that the number of neighbors, k, significantly affects the prediction accuracy, and sufficiently large k is often necessary to achieve the best prediction performance.

8 Conclusion

We introduced safe triplet screening for large-margin metric learning. Three screening rules and six spherical bounds were derived, and the relations among them were analyzed. We further proposed a range-based extension for the regularization path calculation. Our screening technique for metric learning is particularly significant compared with other screening studies because of the large number of triplets and the semidefinite constraint. Our numerical experiments verified the effectiveness of safe triplet screening using several benchmark data sets.

Appendix A: Dual Formulation

To derive the dual problem, we first rewrite the primal problem as

$$\begin{aligned} \min_{\boldsymbol{M}, \boldsymbol{t}} \quad & \sum_{ijl} \ell\left(t_{ijl}\right) + \lambda R(\boldsymbol{M}) \\ \text{s.t.} \quad & \boldsymbol{M} \succeq \boldsymbol{O},\ t_{ijl} = \langle \boldsymbol{M}, \boldsymbol{H}_{ijl} \rangle, \end{aligned}$$

where $\boldsymbol{t}$ is a $|\mathcal{T}|$-dimensional vector that contains all t_{ijl} for $(i, j, l) \in \mathcal{T}$ and

$$R(\boldsymbol{M}) = \frac{1}{2} \|\boldsymbol{M}\|_F^2 . \tag{A.1}$$

The Lagrange function is

$$L(\boldsymbol{M}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}) := \sum_{ijl} \ell\left(t_{ijl}\right) + \lambda R(\boldsymbol{M}) + \sum_{ijl} \alpha_{ijl}(t_{ijl} - \langle \boldsymbol{M}, \boldsymbol{H}_{ijl} \rangle) - \langle \boldsymbol{M}, \boldsymbol{\Gamma} \rangle,$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{T}|}$ and $\boldsymbol{\Gamma} \in \mathbb{R}_+^{d \times d}$ are Lagrange multipliers. Let

$$\ell^*(-\alpha_{ijl}) := \sup_{t_{ijl}} \{(-\alpha_{ijl}) t_{ijl} - \ell\left(t_{ijl}\right)\}, \tag{A.2}$$

$$R^*(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})) := \sup_{\boldsymbol{M}} \{\langle \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}), \boldsymbol{M} \rangle - R(\boldsymbol{M})\}, \tag{A.3}$$

be convex conjugate functions (Boyd & Vandenberghe, 2004) of ℓ and R, where

$$\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) := \frac{1}{\lambda} \left[\sum_{ijl} \alpha_{ijl} \boldsymbol{H}_{ijl} + \boldsymbol{\Gamma} \right]. \tag{A.4}$$

Then the dual function is written as

$$\begin{aligned} D_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) &:= \inf_{\boldsymbol{M}, \boldsymbol{t}} L(\boldsymbol{M}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}) \\ &= -\sum_{ijl} \sup_{t_{ijl}} \{(-\alpha_{ijl}) t_{ijl} - \ell\left(t_{ijl}\right)\} - \lambda \sup_{\boldsymbol{M}} \{\langle \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}), \boldsymbol{M}\rangle - R(\boldsymbol{M})\} \\ &= -\sum_{ijl} \ell^*(-\alpha_{ijl}) - \lambda R^*(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})). \end{aligned}$$

From the Karush-Kuhn-Tucker (KKT) condition, we obtain

$$\nabla_{\boldsymbol{M}} L = \lambda \nabla R(\boldsymbol{M}) - \lambda \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) = \boldsymbol{O}, \tag{A.5a}$$

$$\nabla_{t_{ijl}} L = \nabla \ell(t_{ijl}) + \alpha_{ijl} = 0, \tag{A.5b}$$

$$\boldsymbol{\Gamma} \succeq \boldsymbol{O},\ \boldsymbol{M} \succeq \boldsymbol{O},\ \langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle = t_{ijl},\ \langle \boldsymbol{M}, \boldsymbol{\Gamma}\rangle = 0, \tag{A.5c}$$

where, in the case of hinge loss,

$$\nabla \ell(x) = \begin{cases} 0, & x > 1, \\ -c, & x = 1, \\ -1, & x < 1, \end{cases}$$

where $\forall c \in [0, 1]$, and in the case of smoothed hinge loss,

$$\nabla \ell(x) = \begin{cases} 0, & x > 1, \\ -\frac{1}{\gamma}(1 - x), & 1 - \gamma \le x \le 1, \\ -1, & x < 1 - \gamma. \end{cases}$$

From these two equations and equation A.5b, we see that $\mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1}$. Substituting equation A.5b into equation A.2 and considering the above constraint, the conjugate of the loss function ℓ can be transformed into

$$\ell^*(-\alpha_{ijl}) = \frac{\gamma}{2}\alpha_{ijl}^2 - \alpha_{ijl}.$$

Note that this equation holds for the cases of both hinge loss (by setting $\gamma = 0$) and smoothed hinge loss ($\gamma > 0$). Substituting equation A.5a into A.3, the conjugate of the regularization term R is written as

$$R^*(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})) = R(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})) = \frac{1}{2}\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})\|_F^2.$$

Therefore, the dual problem is

$$\max_{\mathbf{0}\le\boldsymbol{\alpha}\le\mathbf{1},\,\boldsymbol{\Gamma}\succeq\boldsymbol{O}} D_\lambda(\boldsymbol{\alpha},\boldsymbol{\Gamma}) = -\sum_{ijl}\ell^*(-\alpha_{ijl}) - \frac{\lambda}{2}\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha},\boldsymbol{\Gamma})\|_F^2. \tag{Dual1}$$

Because the second term, $\max_{\boldsymbol{\Gamma}\succeq\boldsymbol{O}} -\frac{1}{2}\left\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha},\boldsymbol{\Gamma})\right\|_F^2$, is equivalent to the projection onto a semidefinite cone (Boyd & Xiao, 2005; Malick, 2004), the above problem (Dual1) can be simplified as

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

where

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

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

$$\begin{aligned}
\mathcal{L}^\star &:= \{(i,j,l)\in\mathcal{T} \mid \langle\boldsymbol{H}_{ijl},\boldsymbol{M}^\star\rangle < 1-\gamma\},\\
\mathcal{C}^\star &:= \{(i,j,l)\in\mathcal{T} \mid 1-\gamma \le \langle\boldsymbol{H}_{ijl},\boldsymbol{M}^\star\rangle \le 1\},\\
\mathcal{R}^\star &:= \{(i,j,l)\in\mathcal{T} \mid \langle\boldsymbol{H}_{ijl},\boldsymbol{M}^\star\rangle > 1\}.
\end{aligned} \tag{A.6}$$

Based on equations A.5b and A.5c in becomes clear that $\alpha^\star_{ijl} = -\nabla\ell(\langle\boldsymbol{M}^\star, \boldsymbol{H}_{ijl}\rangle)$, by which the following rules are obtained:

$$\begin{aligned}
&(i,j,l)\in\mathcal{L}^\star \Rightarrow \alpha^\star_{ijl} = 1,\\
&(i,j,l)\in\mathcal{C}^\star \Rightarrow \alpha^\star_{ijl} \in [0,1],\\
&(i,j,l)\in\mathcal{R}^\star \Rightarrow \alpha^\star_{ijl} = 0.
\end{aligned} \tag{A.7}$$

Appendix B: Proof of Lemma 1

The reduced-size problem can be represented by

$$\begin{aligned}
\min_{\boldsymbol{M},t}\quad & \sum_{(i,j,l)\in\tilde{\mathcal{T}}}\ell(t_{ijl}) + \sum_{(i,j,l)\in\hat{\mathcal{L}}}\left(1-\frac{\gamma}{2}-t_{ijl}\right) + \frac{\lambda}{2}\|\boldsymbol{M}\|_F^2\\
\text{s.t.}\quad & t_{ijl} = \langle\boldsymbol{M},\boldsymbol{H}_{ijl}\rangle \qquad (i,j,l)\in\mathcal{T},\\
& \boldsymbol{M}\succeq\boldsymbol{O}.
\end{aligned}$$

