GRAPH METRIC LEARNING VIA GERSHGORIN DISC ALIGNMENT

Cheng Yang†, *Gene Cheung*†, *Wei Hu*‡

†Department of Electrical Engineering & Computer Science, York University, Toronto, Canada
‡Wangxuan Institute of Computer Technology, Peking University, Beijing, China

ABSTRACT

We propose a general projection-free metric learning framework, where the minimization objective $\min_{\mathbf{M}\in\mathcal{S}} Q(\mathbf{M})$ is a convex differentiable function of the metric matrix $\mathbf{M}$, and $\mathbf{M}$ resides in the set $\mathcal{S}$ of generalized graph Laplacian matrices for connected graphs with positive edge weights and node degrees. Unlike low-rank metric matrices common in the literature, $\mathcal{S}$ includes the important positive-diagonal-only matrices as a special case in the limit. The key idea for fast optimization is to rewrite the positive definite cone constraint in $\mathcal{S}$ as signal-adaptive linear constraints via Gershgorin disc alignment, so that the alternating optimization of the diagonal and off-diagonal terms in $\mathbf{M}$ can be solved efficiently as linear programs via Frank-Wolfe iterations. We prove that the Gershgorin discs can be aligned perfectly using the first eigenvector $\mathbf{v}$ of $\mathbf{M}$, which we update iteratively using Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG) with warm start as diagonal / off-diagonal terms are optimized. Experiments show that our efficiently computed graph metric matrices outperform metrics learned using competing methods in terms of classification tasks.

Index Terms— Metric Learning, graph signal processing

1. INTRODUCTION

Given a feature vector $\mathbf{f}_i \in \mathbb{R}^K$ per sample i, a *metric* matrix $\mathbf{M} \in \mathbb{R}^{K\times K}$ defines the *feature distance*—*Mahalanobis distance* [1] between two samples i and j in a feature space as $(\mathbf{f}_i-\mathbf{f}_j)^\top\mathbf{M}(\mathbf{f}_i-\mathbf{f}_j)$, where $\mathbf{M}$ is commonly assumed to be positive definite (PD). *Metric learning*—identifying the best metric $\mathbf{M}$ minimizing a chosen objective function $Q(\mathbf{M})$ subject to $\mathbf{M} \succ 0$—has been the focus of many recent machine learning research efforts [2, 3, 4, 5, 6].

One key challenge in metric learning is to satisfy the positive (semi-)definite (PSD) cone constraint $\mathbf{M} \succ 0$ ($\mathbf{M} \succeq 0$) when minimizing $Q(\mathbf{M})$ in a computation-efficient manner. A standard approach is iterative gradient-descent / projection (*e.g.*, *proximal gradient* (PG) [7]), where a descent step $\boldsymbol{\alpha}$ from current solution $\mathbf{M}^t$ at iteration t in the direction of the negative gradient $-\nabla Q(\mathbf{M}^t)$ is followed by a projection $\text{Pr}()$ back to the PSD cone, *i.e.*, $\mathbf{M}^{t+1} := \text{Pr}\left(\mathbf{M}^t - \boldsymbol{\alpha}\nabla Q(\mathbf{M}^t)\right)$. However, projection $\text{Pr}()$ typically requires eigen-decomposition of $\mathbf{M}$ and soft-thresholding of its eigenvalues, which is computation-expensive.

Recent methods consider alternative search spaces of matrices such as sparse or low-rank matrices to ease optimization [3, 4, 5, 8, 9]. While efficient, the assumed restricted search spaces often degrade the quality of sought metric $\mathbf{M}$ in defining the Mahalanobis distance. For example, low-rank methods explicitly assume reducibility of the K available features to a lower dimension, and hence exclude the simple yet important weighted feature metric case where $\mathbf{M}$ contains only positive diagonal entries [10], *i.e.*, $(\mathbf{f}_i-\mathbf{f}_j)^\top\mathbf{M}(\mathbf{f}_i-\mathbf{f}_j) = \sum_k m_{k,k}(f_i^k - f_j^k)^2, m_{k,k} > 0, \forall k$. We show in our experiments that computed metrics by these methods may result in inferior performance for selected applications.

In this paper, we propose a metric learning framework that is both general and projection-free, capable of optimizing any convex differentiable objective $Q(\mathbf{M})$. Compared to low-rank methods, our framework is more encompassing and includes positive-diagonal metric matrices as a special case in the limit[1]. The main idea is as follows. First, we define a search space $\mathcal{S}$ of *general graph Laplacian* matrices [11], each corresponding to a connected graph with positive edge weights and node degrees. The underlying graph edge weights capture pairwise correlations among the K features, and the self-loops designate relative importance among the features.