Then the Lagrangian is

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

The dual function is written as

$$\begin{aligned}\tilde{D}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) &:= \inf_{\boldsymbol{M}, \boldsymbol{t}} \tilde{L}(\boldsymbol{M}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\Gamma}) \\ &= -\sum_{(i,j,l)\in\tilde{\mathcal{T}}} \sup_{t_{ijl}}\{(-\alpha_{ijl})t_{ijl} - \ell\left(t_{ijl}\right)\} - \sum_{(i,j,l)\in\hat{\mathcal{R}}} \sup_{t_{ijl}}\{(-\alpha_{ijl})t_{ijl}\} \\ &\quad - \sum_{(i,j,l)\in\hat{\mathcal{L}}} \sup_{t_{ijl}}\{(1-\alpha_{ijl})t_{ijl}\} + (1 - \frac{\gamma}{2})|\hat{\mathcal{L}}| \\ &\quad - \lambda \sup_{\boldsymbol{M}}\{\langle \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}), \boldsymbol{M}\rangle - R(\boldsymbol{M})\},\end{aligned}$$

where $R(\boldsymbol{M})$ and and $\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})$ are defined by equations A.1 and A.4, respectively. Based on the second and third terms of the previous equation, we see

$$\alpha_{ijl} = 0, \qquad \forall (i, j, l) \in \hat{\mathcal{R}}, \tag{B.2}$$

$$\alpha_{ijl} = 1, \qquad \forall (i, j, l) \in \hat{\mathcal{L}}, \tag{B.3}$$

which prevent $\tilde{D}_\lambda$ from approaching ∞. Then constraints B.2 and B.3 enable us to further transform the dual objective into

$$\begin{aligned}\tilde{D}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) &= -\sum_{(i,j,l)\in\tilde{\mathcal{T}}} \ell^*(-\alpha_{ijl}) + \sum_{(i,j,l)\in\hat{\mathcal{L}}} \left(1 - \frac{\gamma}{2}\right) - \frac{\lambda}{2}\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})\|_F^2 \\ &= -\frac{\gamma}{2}\|\boldsymbol{\alpha}_{\tilde{\mathcal{T}}}\|_2^2 + \boldsymbol{\alpha}_{\tilde{\mathcal{T}}}^\top 1 + \left(1 - \frac{\gamma}{2}\right)|\hat{\mathcal{L}}| - \frac{\lambda}{2}\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})\|_F^2 \\ &= -\frac{\gamma}{2}\|\boldsymbol{\alpha}\|_2^2 + \boldsymbol{\alpha}^\top 1 - \frac{\lambda}{2}\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})\|_F^2.\end{aligned}$$

Thus, the dual problem is written as

$$\begin{aligned}&\max_{\mathbf{0}\le\boldsymbol{\alpha}\le\mathbf{1},\, \boldsymbol{\Gamma}\succeq O} D_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) = -\frac{\gamma}{2}\|\boldsymbol{\alpha}\|_2^2 + \boldsymbol{\alpha}^\top 1 - \frac{\lambda}{2}\|\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})\|_F^2 \\ &\text{s.t.} \quad \boldsymbol{\alpha}_{\hat{\mathcal{L}}} = \mathbf{1},\ \boldsymbol{\alpha}_{\hat{\mathcal{R}}} = \mathbf{0}.\end{aligned} \tag{B.4}$$

This is the same optimization problem as Dual1 except that $\boldsymbol{\alpha}_{\hat{\mathcal{L}}}$ and $\boldsymbol{\alpha}_{\hat{\mathcal{R}}}$ are fixed as the optimal value in Dual1. This obviously indicates that problems B.4 and Dual1 have the same optimal solution. Given the optimal dual variables $\boldsymbol{\alpha}^\star$ and $\boldsymbol{\Gamma}^\star$, the optimal primal $\boldsymbol{M}^\star$ can be derived by

$$\boldsymbol{M} = \frac{1}{\lambda}\left[\sum_{(i,j,l)\in\mathcal{T}} \alpha_{ijl}\boldsymbol{H}_{ijl} + \boldsymbol{\Gamma}\right], \tag{B.5}$$

which is from $\nabla_{\boldsymbol{M}}\tilde{L} = 0$. Because equation B.5 is exactly the same transformation as equation 2.4, the same optimal primal $\boldsymbol{M}^\star$ must be obtained. □

Appendix C: Proof of Theorem 1 (GB)

The following theorem is a well-known optimality condition for the general convex optimization problem:

Theorem 12 *(Optimality Condition of Convex Optimization, Bertsekas, 1999). In the minimization problem $\min_{\boldsymbol{x}\in\mathcal{F}} f(\boldsymbol{x})$ where the feasible region $\mathcal{F}$ and the function $f(\boldsymbol{x})$ are convex, and the necessary and sufficient condition that $\boldsymbol{x}^\star$ is the optimal solution is*

$$\exists \nabla f(\boldsymbol{x}^\star) \in \partial f(\boldsymbol{x}^\star)\left[\nabla f(\boldsymbol{x}^\star)^\top(\boldsymbol{x}^\star - \boldsymbol{x}) \le 0,\ \forall \boldsymbol{x} \in \mathcal{F}\right],$$

where $\partial f(\boldsymbol{x}^\star)$ represents the set of subgradients in $\boldsymbol{x}^\star$.

From theorem 12, the following holds for the optimal solution $\boldsymbol{M}^\star$:

$$\langle \nabla P_\lambda(\boldsymbol{M}^\star), \boldsymbol{M}^\star - \boldsymbol{M}\rangle \le 0,\ \forall \boldsymbol{M} \succeq \boldsymbol{O}. \tag{C.1}$$

Let $\boldsymbol{\Xi}_{ijl}(\boldsymbol{M})$ be the subgradient of the loss function $\ell(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle)$ at $\boldsymbol{M}$. Then $\nabla P_\lambda(\boldsymbol{M})$ is written as

$$\nabla P_\lambda(\boldsymbol{M}) = \sum_{ijl} \boldsymbol{\Xi}_{ijl}(\boldsymbol{M}) + \lambda \boldsymbol{M}. \tag{C.2}$$

From the convexity of the (smoothed) hinge loss function $\ell(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle)$, we obtain

$$\ell(\langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl}\rangle) \ge \ell(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle) + \langle \boldsymbol{\Xi}_{ijl}(\boldsymbol{M}), \boldsymbol{M}^\star - \boldsymbol{M}\rangle,$$
$$\ell(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle) \ge \ell(\langle \boldsymbol{M}^\star, \boldsymbol{H}_{ijl}\rangle) + \langle \boldsymbol{\Xi}_{ijl}(\boldsymbol{M}^\star), \boldsymbol{M} - \boldsymbol{M}^\star\rangle,$$

for any subgradient. The addition of these two equations shows that

$$\langle \boldsymbol{\Xi}_{ijl}(\boldsymbol{M}^\star), \boldsymbol{M}^\star - \boldsymbol{M}\rangle \ge \langle \boldsymbol{\Xi}_{ijl}(\boldsymbol{M}), \boldsymbol{M}^\star - \boldsymbol{M}\rangle. \tag{C.3}$$

Combining equations C.1, to C.3 results in

$$\left\langle \sum_{ijl} \Xi_{ijl}(\boldsymbol{M}) + \lambda \boldsymbol{M}^\star, \boldsymbol{M}^\star - \boldsymbol{M} \right\rangle \leq 0$$
$$\Leftrightarrow \langle \nabla P_\lambda(\boldsymbol{M}) - \lambda \boldsymbol{M} + \lambda \boldsymbol{M}^\star, \boldsymbol{M}^\star - \boldsymbol{M} \rangle \leq 0.$$