Assuming $\mathbf{M} \in \mathcal{S}$, we next rewrite the PD cone constraint as signal-adaptive linear constraints via *Gershgorin disc alignment* [12, 13]: first compute scalars s_k's from previous solution $\mathbf{M}^t$ that align the Gershgorin disc left-ends of matrix $\mathbf{S}\mathbf{M}^t\mathbf{S}^{-1}$, where $\mathbf{S} = \text{diag}(s_1, \ldots, s_K)$, then derive scaled linear constraints using s_k's to ensure PDness of the next computed metric $\mathbf{M}^{t+1}$ via the Gershgorin Circle Theorem (GCT) [14]. Linear constraints mean that our proposed alternating optimization of the diagonal and off-diagonal terms in $\mathbf{M}$ can be solved speedily as linear programs [15] via Frank-Wolfe iterations [16]. We prove that for any metric $\mathbf{M}^t$ in $\mathcal{S}$, using scalars $s_k = 1/v_k$ can perfectly align Gershgorin disc left-ends for matrix $\mathbf{S}\mathbf{M}^t\mathbf{S}^{-1}$ at the smallest eigenvalue $\lambda_{\min}$, where $\mathbf{M}^t\mathbf{v} = \lambda_{\min}\mathbf{v}$. We efficiently update $\mathbf{v}$ iteratively using *Locally Optimal Block Preconditioned Conjugate Gradient* (LOBPCG) [17] with warm start as diagonal / off-diagonal terms are optimized. Experiments show that our computed graph metrics outperform metrics learned using competing methods in a range of applications.

2. REVIEW OF SPECTRAL GRAPH THEORY

We consider an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{W}\}$ composed of a node set $\mathcal{V}$ of cardinality $|\mathcal{V}| = N$, an edge set $\mathcal{E}$ connecting nodes, and a weighted adjacency matrix $\mathbf{W}$. Each edge $(i, j) \in \mathcal{E}$ has a positive weight $w_{i,j} > 0$ which reflects the degree of similarity between nodes i and j. Specifically, it is common to compute edge weight $w_{i,j}$ as the exponential of the feature distance $\delta_{i,j}$ between nodes i and j [18]:

$$w_{i,j} = \exp\left(-\delta_{i,j}\right) \tag{1}$$

Using (1) means $w_{i,j} \in (0, 1]$ for $\delta_{i,j} \in [0, \infty)$. We discuss feature distance $\delta_{i,j}$ in the next section.

There may be *self-loops* in graph $\mathcal{G}$, *i.e.*, $\exists i$ where $w_{i,i} > 0$, and the corresponding diagonal entries of $\mathbf{W}$ are positive. The *combinatorial graph Laplacian* [18] is defined as $\mathbf{L} := \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is

[1]As the inter-feature correlations tend to zero, only graph self-loops expressing relative importance among the K features remain, and the general graph Laplacian matrix tends to diagonal.

the *degree matrix*—a diagonal matrix where $d_{i,i} = \sum_{j=1}^{N} w_{i,j}$. A *generalized graph Laplacian* [11] accounts for self-loops in $\mathcal{G}$ also and is defined as $\mathbf{L}_g = \mathbf{D} - \mathbf{W} + \text{diag}(\mathbf{W})$, where $\text{diag}(\mathbf{W})$ extracts the diagonal entries of $\mathbf{W}$. Alternatively we can write $\mathbf{L}_g = \mathbf{D}_g - \mathbf{W}$, where the *generalized degree matrix* $\mathbf{D}_g = \mathbf{D} + \text{diag}(\mathbf{W})$ is diagonal.

3. GRAPH METRIC LEARNING

3.1. Graph Metric Matrices

We first define the search space of metric matrices for our optimization framework. We assume that associated with each sample i is a length-K feature vector $\mathbf{f}_i \in \mathbb{R}^K$. A metric matrix $\mathbf{M} \in \mathbb{R}^{K \times K}$ defines the feature distance $\delta_{i,j}(\mathbf{M})$—the *Mahalanobis distance* [1]—between samples i and j as:

$$\delta_{i,j}(\mathbf{M}) = (\mathbf{f}_i - \mathbf{f}_j)^\top \mathbf{M} (\mathbf{f}_i - \mathbf{f}_j) \tag{2}$$

We require $\mathbf{M}$ to be a *positive definite* (PD) matrix[2]. The special case where $\mathbf{M}$ is diagonal with strictly positive entries was studied in [10]. Instead, we study here a more general case: $\mathbf{M}$ must be a *graph metric matrix*, which we define formally as follows.

Definition 1. *A PD symmetric matrix* $\mathbf{M}$ *is a graph metric if it is a generalized graph Laplacian matrix with positive edge weights and node degrees for an irreducible graph.*

For a generalized graph Laplacian $\mathbf{L}_g$ to have positive degrees, each node i may have a self-loop, but its loop weight $w_{i,i}$ must satisfy $w_{i,i} > -\sum_{j \mid j \neq i} w_{i,j}$. Irreducible graph [20] essentially means that any graph node can *commute* with any other node.

3.2. Problem Formulation

Denote by $\mathcal{S}$ the set of all graph metric matrices. We pose an optimization problem for $\mathbf{M}$: find the optimal graph metric $\mathbf{M}$ in $\mathcal{S}$—leading to inter-sample distances $\delta_{i,j}(\mathbf{M})$ in (2)—that yields the smallest value of a convex differential objective $Q(\{\delta_{i,j}(\mathbf{M})\})$:

$$\min_{\mathbf{M} \in \mathcal{S}} Q\left(\{\delta_{i,j}(\mathbf{M})\}\right), \quad \text{s.t. } \text{tr}(\mathbf{M}) \leq C \tag{3}$$

where C is a chosen parameter. Constraint $\text{tr}(\mathbf{M}) \leq C$ is needed to avoid pathological solutions with infinite feature distances, *i.e.*, $\delta_{i,i}(\mathbf{M}) = \infty$. For stability, we assume also that the objective is lower-bounded, *i.e.*, $\min_{\mathbf{M} \in \mathcal{S}} Q(\{\delta_{i,j}(\mathbf{M})\}) \geq \kappa > -\infty$ for some constant κ.

Our strategy to solve (3) is to optimize $\mathbf{M}$'s diagonal and off-diagonal terms alternately using Frank-Wolfe iterations [16], where each iteration is solved as an LP until convergence. We discuss first the initialization of $\mathbf{M}$, then the two optimizations in order. For notation convenience, we will write the objective simply as $Q(\mathbf{M})$, with the understanding that metric $\mathbf{M}$ affects first the feature distances $\delta_{i,j}(\mathbf{M})$, which in turn determine the objective $Q(\{\delta_{i,j}(\mathbf{M})\})$.

3.3. Initialization of $\mathbf{M}$

We first initialize a valid graph metric $\mathbf{M}^0$ as follows:

1. Initialize each diagonal term $m^0_{i,i} := C/K$.
2. Initialize off-diagonal terms $m^0_{i,j}$, $i \neq j$, as:

$$m^0_{i,j} := \begin{cases} -\epsilon & \text{if } j = i \pm 1 \\ 0 & \text{o.w.} \end{cases} \tag{4}$$

where $\epsilon > 0$ is a parameter. Initialization of the diagonal terms ensures that constraints $\text{tr}(\mathbf{M}^0) \leq C$, $\mathbf{M}^0 \succ 0$ and $m^0_{i,i} > 0$ are satisfied. Initialization of the off-diagonal terms ensures that $\mathbf{M}^0$ is symmetric and irreducible, and constraint $m^0_{i,j} \leq 0$, $i \neq j$, is satisfied; *i.e.*, $\mathbf{M}^0$ is a Laplacian matrix for graph with non-negative edge weights. We can hence conclude that initial $\mathbf{M}^0$ is a graph metric, *i.e.*, $\mathbf{M}^0 \in \mathcal{S}$.

3.4. Optimization of Diagonal Terms

When optimizing $\mathbf{M}$'s diagonal terms $m_{i,i}$, (3) becomes

$$\begin{aligned} \min_{\{m_{i,i}\}} \quad & Q(\mathbf{M}) \\ \text{s.t.} \quad & \mathbf{M} \succ 0; \quad \sum_i m_{i,i} \leq C; \quad m_{i,i} > 0, \forall i \end{aligned} \tag{5}$$

where $\text{tr}(\mathbf{M}) = \sum_i m_{i,i}$. Because the diagonal terms do not affect the irreducibility of matrix $\mathbf{M}$, the only requirements for $\mathbf{M}$ to be a graph metric are: i) $\mathbf{M}$ must be PD, and ii) diagonals must be strictly positive.

3.4.1. *Gershgorin-based Reformulation*

To efficiently enforce the PD constraint $\mathbf{M} \succ 0$, we derive sufficient (but not necessary) linear constraints using the *Gershgorin Circle Theorem* (GCT) [14]. By GCT, each eigenvalue λ of a real matrix $\mathbf{M}$ resides in at least one *Gershgorin disc* Ψ_i, corresponding to row i of $\mathbf{M}$, with center $c_i = m_{i,i}$ and radius $r_i = \sum_{j \mid j \neq i} |m_{i,j}|$, *i.e.*,

$$\exists i \text{ s.t. } c_i - r_i \leq \lambda \leq c_i + r_i \tag{6}$$

Thus a sufficient condition to ensure $\mathbf{M}$ is PD (smallest eigenvalue $\lambda_{\min} > 0$) is to ensure that all discs' left-ends are strictly positive, *i.e.*,

$$0 < \min_i c_i - r_i \leq \lambda_{\min} \tag{7}$$

This translates to a linear constraint for each row i:

$$m_{i,i} \geq \sum_{j \mid j \neq i} |m_{i,j}| + \rho, \qquad \forall i \in \{1, \ldots, K\} \tag{8}$$

where $\rho > 0$ is a sufficiently small parameter.