By transforming this inequality based on completing the square, we obtain GB. □

Appendix D: Proof of Theorem 2 (PGB)

Let $\boldsymbol{Q}^{\mathrm{GB}}$ be the center of the GB hypersphere and r_{GB} be the radius. The optimal solution exists in the following set:

$$\{\boldsymbol{X} \mid \|\boldsymbol{X} - \boldsymbol{Q}^{\mathrm{GB}}\|_F^2 \leq r_{\mathrm{GB}}^2,\ \boldsymbol{X} \succeq \boldsymbol{O}\}. \tag{D.1}$$

By transforming the sphere of GB, we obtain

$$\begin{aligned}\|\boldsymbol{X} - \boldsymbol{Q}^{\mathrm{GB}}\|_F^2 &= \|\boldsymbol{X} - (\boldsymbol{Q}_+^{\mathrm{GB}} + \boldsymbol{Q}_-^{\mathrm{GB}})\|_F^2 \\ &= \|\boldsymbol{X} - \boldsymbol{Q}_+^{\mathrm{GB}}\|_F^2 + 2\langle \boldsymbol{X}, -\boldsymbol{Q}_-^{\mathrm{GB}} \rangle + 2\langle \boldsymbol{Q}_+^{\mathrm{GB}}, \boldsymbol{Q}_-^{\mathrm{GB}} \rangle + \|\boldsymbol{Q}_-^{\mathrm{GB}}\|_F^2.\end{aligned}$$

Because $\boldsymbol{X} \succeq \boldsymbol{O}$ and $-\boldsymbol{Q}_-^{\mathrm{GB}} \succeq \boldsymbol{O}$, we see $\langle \boldsymbol{X}, -\boldsymbol{Q}_-^{\mathrm{GB}} \rangle \geq 0$. Furthermore, using $\langle \boldsymbol{Q}_+^{\mathrm{GB}}, \boldsymbol{Q}_-^{\mathrm{GB}} \rangle = 0$, we obtain the following sphere:

$$\begin{aligned}r_{\mathrm{GB}}^2 \geq \|\boldsymbol{X} - \boldsymbol{Q}^{\mathrm{GB}}\|_F^2 \geq \|\boldsymbol{X} - \boldsymbol{Q}_+^{\mathrm{GB}}\|_F^2 + \|\boldsymbol{Q}_-^{\mathrm{GB}}\|_F^2. \\ \therefore\ \|\boldsymbol{X} - \boldsymbol{Q}_+^{\mathrm{GB}}\|_F^2 \leq r_{\mathrm{GB}}^2 - \|\boldsymbol{Q}_-^{\mathrm{GB}}\|_F^2.\end{aligned}$$

Letting $\boldsymbol{Q}^{\mathrm{PGB}} := \boldsymbol{Q}_+^{\mathrm{GB}}$ and $r_{\mathrm{PGB}}^2 := r_{\mathrm{GB}}^2 - \|\boldsymbol{Q}_-^{\mathrm{GB}}\|_F^2$, PGB is obtained. Note that by considering $\langle \boldsymbol{X}, -\boldsymbol{Q}_-^{\mathrm{GB}} \rangle \geq 0$ instead of $\boldsymbol{X} \succeq \boldsymbol{O}$ in equation D.1, we can immediately see that GB with the linear constraint $\langle \boldsymbol{X}, -\boldsymbol{Q}_-^{\mathrm{GB}} \rangle \geq 0$ is tighter than PGB. □

Appendix E: Proof of Theorem 4 (DGB)

In general, a function $f(\boldsymbol{x})$ is an m-strongly convex function if $f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|_2^2$ is convex. Because 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, equation C.1, the second term on the right-hand side is greater than or equal to 0, and from the weak duality, $P_\lambda(\boldsymbol{M}^\star) \geq D_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})$. Therefore, we obtain theorem 4. □

Appendix F: Constrained Duality Gap Bound (CDGB)

For the DGB, we show that if the primal and dual reference solutions satisfy equation 2.4, the radius can be $\sqrt{2}$ times smaller. We extend the dual-based screening of SVM (Zimmert et al., 2015) for RTLM.

Theorem 14 *(CDGB). Let $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$ be the feasible solutions of the dual problem. Then the optimal solution of the primal problem $\boldsymbol{M}^\star$ exists in the following hypersphere:*

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

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 $\boldsymbol{\Gamma}$. The following equation is the duality gap as a function of the primal feasible solution $\boldsymbol{M}$ in which the dual solutions are optimized:

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

From the definition, we obtain

$$G_{D_\lambda}(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) \geq G_{P_\lambda}(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})). \tag{F.1}$$

From the strong convexity of G_{P_λ} shown in section F.1, the following holds for any $\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma})$ and $\boldsymbol{M}^\star \succeq \boldsymbol{O}$:

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

We assume that $\boldsymbol{M}^\star$ is the optimal solution of the primal problem. Then, because $\boldsymbol{M}^\star$ is also a solution to the convex optimization problem $\min_{\boldsymbol{M} \succeq \boldsymbol{O}} G_{P_\lambda}(\boldsymbol{M})$, it becomes clear that $\langle \nabla G_{P_\lambda}(\boldsymbol{M}^\star), \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) - \boldsymbol{M}^\star \rangle \geq 0$ from theorem 12. Considering $G_{P_\lambda}(\boldsymbol{M}^\star) = 0$ and $G_{D_\lambda}(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) \geq G_{P_\lambda}(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}))$, both of which are from the definition, we obtain

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

Dividing by λ, CDGB is derived. □

We name this bound the constrained duality gap bound (CDGB), of which the radius converges to 0 at the optimal solution, because the CDGB also has a radius proportional to the square root of the duality gap. For primal-based optimizers, additional calculation is necessary for $P_\lambda(\boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}))$, whereas dual-based optimizers calculate this term in the optimization process.

F.1 Proof of Strong Convexity of G_{P_λ}. We first define an m-strongly convex function as follows:

Definition 1 *(m-strongly Convex Function). When $f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|_2^2$ is a convex function, $f(\boldsymbol{x})$ is an m-strongly convex function.*

According to definition 1, to show that G_{P_λ} is strongly convex, we need to show that the term other than $\lambda\|\boldsymbol{M}\|_F^2$ is convex:

$$G_{P_\lambda}(\boldsymbol{M}) = \underbrace{\sum_{ijl} \ell(\langle \boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle)}_{\text{convex}} + \lambda\|\boldsymbol{M}\|_F^2 + \underbrace{\min_{\mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1},\, \boldsymbol{\Gamma} \succeq \boldsymbol{O},\, \boldsymbol{M}_\lambda(\boldsymbol{\alpha}, \boldsymbol{\Gamma}) = \boldsymbol{M}} \underbrace{\sum_{ijl} \ell^*(-\alpha_{ijl})}_{:= g(\boldsymbol{\alpha})}}_{:= f(\boldsymbol{M})}.$$

Because the loss ℓ is convex, we need to show that $f(\boldsymbol{M})$ is convex. This can be shown as

$$f(\boldsymbol{M}) = \min_{\substack{\mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1},\, \boldsymbol{\Gamma} \succeq \boldsymbol{O}, \\ \frac{1}{\lambda}\left[\sum_{ijl} \alpha_{ijl} \boldsymbol{H}_{ijl} + \boldsymbol{\Gamma}\right] = \boldsymbol{M}}} g(\boldsymbol{\alpha}) = \min_{\substack{\mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1}, \\ \frac{1}{\lambda}\sum_{ijl} \alpha_{ijl} \boldsymbol{H}_{ijl} \preceq \boldsymbol{M}}} g(\boldsymbol{\alpha}).$$