However, GCT lower bound $\min_i c_i - r_i$ for $\lambda_{\min}$ is often loose. When optimizing $\mathbf{M}$'s diagonal terms, enforcing (8) directly means that we are searching for $\{m_{i,i}\}$ in a smaller space than the original space $\{\mathbf{M} \mid \mathbf{M} \succ 0\}$ in (5), resulting in an inferior solution. As an illustration, consider the following example matrix $\mathbf{M}$:

$$\mathbf{M} = \begin{bmatrix} 2 & -2 & -1 \\ -2 & 5 & -2 \\ -1 & -2 & 4 \end{bmatrix} \tag{9}$$

Gershgorin disc left-ends $m_{i,i} - \sum_{j \mid j \neq i} |m_{i,j}|$ for this matrix are $\{-1, 1, 1\}$, of which -1 is the smallest. Thus the diagonal terms $\{2, 5, 4\}$ do not meet constraints (8). However, $\mathbf{M}$ is PD, since its smallest eigenvalue is $\lambda_{\min} = 0.1078 > 0$.

[2] By definition of a metric [19], $(\mathbf{f}_i - \mathbf{f}_j)^\top \mathbf{M} (\mathbf{f}_i - \mathbf{f}_j) > 0$ if $\mathbf{f}_i - \mathbf{f}_j \neq \mathbf{0}$.

3.4.2. Gershgorin Disc Alignment

To derive more appropriate linear constraints—thus more suitable search space when solving $\min_{\mathbf{M}\in\mathcal{S}} \mathbf{Q}(\mathbf{M})$, we examine instead the Gershgorin discs of a similar-transformed matrix $\mathbf{B}$ from $\mathbf{M}$, *i.e.*,

$$\mathbf{B} = \mathbf{S}\mathbf{M}\mathbf{S}^{-1} \tag{10}$$

where $\mathbf{S} = \text{diag}(s_1, \ldots, s_K)$ is a diagonal matrix with scalars $s_1, \ldots, s_K$ along its diagonal, $s_k > 0, \forall k$. $\mathbf{B}$ has the same eigenvalues as $\mathbf{M}$, and thus the smallest Gershgorin disc left-end, $\min_i b_{i,i} - \sum_{j \mid j\neq i} |b_{i,j}|$, for $\mathbf{B}$ is also a lower bound for $\mathbf{M}$'s smallest eigenvalue $\lambda_{\min}$. *Our goal is then to derive tight $\lambda_{\min}$ lower bounds by adapting to good solutions to* (5)*—by appropriately choosing $s_1, \ldots, s_K$ used to define $\mathbf{B}$ in* (10).

Specifically, given scalars $s_1, \ldots, s_K$, a disc Ψ_i for $\mathbf{B}$ has center $m_{i,i}$ and radius $s_i \sum_{j \mid j\neq i} |m_{i,j}|/s_j$. Thus to ensure $\mathbf{B}$ is PD (and hence $\mathbf{M}$ is PD), we can write similar linear constraints as (8):

$$m_{i,i} \geq s_i \sum_{j \mid j\neq i} \frac{|m_{i,j}|}{s_j} + \rho, \quad \forall i \in \{1, \ldots, K\} \tag{11}$$

It turns out that given a graph metric $\mathbf{M}$, there exist scalars $s_1, \ldots, s_K$ such that all disc left-ends are aligned at the same value $\lambda_{\min}$. We state this formally as a theorem.

Theorem 1. *Let $\mathbf{M}$ be a graph metric matrix. There exist strictly positive scalars $s_1, \ldots, s_K$ such that all Gershgorin disc left-ends of $\mathbf{B} = \mathbf{S}\mathbf{M}\mathbf{S}^{-1}$ are aligned exactly at the smallest eigenvalue, i.e., $b_{i,i} - \sum_{j \mid j\neq i} |b_{i,j}| = \lambda_{\min}, \forall i$.*

In other words, for matrix $\mathbf{B}$ the Gershgorin lower bound $\min_i c_i - r_i$ is *exactly* $\lambda_{\min}$, and the bound is the tightest possible. The important corollary is the following:

Corollary 1. *For any graph metric $\mathbf{M}$, which by definition is PD, there exist scalars $s_1, \ldots, s_K$ where $\mathbf{M}$ is feasible using linear constraints in* (11).

Proof. By Theorem 1, let $s_1, \ldots, s_K$ be scalars such that all Gershgorin disc left-ends of $\mathbf{B} = \mathbf{S}\mathbf{M}\mathbf{S}^{-1}$ align at $\lambda_{\min}$. Thus

$$\forall i, \;\; m_{i,i} - s_i \sum_{j \mid j\neq i} \frac{|m_{i,j}|}{s_j} = \lambda_{\min} > 0 \tag{12}$$

where $\lambda_{\min} > 0$ since $\mathbf{M}$ is PD. Hence $\mathbf{M}$ must also satisfy (11) for all i for sufficiently small $\rho > 0$. □

Continuing our earlier example, using $s_1 = 0.7511$, $s_2 = 0.4886$ and $s_3 0.4440$, we see that $\mathbf{B} = \mathbf{S}\mathbf{M}\mathbf{S}^{-1}$ for $\mathbf{M}$ in (9) has all disc left-ends aligned at $\lambda_{\min} = 0.1078$. Hence using these scalars and constraints (11), diagonal terms $\{2, 5, 4\}$ now constitute a feasible solution.

To prove Theorem 1, we first establish the following lemma.

Lemma 1. *There exists a first eigenvector $\mathbf{v}$ with strictly positive entries for a graph metric matrix $\mathbf{M}$.*

Proof. By definition, graph metric matrix $\mathbf{M}$ is a generalized graph Laplacian $\mathbf{L}_g = \mathbf{D}_g - \mathbf{W}$ with positive edge weights in $\mathbf{W}$ and positive degrees in $\mathbf{D}_g$. Let $\mathbf{v}$ be the first eigenvector of $\mathbf{M}$, *i.e.*,

$$\begin{aligned}
\mathbf{M}\mathbf{v} &= \lambda_{\min}\mathbf{v} \\
(\mathbf{D}_g - \mathbf{W})\mathbf{v} &= (\lambda_{\min}\mathbf{I})\mathbf{v} \\
\mathbf{D}_g\mathbf{v} &= (\mathbf{W} + \lambda_{\min}\mathbf{I})\mathbf{v} \\
\mathbf{v} &= \mathbf{D}_g^{-1}(\mathbf{W} + \lambda_{\min}\mathbf{I})\mathbf{v}
\end{aligned}$$

where $\lambda_{\min} > 0$ since $\mathbf{M}$ is PD. Since the matrix on the right contains only non-negative entries and $\mathbf{W}$ is an irreducible matrix, $\mathbf{v}$ is a positive eigenvector by the Perron-Frobenius Theorem [21]. □

We now prove Theorem 1 as follows.

Proof. Denote by $\mathbf{v}$ a strictly positive eigenvector corresponding to graph metric matrix $\mathbf{M}$'s smallest eigenvalue $\lambda_{\min}$. Define $\mathbf{S} = \text{diag}(1/v_1, \ldots, 1/v_K)$. Then,

$$\mathbf{S}\mathbf{M}\mathbf{S}^{-1}\mathbf{S}\mathbf{v} = \lambda_{\min}\mathbf{S}\mathbf{v} \tag{13}$$

where $\mathbf{S}\mathbf{v} = \mathbf{1} = [1, \ldots, 1]^\top$. Let $\mathbf{B} = \mathbf{S}\mathbf{M}\mathbf{S}^{-1}$. Then,

$$\mathbf{B}\mathbf{1} = \lambda_{\min}\mathbf{1} \tag{14}$$

(14) means that

$$b_{i,i} + \sum_{j \mid j\neq i} b_{i,j} = \lambda_{\min}, \quad \forall i$$

Note that the off-diagonal terms $b_{i,j} = (v_i/v_j)m_{i,j} \leq 0$, since i) $\mathbf{v}$ is strictly positive and ii) off-diagonal terms of graph metric $\mathbf{M}$ satisfy $m_{i,j} \leq 0$. Thus,

$$b_{i,i} - \sum_{j \mid j\neq i} |b_{i,j}| = \lambda_{\min}, \quad \forall i \tag{15}$$

Thus defining $\mathbf{S} = \text{diag}(1/v_1, \ldots, 1/v_K)$ means $\mathbf{B} = \mathbf{S}\mathbf{M}\mathbf{S}^{-1}$ has all its Gershgorin disc left-ends aligned at $\lambda_{\min}$. □

Thus, using a positive first eigenvector $\mathbf{v}$ of a graph metric $\mathbf{M}$, one can compute corresponding scalars $s_k = 1/v_k$ to align all disc left-ends of $\mathbf{B} = \mathbf{S}\mathbf{M}\mathbf{S}^{-1}$ at $\lambda_{\min}$, and $\mathbf{M}$ satisfies (11) by Corollary 1. Note that these scalars are *signal-adaptive*, *i.e.*, s_k's depend on $\mathbf{v}$, which is computed from $\mathbf{M}$. Our strategy then is to derive scalars s_k^t's from a good solution $\mathbf{M}^{t-1}$, optimize for a better solution $\mathbf{M}^t$ using scaled Gershgorin linear constraints (11), derive new scalars again until convergence. Specifically,