Consider a point $\boldsymbol{M}_2 = t\,\boldsymbol{M}_0 + (1-t)\,\boldsymbol{M}_1$ ($t \in [0, 1]$), which internally divides two points $\boldsymbol{M}_0$ and $\boldsymbol{M}_1$. Let

$$\boldsymbol{\alpha}_i^\star := \operatorname*{argmin}_{\substack{\mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1}, \\ \frac{1}{\lambda}\sum_{ijl} \alpha_{ijl} \boldsymbol{H}_{ijl} \preceq \boldsymbol{M}_i,}} g(\boldsymbol{\alpha}),$$

which means that $\boldsymbol{\alpha}_i^\star$ is the minimizer of this problem for a given $\boldsymbol{M}_i (i \in \{0, 1, 2\})$, and from the definition, we see $f(\boldsymbol{M}_i) = g(\boldsymbol{\alpha}_i^\star)$. Further, let $\boldsymbol{\alpha}_2 = t\,\boldsymbol{\alpha}_0^\star + (1-t)\,\boldsymbol{\alpha}_1^\star$. Then, $\mathbf{0} \le \boldsymbol{\alpha}_2 \le \mathbf{1}$ and $\frac{1}{\lambda}\sum_{ijl} \alpha_{2,ijl} \boldsymbol{H}_{ijl} \preceq \boldsymbol{M}_2$. Because g is convex because of the convexity of ℓ^*, we have

$$t\,f(\boldsymbol{M}_0) + (1-t)\,f(\boldsymbol{M}_1) = t\,g(\boldsymbol{\alpha}_0^\star) + (1-t)\,g(\boldsymbol{\alpha}_1^\star)$$

$$\geq g(\underbrace{t\boldsymbol{\alpha}_0^\star + (1-t)\boldsymbol{\alpha}_1^\star}_{\boldsymbol{\alpha}_2})$$
$$\geq g(\boldsymbol{\alpha}_2^\star) = f(\underbrace{t\,\boldsymbol{M}_0 + (1-t)\,\boldsymbol{M}_1}_{\boldsymbol{M}_2}).$$

Hence, $f(\boldsymbol{M})$ is convex and G_{P_λ} is a strongly convex function.

Appendix G: Proof of Theorem 5 (RPB)

The optimality condition, theorem 12, in the dual problem, Dual1, for λ_0, λ_1 determines that

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

By adding these two equations, we obtain

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

Next, we consider the following difference of gradient:

$$\nabla_{\boldsymbol{\alpha}_{ijl}} D_{\lambda_0}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star) - \nabla_{\boldsymbol{\alpha}_{ijl}} D_{\lambda_1}(\boldsymbol{\alpha}_1^\star, \boldsymbol{\Gamma}_1^\star) = -\gamma(\alpha_{0ijl}^\star - \alpha_{1ijl}^\star) - \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0^\star - \boldsymbol{M}_1^\star\rangle,$$
$$\nabla_{\boldsymbol{\Gamma}} D_{\lambda_0}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star) - \nabla_{\boldsymbol{\Gamma}} D_{\lambda_1}(\boldsymbol{\alpha}_1^\star, \boldsymbol{\Gamma}_1^\star) = -(\boldsymbol{M}_0^\star - \boldsymbol{M}_1^\star).$$

Defining $\boldsymbol{H}_t^\star := \sum_{ijl} \alpha_{t\,ijl}^\star \boldsymbol{H}_{ijl}$, $\boldsymbol{M}_t^\star$ is rewritten as $\boldsymbol{M}_t^\star = \frac{1}{\lambda_t}[\boldsymbol{H}_t^\star + \boldsymbol{\Gamma}_t^\star]$. Then

$$\gamma\|\boldsymbol{\alpha}_1^\star - \boldsymbol{\alpha}_0^\star\|_2^2 - \langle \boldsymbol{H}_1^\star - \boldsymbol{H}_0^\star, \boldsymbol{M}_0^\star - \boldsymbol{M}_1^\star\rangle - \langle \boldsymbol{M}_0^\star - \boldsymbol{M}_1^\star, \boldsymbol{\Gamma}_1^\star - \boldsymbol{\Gamma}_0^\star\rangle \leq 0$$
$$\Leftrightarrow \gamma\|\boldsymbol{\alpha}_1^\star - \boldsymbol{\alpha}_0^\star\|_2^2 - \langle \lambda_1\boldsymbol{M}_1^\star - \lambda_0\boldsymbol{M}_0^\star, \boldsymbol{M}_0^\star - \boldsymbol{M}_1^\star\rangle \leq 0$$
$$\Rightarrow -\langle \lambda_1\boldsymbol{M}_1^\star - \lambda_0\boldsymbol{M}_0^\star, \boldsymbol{M}_0^\star - \boldsymbol{M}_1^\star\rangle \leq 0.$$

Transformation of this inequality based on completing the square allows the RPB to be obtained. □

Appendix H: Proof of Theorem 6 (RRPB)

Considering a hypersphere that expands the RPB radius by $\frac{\lambda_0+\lambda_1}{2\lambda_1}\epsilon$ and replaces the RPB center with $\frac{\lambda_0+\lambda_1}{2\lambda_1}\boldsymbol{M}_0$, we obtain

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

Because ϵ is defined by $\|M_0^\star - M_0\|_F \leq \epsilon$, this sphere covers any RPB created by $M_0^\star$, which satisfies $\|M_0^\star - M_0\|_F \leq \epsilon$ (see Figure 2d for a geometrical illustration). Using the reverse triangle inequality,

$$\|M_0^\star\|_F - \|M_0\|_F \leq \|M_0^\star - M_0\|_F \leq \epsilon,$$

we obtain

$$\left\| M_1^\star - \frac{\lambda_0 + \lambda_1}{2\lambda_1} M_0 \right\|_F \leq \frac{|\lambda_0 - \lambda_1|}{2\lambda_1}(\|M_0\|_F + \epsilon) + \frac{\lambda_0 + \lambda_1}{2\lambda_1}\epsilon.$$

By rearranging this, RRPB is obtained. □

Appendix I: Proof of Theorem 7 (Relationship between PGB and RPB)

When the dual variable is used as the subgradient of the (smoothed) hinge loss at the optimal solution $M_0^\star$ of λ_0 (from equation A.7, the optimal dual variable provides a valid subgradient), the gradient of the objective function in the case of λ_1 is written as

$$\nabla P_{\lambda_1}(M_0^\star) = -H_0^\star + \lambda_1 M_0^\star,$$

where

$$H_0^\star := -\sum_{ijl} \nabla \ell(\langle M_0^\star, H_{ijl}\rangle) H_{ijl} = \sum_{ijl} \alpha_{0ijl}^\star H_{ijl}.$$

Because $\lambda_0 M_0^\star = H_{0+}^\star$,

$$\begin{aligned}\nabla P_{\lambda_1}(M_0^\star) &= -(H_{0+}^\star + H_{0-}^\star) + \lambda_1 M_0^\star \\ &= (\lambda_1 - \lambda_0) M_0^\star - H_{0-}^\star.\end{aligned}$$

Then the center and radius of GB are

$$Q^{\mathrm{GB}} = M_0^\star - \frac{1}{2\lambda_1}\nabla P_{\lambda_1}(M_0^\star) = \frac{(\lambda_0 + \lambda_1) M_0^\star + H_{0-}^\star}{2\lambda_1},$$

$$\begin{aligned} r_{\mathrm{GB}}^2 &= \frac{\left\|(\lambda_1 - \lambda_0) M_0^\star - H_{0-}^\star\right\|_F^2}{4\lambda_1^2} \\ &= \frac{\left\|(\lambda_1 - \lambda_0) M_0^\star\right\|_F^2 - 2(\lambda_1 - \lambda_0)\langle M_0^\star, H_{0-}^\star\rangle + \left\|H_{0-}^\star\right\|_F^2}{4\lambda_1^2}\end{aligned}$$

$$= \frac{\left\|(\lambda_0 - \lambda_1)\boldsymbol{M}_0^\star\right\|_F^2 + \left\|\boldsymbol{H}_{0-}^\star\right\|_F^2}{4\lambda_1^2}.$$