1. Given scalars s_k^t's, identify a good solution $\mathbf{M}^t$ minimizing objective $Q(\mathbf{M})$ subject to (11), *i.e.*,

$$\min_{\{m_{i,i}\}} Q(\mathbf{M}) \tag{16}$$
$$\text{s.t. } m_{i,i} \geq s_i \sum_{j \mid j\neq i} \frac{|m_{i,j}|}{s_j} + \rho, \forall i; \quad \sum_i m_{i,i} \leq C$$

2. Given $\mathbf{M}^t$, update scalars $s_k^{t+1} = 1/v_k^t$ where $\mathbf{v}^t$ is the first eigenvector of $\mathbf{M}^t$.
3. Increment t and repeat until convergence.

When the scalars in (16) are updated as $s_k^{t+1} = 1/v_k^t$ for iteration $t+1$, we show that previous solution $\mathbf{M}^t$ at iteration t remains feasible at iteration $t+1$:

Lemma 2. *Solution $\mathbf{M}^t$ to* (16) *in iteration t remains feasible in iteration $t+1$, when scalars s_i^{t+1} for the linear constraints in* (16) *are updated as $s_i^{t+1} = 1/v_i^t, \forall i$, where $\mathbf{v}^t$ is the first eigenvector of $\mathbf{M}^t$.*

Proof. Using the first eigenvector $\mathbf{v}^t$ of graph metric $\mathbf{M}^t$ at iteration t, by the proof of Theorem 1 we know that the Gershgorin disc left-ends of $\mathbf{B} = \mathbf{S}\mathbf{M}^t\mathbf{S}^{-1}$ are aligned at $\lambda_{\min}$. Since $\mathbf{M}^t$ is a feasible solution in (16), $\mathbf{M}^t \succ 0$ and $\lambda_{\min} > 0$. Thus $\mathbf{M}^t$ is also a feasible solution when scalars are updated as $s_i = 1/v_i^t, \forall i$. □

The remaining issue is how to best compute first eigenvector $\mathbf{v}^t$ given solution $\mathbf{M}^t$ repeatedly. For this task, we employ *Locally Optimal Block Preconditioned Conjugate Gradient* (LOBPCG) [17], a state-of-the-art iterative algorithm known to compute extreme eigenpairs efficiently. Further, using previously computed eigenvector $\mathbf{v}^{t-1}$ as an initial guess, LOBPCG benefits from warm start when computing $\mathbf{v}^t$, reducing its complexity in subsequent iterations [17].

3.4.3. Frank-Wolfe Algorithm

To solve (16), we employ the Frank-Wolfe algorithm [16] that iteratively linearizes the objective $Q(\mathbf{M})$ using its gradient $\nabla Q(\mathbf{M}^t)$ with respect to diagonal terms $\{m_{i,i}\}$, computed using previous solution $\mathbf{M}^t$, *i.e.*,

$$\nabla Q(\mathbf{M}^t) = \left.\begin{bmatrix} \frac{\partial Q(\mathbf{M})}{\partial m_{1,1}} \\ \vdots \\ \frac{\partial Q(\mathbf{M})}{\partial m_{K,K}} \end{bmatrix}\right|_{\mathbf{M}^t} \tag{17}$$

Given gradient $\nabla Q(\mathbf{M}^t)$, optimization (16) becomes a *linear program* (LP) at each iteration t:

$$\min_{\{m_{i,i}\}} \text{vec}(\{m_{i,i}\})^\top \nabla Q(\mathbf{M}^t) \tag{18}$$

$$\text{s.t. } m_{i,i} \geq s_i \sum_{j \mid j \neq i} \frac{|m^t_{i,j}|}{s_j} + \rho, \;\; \forall i; \quad \sum_i m_{i,i} \leq C.$$

where $\text{vec}(\{m_{i,i}\}) = [m_{1,1} \; m_{2,2} \; \ldots \; m_{K,K}]^\top$ is a vector composed of diagonal terms $\{m_{i,i}\}$, and $m^t_{i,j}$ are off-diagonal terms of previous solution $\mathbf{M}^t$. LP (18) can be solved efficiently using known fast algorithms such as Simplex [15] and interior point method [22]. When a new solution $\{m^{t+1}_{i,i}\}$ is obtained, gradient $\nabla Q(\mathbf{M}^{t+1})$ is updated, and LP (18) is solved again until convergence.

3.5. Optimization of Off-diagonal Entries

For off-diagonal entries of $\mathbf{M}$, we design a block coordinate descent algorithm, which optimizes one row / column at a time.