Here, the last equation of r_{GB}^2 uses the fact that $\boldsymbol{M}_0^*$ and $\boldsymbol{H}_{0-}^*$ are orthogonal. Using $\boldsymbol{Q}^{\mathrm{GB}}$ and r_{GB}^2, the center and radius of PGB are found to be

$$\boldsymbol{Q}^{\mathrm{PGB}} = \boldsymbol{Q}_+^{\mathrm{GB}} = \frac{(\lambda_0 + \lambda_1)\boldsymbol{M}_0^\star}{2\lambda_1},\ \boldsymbol{Q}_-^{\mathrm{GB}} = \frac{\boldsymbol{H}_{0-}^\star}{2\lambda_1},$$

$$r_{\mathrm{PGB}}^2 = r_{\mathrm{GB}}^2 - \left\|\boldsymbol{Q}_-^{\mathrm{GB}}\right\|_F^2 = \frac{\left\|(\lambda_0 - \lambda_1)\boldsymbol{M}_0^\star\right\|_F^2}{4\lambda_1^2}.$$

Therefore, PGB coincides with RPB. □

Appendix J: Proof of Theorem 8 (Relationship between DGB and RPB)

At the optimal solution $\boldsymbol{M}_0^\star$, $\boldsymbol{\alpha}_0^\star$ and $\boldsymbol{\Gamma}_0^\star$ of λ_0, we obtain the following equation from $P_{\lambda_0}(\boldsymbol{M}_0^\star) = D_{\lambda_0}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star)$ and $\boldsymbol{M}_{\lambda_0}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star) = \boldsymbol{M}_0^\star$:

$$\sum_{ijl} \ell(\langle \boldsymbol{M}_0^\star, \boldsymbol{H}_{ijl}\rangle) + \sum_{ijl} \ell^*(-\alpha_{0ijl}^\star) = -\lambda_0 \left\|\boldsymbol{M}_0^\star\right\|_F^2.$$

We also see $\boldsymbol{M}_{\lambda_1}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star) = \frac{\lambda_0}{\lambda_1}\boldsymbol{M}_{\lambda_0}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star) = \frac{\lambda_0}{\lambda_1}\boldsymbol{M}_0^\star$. Using these results, the value of the duality gap for λ_1 is

$$P_{\lambda_1}(\boldsymbol{M}_0^\star) - D_{\lambda_1}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star) = \frac{(\lambda_0 - \lambda_1)^2}{2\lambda_1}\left\|\boldsymbol{M}_0^\star\right\|_F^2.$$

Therefore, the radius of DGB r_{DGB} and the radius of RPB r_{RPB} satisfy the following relationship:

$$r_{\mathrm{DGB}}^2 = \frac{2(P_{\lambda_1}(\boldsymbol{M}_0^\star) - D_{\lambda_1}(\boldsymbol{\alpha}_0^\star, \boldsymbol{\Gamma}_0^\star))}{\lambda_1}$$

$$= \frac{(\lambda_0 - \lambda_1)^2}{\lambda_1^2}\left\|\boldsymbol{M}_0^\star\right\|_F^2 = 4\, r_{\mathrm{RPB}}^2.$$

Furthermore, the centers of these hyperspheres are

$$\boldsymbol{Q}^{\mathrm{DGB}} = \boldsymbol{M}_0^\star,\ \boldsymbol{Q}^{\mathrm{RPB}} = \frac{\lambda_0 + \lambda_1}{2\lambda_1}\boldsymbol{M}_0^\star,$$

and the distance between the centers is

$$\|Q^{\mathrm{DGB}} - Q^{\mathrm{RPB}}\|_F = \frac{|\lambda_0 - \lambda_1|}{2\lambda_1}\|M_0^\star\|_F = r_{\mathrm{RPB}}.$$

Thus, the DGB includes the RPB as illustrated in Figure 2c. □

Appendix K: Proof of Theorem 9

The Lagrange function is defined as

$$L(X, \alpha, \beta) := \langle X, H_{ijl}\rangle - \alpha\frac{1}{2}(r^2 - \|X - Q\|_F^2) - \beta\langle P, X\rangle.$$

From the KKT condition, we obtain

$$\partial L/\partial X = H_{ijl} + \alpha(X - Q) - \beta P = O. \tag{K.1a}$$

$$\alpha \geq 0,\ \beta \geq 0,\ \|X - Q\|_F^2 \leq r^2,\ \langle P, X\rangle \geq 0. \tag{K.1b}$$

$$\alpha(r^2 - \|X - Q\|_F^2) = 0,\ \beta\langle P, X\rangle = 0. \tag{K.1c}$$

If $\alpha = 0$, then $H_{ijl} = \beta P$ from equation K.1a, and the value of the objective function becomes $\langle X, H_{ijl}\rangle = \beta\langle X, P\rangle = 0$ from equation K.1c. Let us consider the case of $\alpha \neq 0$. From equation K.1c, it becomes clear that $\|X - Q\|_F^2 = r^2$. If $\beta = 0$, the linear constraint is not an active constraint (i.e., $\langle P, X\rangle > 0$ at the optimal); hence, it is the same as problem P1, which can be analytically solved. If this solution satisfies the linear constraint $\langle P, X\rangle \geq 0$, it becomes the optimal solution. Next, we consider the case of $\beta \neq 0$. From equations K.1a and K.1c, α and β are obtained as

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

Of the solutions of the two values of $\alpha, \alpha > 0$ gives the minimum value from equation K.1b. □

Appendix L: Range-Based Extension

L.1 Generalized Form of GB, DGB, RPB, and RRPB.

L.1.1 GB. The gradient is written as

$$\nabla P_\lambda(M) = \Xi + \lambda M.$$

Then, the squared norm of this gradient is

$$\|\nabla P_\lambda(\boldsymbol{M})\|_F^2 = \|\boldsymbol{\Xi}\|_F^2 + 2\lambda\langle\boldsymbol{\Xi}, \boldsymbol{M}\rangle + \lambda^2\|\boldsymbol{M}\|_F^2.$$

By substituting this into the center and the radius of GB, we obtain

$$\begin{aligned} r_{\mathrm{GB}}^2 &= \frac{1}{4\lambda^2}\|\nabla P_\lambda(\boldsymbol{M})\|_F^2 \\ &= \frac{1}{4\lambda^2}\left(\|\boldsymbol{\Xi}\|_F^2 + 2\lambda\langle\boldsymbol{\Xi}, \boldsymbol{M}\rangle + \lambda^2\|\boldsymbol{M}\|_F^2\right) \\ &= \frac{1}{4}\|\boldsymbol{M}\|_F^2 + \frac{1}{2\lambda}\langle\boldsymbol{\Xi}, \boldsymbol{M}\rangle\frac{1}{4\lambda^2}\|\boldsymbol{\Xi}\|_F^2, \\ \boldsymbol{Q}^{\mathrm{GB}} &= \boldsymbol{M} - \frac{1}{2\lambda}(\boldsymbol{\Xi} + \lambda\boldsymbol{M}) \\ &= \frac{1}{2}\boldsymbol{M} - \frac{1}{2\lambda}\boldsymbol{\Xi}. \end{aligned}$$

L.1.2 DGB. The duality gap is written as

$$\mathrm{gap} = \sum_{ijl}(\ell(\langle\boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle) + \ell^*(-\alpha_{ijl})) + \frac{\lambda}{2}\|\boldsymbol{M}\|_F^2 + \frac{1}{2\lambda}\|\sum_{ijl}\alpha_{ijl}\boldsymbol{H}_{ijl} + \boldsymbol{\Gamma}\|_F^2.$$