3.5.1. Block Coordinate Iteration

First, we divide $\mathbf{M}$ into four sub-matrices:

$$\mathbf{M} = \begin{bmatrix} m_{1,1} & \mathbf{M}_{1,2} \\ \mathbf{M}_{2,1} & \mathbf{M}_{2,2} \end{bmatrix}, \tag{19}$$

where $m_{1,1} \in \mathbb{R}$, $\mathbf{M}_{1,2} \in \mathbb{R}^{1\times(K-1)}$, $\mathbf{M}_{2,1} \in \mathbb{R}^{(K-1)\times 1}$ and $\mathbf{M}_{2,2} \in \mathbb{R}^{(K-1)\times(K-1)}$. Assuming $\mathbf{M}$ is symmetric, $\mathbf{M}_{1,2} = \mathbf{M}_{2,1}^\top$. We optimize $\mathbf{M}_{2,1}$ in one iteration, *i.e.*,

$$\min_{\mathbf{M}_{2,1}} Q(\mathbf{M}), \quad \text{s.t. } \mathbf{M} \in \mathcal{S} \tag{20}$$

In the next iteration, a different row / column i is selected, and with appropriate row / column permutation, we still optimize the first column off-diagonal terms $\mathbf{M}_{2,1}$ as in (20).

Note that the constraint $\text{tr}(\mathbf{M}) \leq C$ in (3) can be ignored, since it does not involved optimization variable $\mathbf{M}_{2,1}$. For $\mathbf{M}$ to remain in the set $\mathcal{S}$ of graph metric matrices, i) $\mathbf{M}$ must be PD, ii) $\mathbf{M}$ must be irreducible, and iii) $\mathbf{M}_{2,1} \leq \mathbf{0}$.

As done for the diagonal terms optimization, we replace the PD constraint with Gershgorin-based linear constraints. To ensure irreducibility (*i.e.*, the graph remains connected), we ensure that *at least one* off-diagonal term (say index s) in column 1 has magnitude at least $\epsilon > 0$. The optimization thus becomes:

$$\min_{\mathbf{M}_{2,1}} Q(\mathbf{M}) \tag{21}$$

$$\text{s.t. } m_{i,i} \geq s_i \sum_{j \mid j \neq i} \frac{|m_{i,j}|}{s_j} + \rho, \;\; \forall i$$

$$m_{s,1} \leq -\epsilon; \quad \mathbf{M}_{2,1} \leq \mathbf{0}$$

Essentially any selection of s in (21) can ensure $\mathbf{M}$ is irreducible. To encourage solution convergence, we select s as the index of the previously optimized $\mathbf{M}^t_{2,1}$ with the largest magnitude.

(21) also has a convex differentiable objective with a set of linear constraints. We thus employ the Frank-Wolfe algorithm again to iteratively linearize the objective using gradient $\nabla Q(\mathbf{M}^t)$ with respect to off-diagonal $\mathbf{M}_{2,1}$, where the solution in each iteration is solved as an LP. We omit the details for brevity.

4. EXPERIMENTS

We evaluate our proposed metric learning method by classification performance. Specifically, the objective function $Q(\mathbf{M})$ we consider here is the *graph Laplacian Regularizer* (GLR) [18, 23]:

$$Q(\mathbf{M}) = \mathbf{z}^\top \mathbf{L}(\mathbf{M})\mathbf{z} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(z_i - z_j)^2$$

$$= \exp\left\{-(\mathbf{f}_i - \mathbf{f}_j)^\top \mathbf{M}(\mathbf{f}_i - \mathbf{f}_j)\right\}(z_i - z_j)^2 \tag{22}$$

A small GLR means that signal $\mathbf{z}$ at connected node pairs (z_i, z_j) are similar for a large edge weight $w_{i,j}$, *i.e.z* is *smooth* w.r.t. the variation operator $\mathbf{L}(\mathbf{M})$. GLR has been used in the GSP literature to solve a range of inverse problems, including image denoising [23] and deblurring [24].

We evaluate our method with the following competing schemes: three metric learning methods that only learn the diagonals of $\mathbf{M}$, *i.e.*, [25], [26], and [10], and two methods that learn the full matrix $\mathbf{M}$, *i.e.*, [6] and [27]. We do this by performing classificaiton tasks via the following two classifiers: 1) a k-nearest-neighbour classifier, and 2) a graph-based classifier with quadratic formulation $\min_{\mathbf{z}} \mathbf{z}^\top \mathbf{L}(\mathbf{M})\mathbf{z}$ s.t. $z_i = \hat{z}_i, i \in \mathcal{F}, \mathcal{F} \subset \{1, \ldots, J\}$, where $\hat{z}_i$ in subset $\mathcal{F}$ are the observed labels. We evaluate all classifiers on `wine` (3 classes, 13 features and 178 samples), `iris` (3 classes, 4 features and 150 samples), `seeds` (3 classes, 7 features and 210 samples), and `pb` (2 classes, 10 features and 300 samples). All experiments were performed in Matlab R2017a on an i5-7500, 8GB of RAM, Windows 10 PC. We perform 2-fold cross validation 50 times using 50 random seeds (0 to 49) with one-against-all classification strategy. As shown in Tables 1, our proposed metric learning method has the lowest classification error rates with a graph-based classifier.