Then the center and radius of DGB are

$$\begin{aligned} r_{\mathrm{DGB}}^2 &= \frac{2\mathrm{gap}}{\lambda} \\ &= \frac{2}{\lambda}\left(\sum_{ijl}(\ell(\langle\boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle) + \ell^*(-\alpha_{ijl})) + \frac{\lambda}{2}\|\boldsymbol{M}\|_F^2\right. \\ &\quad \left. + \frac{1}{2\lambda}\left\|\sum_{ijl}\alpha_{ijl}\boldsymbol{H}_{ijl} + \boldsymbol{\Gamma}\right\|_F^2\right) \\ &= \|\boldsymbol{M}\|_F^2 + \frac{2}{\lambda}\left(\sum_{ijl}(\ell(\langle\boldsymbol{M}, \boldsymbol{H}_{ijl}\rangle) + \ell^*(-\alpha_{ijl}))\right) \\ &\quad + \frac{1}{\lambda^2}\left\|\sum_{ijl}\alpha_{ijl}\boldsymbol{H}_{ijl} + \boldsymbol{\Gamma}\right\|_F^2, \\ \boldsymbol{Q}^{\mathrm{DGB}} &= \boldsymbol{M}. \end{aligned}$$

L.1.3 RPB. With respect to RPB, we regard λ_1 as the target λ for which we consider the range. From the definition, we see

$$\begin{aligned} \boldsymbol{Q}^{\mathrm{RPB}} &= \frac{\lambda_0+\lambda}{2\lambda}\boldsymbol{M}_0^\star \\ &= \frac{1}{2}\boldsymbol{M}_0^\star + \frac{\lambda_0}{2\lambda}\boldsymbol{M}_0^\star, \\ r_{\mathrm{RPB}} &= \frac{\lambda_0-\lambda}{2\lambda}\left\|\boldsymbol{M}_0^\star\right\|_F \\ &= -\frac{1}{2}\left\|\boldsymbol{M}_0^\star\right\|_F + \frac{\lambda_0}{2\lambda}\left\|\boldsymbol{M}_0^\star\right\|_F. \end{aligned}$$

L.1.4 RRPB. Here again, we regard λ_1 as the target λ for which we consider the range. First, we assume $\lambda \le \lambda_0$. Then we have

$$\begin{aligned} \boldsymbol{Q}^{\mathrm{RRPB}} &= \frac{\lambda_0+\lambda}{2\lambda}\boldsymbol{M}_0 \\ &= \frac{1}{2}\boldsymbol{M}_0 + \frac{\lambda_0}{2\lambda}\boldsymbol{M}_0, \\ r_{\mathrm{RRPB}} &= \frac{\lambda_0-\lambda}{2\lambda}\|\boldsymbol{M}_0\|_F + \frac{\lambda_0}{\lambda}\epsilon \\ &= -\frac{1}{2}\|\boldsymbol{M}_0\|_F + \frac{1}{\lambda}\left(\frac{\lambda_0}{2}\|\boldsymbol{M}_0\|_F + \lambda_0\epsilon\right). \end{aligned}$$

In the case of $\lambda \ge \lambda_0$, we have

$$\begin{aligned} \boldsymbol{Q}^{\mathrm{RRPB}} &= \frac{1}{2}\boldsymbol{M}_0 + \frac{\lambda_0}{2\lambda}\boldsymbol{M}_0, \\ r_{\mathrm{RRPB}} &= \frac{\lambda-\lambda_0}{2\lambda}\|\boldsymbol{M}_0\|_F + \epsilon \\ &= \left(\epsilon + \frac{1}{2}\|\boldsymbol{M}_0\|_F\right) - \frac{\lambda_0}{2\lambda}\|\boldsymbol{M}_0\|_F. \end{aligned}$$

L.2 Proof of Theorem 10 (Range-Based Extension of RRPB). In RRPB, we replace λ_1 with λ and assume $\lambda \le \lambda_0$. Then,

$$\boldsymbol{Q}^{\mathrm{RRPB}} = \frac{\lambda_0+\lambda}{2\lambda}\boldsymbol{M}_0,\ r_{\mathrm{RRPB}} = \frac{\lambda_0-\lambda}{2\lambda}\|\boldsymbol{M}_0\|_F + \frac{\lambda_0}{\lambda}\epsilon.$$

From the spherical rule, equation 4.1, we obtain

$$\frac{\lambda+\lambda_0}{2\lambda}\langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0\rangle - \left(\frac{\lambda_0-\lambda}{2\lambda}\|\boldsymbol{M}_0\|_F + \frac{\lambda_0}{\lambda}\epsilon\right)\|\boldsymbol{H}_{ijl}\|_F > 1$$

$$\Leftrightarrow \underbrace{(\langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle - 2 + \|\boldsymbol{H}_{ijl}\|_F \|\boldsymbol{M}_0\|_F)}_{>0 \text{ is required.}} \lambda$$
$$> \lambda_0 (\underbrace{\|\boldsymbol{M}_0\|_F \|\boldsymbol{H}_{ijl}\|_F - \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle}_{\geq 0} + 2\epsilon \|\boldsymbol{H}_{ijl}\|_F).$$

The Cauchy-Schwarz inequality determines that the right-hand side is equal to or greater than 0; therefore, the left-hand side must be greater than 0:

$$\therefore \ \lambda_0 \geq \lambda > \frac{\lambda_0 (\|\boldsymbol{M}_0\|_F \|\boldsymbol{H}_{ijl}\|_F - \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle + 2\epsilon \|\boldsymbol{H}_{ijl}\|_F)}{\langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle - 2 + \|\boldsymbol{H}_{ijl}\|_F \|\boldsymbol{M}_0\|_F}.$$

In the case of $\lambda \geq \lambda_0$,

$$Q^{\mathrm{RRPB}} = \frac{\lambda_0 + \lambda}{2\lambda} \boldsymbol{M}_0, \ r_{\mathrm{RRPB}} = \frac{\lambda - \lambda_0}{2\lambda} \|\boldsymbol{M}_0\|_F + \epsilon.$$

From spherical rule 4.1,

$$\frac{\lambda + \lambda_0}{2\lambda} \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle - \left(\frac{\lambda - \lambda_0}{2\lambda} \|\boldsymbol{M}_0\|_F + \epsilon \right) \|\boldsymbol{H}_{ijl}\|_F > 1$$
$$\Leftrightarrow (\underbrace{\|\boldsymbol{H}_{ijl}\|_F \|\boldsymbol{M}_0\|_F - \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle}_{\geq 0} + 2 + 2\epsilon \|\boldsymbol{H}_{ijl}\|_F) \lambda$$
$$< \lambda_0 (\|\boldsymbol{M}_0\|_F \|\boldsymbol{H}_{ijl}\|_F + \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle).$$

Similarly, the Cauchy-Schwarz inequality determines that the left-hand side is greater than 0:

$$\therefore \ \lambda_0 \leq \lambda < \frac{\lambda_0 (\|\boldsymbol{M}_0\|_F \|\boldsymbol{H}_{ijl}\|_F + \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle)}{\|\boldsymbol{H}_{ijl}\|_F \|\boldsymbol{M}_0\|_F - \langle \boldsymbol{H}_{ijl}, \boldsymbol{M}_0 \rangle + 2 + 2\epsilon \|\boldsymbol{H}_{ijl}\|_F}.$$

□

Appendix M: Calculation for Range-Based Extension Other Than RPB and RRPB

The spherical rule is written as

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

This inequality is equivalent to the following two inequalities:

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

By using equations 6.1 and 6.3, these inequalities can be transformed into

$$\left(\left\langle \boldsymbol{H}_{ijl}, \boldsymbol{A} + \boldsymbol{B}\frac{1}{\lambda}\right\rangle - 1\right)^2 > \left(\sqrt{a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2}} + \sqrt{\frac{2\epsilon}{\lambda}}\right)^2 \|\boldsymbol{H}_{ijl}\|_F^2, \quad \text{(M.1)}$$

$$\left\langle \boldsymbol{H}_{ijl}, \boldsymbol{A} + \boldsymbol{B}\frac{1}{\lambda}\right\rangle > 1. \quad \text{(M.2)}$$

Note that the definitions of a, b, c, $\boldsymbol{A}$, and $\boldsymbol{B}$ for each bound are shown in section L.1. Because inequality M.2 can be written as a linear inequality of λ, we can easily obtain the range of λ that satisfies the inequality. On the other hand, inequality M.1 is equivalent to

$$\begin{aligned}
&\left(\left\langle \boldsymbol{H}_{ijl}, \boldsymbol{A} + \boldsymbol{B}\frac{1}{\lambda}\right\rangle - 1\right)^2 \\
&\quad > \left(a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2} + \frac{2\epsilon}{\lambda} + 2\sqrt{a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2}}\sqrt{\frac{2\epsilon}{\lambda}}\right)\|\boldsymbol{H}_{ijl}\|_F^2 \\
\Leftrightarrow &\left(\left\langle \boldsymbol{H}_{ijl}, \boldsymbol{A} + \boldsymbol{B}\frac{1}{\lambda}\right\rangle - 1\right)^2 - \left(a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2} + \frac{2\epsilon}{\lambda}\right)\|\boldsymbol{H}_{ijl}\|_F^2 \\
&\quad > \left(2\sqrt{a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2}}\sqrt{\frac{2\epsilon}{\lambda}}\right)\|\boldsymbol{H}_{ijl}\|_F^2.
\end{aligned}$$

The last inequality can be transformed into the following two inequalities:

$$\begin{aligned}
&\left(\left(\left\langle \boldsymbol{H}_{ijl}, \boldsymbol{A} + \boldsymbol{B}\frac{1}{\lambda}\right\rangle - 1\right)^2 - \left(a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2} + \frac{2\epsilon}{\lambda}\right)\|\boldsymbol{H}_{ijl}\|_F^2\right)^2 \\
&\quad > 4\left(a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2}\right)\left(\frac{2\epsilon}{\lambda}\right)\|\boldsymbol{H}_{ijl}\|_F^4, \quad \text{(M.3)}
\end{aligned}$$

$$\left(\left\langle \boldsymbol{H}_{ijl}, \boldsymbol{A} + \boldsymbol{B}\frac{1}{\lambda}\right\rangle - 1\right)^2 > \left(a + b\frac{1}{\lambda} + c\frac{1}{\lambda^2} + \frac{2\epsilon}{\lambda}\right)\|\boldsymbol{H}_{ijl}\|_F^2. \quad \text{(M.4)}$$

Inequality M.4 is also a quadratic inequality for which we can obtain the range of λ that satisfies the inequality. Although inequality M.3 is a fourth-order inequality, the range of λ can be calculated by using a fourth-order equation solver. Then we obtain the range of λ as the intersection of the ranges derived from equations M.2 to M.4.

Appendix N: Spherical Rule with Semidefinite Constraint for Diagonal Case

N.1 Proof of Theorem 11. Rearranging equation 6.4a, we obtain

$$\beta_k = 2\alpha x_k + (h_{ijl,k} - 2\alpha q_k).$$

When we assume $h_{ijl,k} - 2\alpha q_k > 0$, we see

$$\begin{aligned} & h_{ijl,k} - 2\alpha q_k > 0 \\ \Rightarrow\ & \beta_k = \underbrace{2\alpha x_k}_{\geq 0} + \underbrace{(h_{ijl,k} - 2\alpha q_k)}_{>0} > 0 \\ \Rightarrow\ & x_k = 0. \end{aligned}$$

The previous equation is derived from the complementary condition $\beta_k x_k = 0$. When we assume $h_{ijl,k} - 2\alpha q_k \leq 0$, we have

$$\begin{aligned} & h_{ijl,k} - 2\alpha q_k \leq 0 \\ \Rightarrow\ & 2\alpha x_k - \beta_k = -(h_{ijl,k} - 2\alpha q_k) \geq 0 \\ \Rightarrow\ & \beta_k = 0 \\ \Rightarrow\ & x_k = q_k - h_{ijl,k}/2\alpha. \end{aligned}$$

The third equation, $\beta_k = 0$, is derived from $x_k \geq 0$, $\beta_k \geq 0$ and the complementary condition $\beta_k x_k = 0$, and in the previous equation, the assumption $\alpha > 0$ is used. Using the above two derivations, we obtain equation 6.5. Further, from the complementary condition $\alpha(r^2 - \|\boldsymbol{x} - \boldsymbol{q}\|_2^2) = 0$, it is clear that $\|\boldsymbol{x} - \boldsymbol{q}\|_2^2 = r^2$ because of the assumption $\alpha > 0$. □

N.2 Proof of Theorem 12. Because we assume that $\alpha = 0$, we obtain

$$\boldsymbol{\beta} = \boldsymbol{h}_{ijl}$$

by using the KKT condition, equation 6.4a. Note that this implicitly indicates that $\boldsymbol{h}_{ijl} \geq 0$ should be satisfied because of the nonnegativity of $\boldsymbol{\beta}$. The complementary condition $x_k \beta_k = 0$ determines that

$$x_k = 0 \ \text{ if } h_{ijl,k} > 0. \tag{N.1}$$

To satisfy all the KKT conditions, equation 6.4, we need to set the other x_k in such a way that $\|\boldsymbol{x} - \boldsymbol{q}\|_2^2 \leq r^2$ and $\boldsymbol{x} \geq 0$ are satisfied. Note that the other conditions in equation 6.4 are satisfied for any $\boldsymbol{x}$ because of the assumption $\alpha = 0$. By setting

$$x_k = \max\{q_k, 0\} \text{ for } h_{ijl,k} = 0, \tag{N.2}$$

Algorithm 3: Analytical Procedure for Calculating the Optimal Solution of Equation P.

if the KKT conditions (6.4) holds for the solution with $\alpha = 0$ defined by equations 6.8 and 6.9 **then**

Return

end if

$\alpha_0 \leftarrow 0$

Calculate the set of change points $\alpha_1, \ldots, \alpha_{d'}$ by sorting the set (N.3) ascendingly

for $k = 0$ to $d' - 1$ **do**

Create the set $\mathcal{S}_k$ for the interval (α_k, α_{k+1})

Calculate α which satisfies the equality (6.7) by assuming the support of $\boldsymbol{x}$ is $\mathcal{S}_k$

if the KKT conditions (6.4) hold for the solution defined by equations 6.5 and 6.6 **then**

Return

end if

end for

$\|\boldsymbol{x} - \boldsymbol{q}\|_2^2$ is minimized under the conditions $\boldsymbol{x} \geq 0$ and equation N.1, and thus the condition $\|\boldsymbol{x} - \boldsymbol{q}\|_2^2 \leq r^2$ should be satisfied when the optimal α is 0.