Table 1. Classification error rates. (GB=Graph-based classifier.)

methods	`iris`		`wine`		`seeds`		`pb`	
	kNN	GB	kNN	GB	kNN	GB	kNN	GB
[25]	4.61	4.41	3.84	4.88	7.30	7.20	-	-
[26]	4.97	4.57	4.61	5.18	7.15	6.93	4.46	5.04
[10]	5.45	5.49	4.35	4.96	7.78	7.40	5.33	4.51
[6]	6.12	10.40	**3.58**	4.37	**6.92**	6.63	4.55	4.96
[27]	**4.35**	4.80	4.12	4.36	7.77	7.47	**4.44**	4.24
Prop.	**4.35**	**4.12**	4.27	**4.19**	7.10	**6.61**	4.8	**4.23**

5. REFERENCES

[1] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, no. 1, pp. 49–55, 1936.

[2] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[3] Guo-Jun Qi, Jinhui Tang, Zheng-Jun Zha, Tat-Seng Chua, and Hong-Jiang Zhang, "An efficient sparse metric learning in high-dimensional space via l 1-penalized log-determinant regularization," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 841–848.

[4] Daryl Lim, Gert Lanckriet, and Brian McFee, "Robust structural metric learning," in *International conference on machine learning*, 2013, pp. 615–623.

[5] Wei Liu, Cun Mu, Rongrong Ji, Shiqian Ma, John R Smith, and Shih-Fu Chang, "Low-rank similarity metric learning in high dimensions," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[6] Pourya Zadeh, Reshad Hosseini, and Suvrit Sra, "Geometric mean metric learning," in *International conference on machine learning*, 2016, pp. 2464–2471.

[7] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[8] Yadong Mu, "Fixed-rank supervised metric learning on riemannian manifold," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[9] Jie Zhang and Lijun Zhang, "Efficient stochastic optimization for low-rank distance metric learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[10] Cheng Yang, Gene Cheung, and Vladimir Stankovic, "Alternating binary classifier and graph learning from partial labels," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1137–1140.

[11] Türker Biyikoglu, Josef Leydold, and Peter F Stadler, "Nodal domain theorems and bipartite subgraphs," 2005.

[12] Yuanchao Bai, Gene Cheung, Fen Wang, Xianming Liu, and Wen Gao, "Reconstruction-cognizant graph sampling using gershgorin disc alignment," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019.

[13] Yuanchao Bai, Fen Wang, Gene Cheung, Yuji Nakatsukasa, and Wen Gao, "Fast graph sampling set selection using gershgorin disc alignment," *arXiv*, 2019.

[14] Semyon Aranovich Gershgorin, "Über die abgrenzung der eigenwerte einer matrix," *Proceedings of the Russian Academy of Sciences*, , no. 6, pp. 749–754, 1931.

[15] Christos Papadimitriou and Kenneth Steiglitz, *Combinatorial Optimization*, Dover Publications, Inc, 1998.

[16] Martin Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *ICML*, Atlanta, Georgia, USA, Jun 2013, pp. 427–435.

[17] A. V. Knyazev, "Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method," *SIAM Journal on Scientific Computing*, vol. 23, no. 2, pp. 517–541, 2001.

[18] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, pp. 83–98, May 2013.

[19] Martin Vetterli, Jelena Kovacevic, and Vivek Goyal, *Foundations of Signal Processing*, Cambridge University Press, 2014.

[20] M. Milgram, "Irreducible graphs," *Journal Of Combinatorial Theory (B)*, vol. 12, pp. 6–31, 1972.

[21] Roger Horn and Charles Johnson, *Matrix Analysis*, Cambridge University Press, 2012.

[22] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2009.

[23] Jiahao Pang and Gene Cheung, "Graph laplacian regularization for image denoising: Analysis in the continuous domain," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 4, pp. 1770–1785, 2017.

[24] Y. Bai, G. Cheung, X. Liu, and W. Gao, "Graph-based blind image deblurring from a single photograph," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 3, pp. 1404–1418, 2019.

[25] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, 2003, pp. 912–919.

[26] Y. Mao, G. Cheung, C.-W. Lin, and Y. Ji, "Joint learning of similarity graph and image classifier from partial labels," in *APSIPA*, Jeju, South Korea, Dec. 2016.

[27] Wei Hu, Xiang Gao, Gene Cheung, and Zongming Guo, "Feature graph learning for 3d point cloud denoising," *CoRR*, vol. abs/1907.09138, 2019.