N.3 Analytical Procedure of Rule Evaluation for the Diagonal Case. We first verify the case $\alpha = 0$. If the solution equations 6.8 and 6.9 satisfy all the KKT conditions, equation 6.4, then the solution is optimal. Otherwise, we consider the case $\alpha > 0$. Let $\mathcal{S} := \{k \mid x_k > 0\}$ be the support set of $\boldsymbol{x}$, where x_k is defined by equation 6.5. When $\mathcal{S}$ is regarded as a function of α, an element of $\mathcal{S}$ can change at which α satisfies

$$h_{ijl,k} - 2\alpha q_k = 0$$

for some $k \in [d]$. Let $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{d'}$ for $d' \leq d$ be a sequence of those change points that can be found by sorting

$$\left\{ \frac{h_{ijl,k}}{2q_k} \;\middle|\; \frac{h_{ijl,k}}{2q_k} > 0, q_k \neq 0, k \in [d] \right\}. \tag{N.3}$$

For notational convenience, we define $\alpha_0 := 0$. Based on the definition, $\mathcal{S}$ is fixed for any α in an interval (α_k, α_{k+1}), to which we refer as $\mathcal{S}_k$. This means

that the support set of the optimal $\boldsymbol{x}$ should be one of $\mathcal{S}_k$ for $k = 1, \ldots, d'$. Algorithm 3 shows an analytical procedure for calculating the optimal $\boldsymbol{x}$, which verifies the optimality of each one of $\mathcal{S}_k$ after considering the case of $\alpha = 0$. For each iterative cycle in algorithm 3, $O(d)$ computation is required, and thus the solution can be found by $O(d^2)$.

Acknowledgments

This work was financially supported by grants from the Japanese Ministry of Education, Culture, Sports, Science and Technology awarded to I.T. (16H06538, 17H00758) and M.K. (16H06538, 17H04694); from the Japan Science and Technology Agency (JST) CREST awarded to I.T. (JPMJCR1302, JPMJCR1502) and PRESTO awarded to M.K. (JPMJPR15N2); from the Materials Research by Information Integration Initiative (MI2I) project of the Support Program for Starting Up Innovation Hub from the JST awarded to I.T. and M.K.; and from the RIKEN Center for Advanced Intelligence Project awarded to I.T.

References

Barzilai, J., & Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, *8*(1), 141–148.

Bertsekas, D. P. (1999). *Nonlinear programming*. Belmont: Athena Scientific.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Boyd, S., & Xiao, L. (2005). Least-squares covariance matrix adjustment. *SIAM Journal on Matrix Analysis and Applications*, *27*(2), 532–546.

Capitaine, H. L. (2016). *Constraint selection in metric learning*. arXiv:1612.04853.

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27.

Chollet, F., et al. (2015). *Keras*. https://github.com/keras-team/keras.

Davis, J. V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 209–216). New York: ACM.

Fercoq, O., Gramfort, A., & Salmon, J. (2015). Mind the duality gap: Safer rules for the lasso. In *Proceedings of the 32nd International Conference on Machine Learning*, (pp. 333–342).

Ghaoui, L. E., Viallon, V., & Rabbani, T. (2010). *Safe feature elimination for the lasso and sparse supervised learning problems*. arXiv:1009.4219.

Hanada, H., Shibagaki, A., Sakuma, J., & Takeuchi, I. (2018). Efficiently monitoring small data modification effect for large-scale learning in changing environment. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (pp. 1314–1321). Palo Alto, CA: AAAI Press.

Hoffer, E., & Ailon, N. (2015). Deep metric learning using triplet network. In *Proceedings of the International Workshop on Similarity-Based Pattern Recognition* (pp. 84–92). Berlin: Springer.

Jain, L., Mason, B., & Nowak, R. (2017). Learning low-dimensional metrics. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*, *30* (pp. 4139–4147). Red Hook, NY: Curran.

Jain, P., Kulis, B., Dhillon, I. S., & Grauman, K. (2009). Online metric learning and fast similarity search. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems*, *21* (pp. 761–768). Cambridge, MA: MIT Press.

Jamieson, K. G., & Nowak, R. D. (2011). Low-dimensional embedding using adaptively selected ordinal data. In *Proceedings of the 2011 49th Annual Allerton Conference on Communication, Control, and Computing* (pp. 1077–1084). Piscataway, NJ: IEEE.

Kulis, B. (2013). *Metric learning: A survey*. Boston: Now Publishers.

Law, M. T., Thome, N., & Cord, M. (2013). Quadruplet-wise image similarity learning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 249–256). Piscataway, NJ: IEEE.

Lee, S., & Xing, E. P. (2014). *Screening rules for overlapping group lasso*. arXiv:1410.6880.

Lehoucq, R. B., & Sorensen, D. C. (1996). Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, *17*(4), 789–821.

Li, D., & Tian, Y. (2018). Survey and experimental study on metric learning methods. *Neural Networks*, *105*, 447–462.

Liu, J., Zhao, Z., Wang, J., & Ye, J. (2014). Safe screening with variational inequalities and its application to lasso. In *Proceedings of the International Conference on Machine Learning* (pp. 289–297).

Malick, J. (2004). A dual approach to semidefinite least-squares problems. *SIAM Journal on Matrix Analysis and Applications*, *26*(1), 272–284.

McFee, B., & Lanckriet, G. R. (2010). Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 775–782). Madison, WI: Omnipress.

Nakagawa, K., Suzumura, S., Karasuyama, M., Tsuda, K., & Takeuchi, I. (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* (pp. 1785–1794). New York: ACM.

Ndiaye, E., Fercoq, O., Gramfort, A., & Salmon, J. (2016). Gap safe screening rules for sparse-group lasso. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*, *29* (pp. 388–396). Red Hook, NY: Curran.

Ogawa, K., Suzuki, Y., & Takeuchi, I. (2013). Safe screening of non-support vectors in pathwise SVM computation. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 1382–1390).

Okumura, S., Suzuki, Y., & Takeuchi, I. (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* (pp. 885–894). New York: ACM.

Perrot, M., & Habrard, A. (2015). Regressive virtual metric learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 1810–1818). Red Hook, NY: Curran.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823). Piscataway, NJ: IEEE.

Schultz, M., & Joachims, T. (2004). Learning a distance metric from relative comparisons. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems*, *16* (pp. 41–48). Cambridge, MA: MIT Press.

Shen, C., Kim, J., Liu, F., Wang, L., & Van Den Hengel, A. (2014). Efficient dual approach to distance metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(2), 394–406.

Shi, Y., Bellet, A., & Sha, F. (2014). Sparse compositional metric learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

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

Shibagaki, A., Suzuki, Y., Karasuyama, M., & Takeuchi, I. (2015). Regularization path of cross-validation error lower bounds. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems, 28* (pp. 1675–1683). Red Hook, NY: Curran.

Takada, T., Hanada, H., Yamada, Y., Sakuma, J., & Takeuchi, I. (2016). Secure approximation guarantee for cryptographically private empirical risk minimization. In *Proceedings of the 8th Asian Conference on Machine Learning* (pp. 126–141).

Wang, J., Wonka, P., & Ye, J. (2014). Scaling SVM and least absolute deviations via exact data reduction. In *Proceedings of the International Conference on Machine Learning* (pp. 523–531).

Wang, J., Zhou, J., Wonka, P., & Ye, J. (2013). Lasso screening rules via dual polytope projection. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *26* (pp. 1070–1078). Red Hook, NY: Curran.

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

Xiang, Z. J., Wang, Y., & Ramadge, P. J. (2017). Screening tests for lasso problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(5), 1008–1027.

Xing, E. P., Jordan, M. I., Russell, S. J., & Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems*, *15* (pp. 521–528). Cambridge, MA: MIT Press.

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

Zhang, W., Hong, B., Liu, W., Ye, J., Cai, D., He, X., & Wang, J. (2016). *Scaling up sparse support vector machines by simultaneous feature and sample reduction*. arXiv:1607.06996.

Zhou, Q., & Zhao, Q. (2015). Safe subspace screening for nuclear norm regularized least squares problems. In *Proceedings of the International Conference on Machine Learning* (pp. 1103–1112).

Zimmert, J., de Witt, C. S., Kerg, G., & Kloft, M. (2015). Safe screening for support vector machines. In *NIPS 2015 workshop on optimization in machine learning*. Red Hook, NY: Curran.

Received November 14, 2018; accepted July 29, 2019.