Orthogonality-Promoting Distance Metric Learning: Convex Relaxation and Theoretical Analysis

Pengtao Xie1,2 Wei Wu2 Yichen Zhu3 Eric P. Xing1

Abstract

Distance metric learning (DML), which learns a distance metric from labeled “similar” and “dissimilar” data pairs, is widely utilized. Recently, several works investigate orthogonality-promoting regularization (OPR), which encourages the projection vectors in DML to be close to being orthogonal, to achieve three effects: (1) high balancedness – achieving comparable performance on both frequent and infrequent classes; (2) high compactness – using a small number of projection vectors to achieve a “good” metric; (3) good generalizability – alleviating overfitting to training data. While showing promising results, these approaches suffer three problems. First, they involve solving nonconvex optimization problems where achieving the global optimal is NP-hard. Second, it lacks a theoretical understanding why OPR can lead to balancedness. Third, the current generalization error analysis of OPR is not directly on the regularizer. In this paper, we address these three issues by (1) seeking convex relaxations of the original nonconvex problems so that the global optimal is guaranteed to be achievable; (2) providing a formal analysis on OPR’s capability of promoting balancedness; (3) providing a theoretical analysis that directly reveals the relationship between OPR and generalization performance. Experiments on various datasets demonstrate that our convex methods are more effective in promoting balancedness, compactness, and generalization, and are computationally more efficient, compared with the nonconvex methods.

1. Introduction

Given data pairs labeled as either “similar” or “dissimilar”, distance metric learning (Xing et al., 2002; Weinberger et al., 2005; Davis et al., 2007) learns a distance measure in such a way that similar examples are placed close to each other while dissimilar ones are separated apart. The learned distance metrics are important to many downstream tasks, such as retrieval (Chen et al., 2017), classification (Weinberger et al., 2005) and clustering (Xing et al., 2002). One commonly used distance metric between two examples $x, y \in \mathbb{R}^D$ is: $\|Ax - Ay\|_2$ (Weinberger et al., 2005; Xie, 2015; Chen et al., 2017), which is parameterized by $R$ projection vectors (in $A \in \mathbb{R}^{R \times D}$).

Many works (Wang et al., 2012; Xie, 2015; Wang et al., 2015; Raziperchikolaei & Carreira-Perpinán, 2016; Chen et al., 2017) have proposed orthogonality-promoting DML to learn distance metrics that are (1) balanced: performing equally well on data instances belonging to frequent and infrequent classes; (2) compact: using a small number of projection vectors to achieve a “good” metric, (i.e., capturing well the relative distances of the data pairs); (3) generalizable: reducing the overfitting to training data. Regarding balancedness, under many circumstances, the frequency of classes, defined as the number of examples belonging to each class, can be highly imbalanced. Classic DML methods are sensitive to the skewness of the frequency of the classes: they perform favorably on frequent classes whereas less well on infrequent classes — a phenomenon also confirmed in our experiments in Section 7. However, infrequent classes are of crucial importance in many applications, and should not be ignored. For example, in a clinical setting, many diseases occur infrequently, but are life-threatening. Regarding compactness, the number of the projection vectors $R$ entails a tradeoff between performance and computational complexity (Ge et al., 2014b; Xie, 2015; Raziperchikolaei & Carreira-Perpinán, 2016). On one hand, more projection vectors bring in more expressiveness in measuring distance. On the other hand, a larger $R$ incurs a higher computational overhead since the number of weight parameters in $A$ grows linearly with $R$. It is therefore desirable to keep $R$ small without hurting much ML performance. Regarding generalization perfor-
We apply the learned distance metrics for information retrieval to healthcare, texts, images, and sensory data. Compared with non-convex baseline methods, our approaches achieve higher computational efficiency and are more capable of improving balancedness, compactness and generalizability.

### 2. Related Works

Many studies (Xing et al., 2002; Weinberger et al., 2005; Davis et al., 2007; Guillaumin et al., 2009; Ying & Li, 2012; Kostinger et al., 2012; Zadeh et al., 2016) have investigated DML (for a detailed review, please refer to the supplements and (Kulis et al., 2013; Wang & Sun, 2015)). To avoid overfitting in DML, various regularization approaches have been explored, which include KL-divergence (Davis et al., 2007), $\ell_1$ norm, trace norm (Niu et al., 2012; Liu et al., 2015), and dropout (Qian et al., 2014). Many works (Liu et al., 2008; Weiss et al., 2009; Kong & Li, 2012; Wang et al., 2012; Gong et al., 2013; Fu et al., 2014; Ge et al., 2014b; a; Ji et al., 2014; Wang et al., 2015; Xie, 2015; Carreira-Perpinán & Raziperchikolaei, 2016; Raziperchikolaei & Carreira-Perpinán, 2016; Yao et al., 2016; Chen et al., 2017) study orthogonality-promoting regularization in the context of DML or hashing. They define regularizers based on squared Frobenius norm (Wang et al., 2012; Fu et al., 2014; Ge et al., 2014b; Chen et al., 2017) or angles (Xie, 2015; Yao et al., 2016) to encourage the projection vectors to approach orthogonal.

### 3. Preliminaries

We review a DML method (Xie et al., 2017) that uses BMD (Kulis et al., 2009) to promote orthogonality.

**Distance Metric Learning**

Given data pairs labeled either as “similar” $S = \{(x_i, y_i)\}_{i=1}^{|S|}$ or “dissimilar” $D = \{(x_i, y_i)\}_{i=1}^{|D|}$. DML (Xing et al., 2002; Weinberger et al., 2005; Davis et al., 2007) aims to learn a distance metric under which similar examples are close to each other and dissimilar ones are separated far apart. There are many ways to define a distance metric. Here, we present two popular choices. One is based on linear projection (Weinberger et al., 2005; Xie, 2015; Chen et al., 2017). Given two examples $x, y \in \mathbb{R}^D$, a linear projection matrix $A \in \mathbb{R}^{R \times D}$ can be utilized to map them into a $R$-dimensional latent space. The distance metric is then defined as their squared Euclidean distance in the latent space: $\|Ax - Ay\|_2^2$. $A$ can be learned by minimizing (Xing et al., 2002):

$$\frac{1}{|S|} \sum_{(x,y) \in S} \|Ax - Ay\|_2^2 + \frac{1}{|D|} \sum_{(x,y) \in D} \max\{0, \tau - \|Ax - Ay\|_2^2\},$$

which aims at making the distances between similar examples as small as possible while separating dissimilar examples with a margin $\tau$ using a hinge loss. We call this formulation as projection matrix-based DML (PDML). PDML is a non-convex problem where the global optimal is difficult to achieve. Moreover, one needs to manually tune the number of projection vectors, typically via cross-validation, which incurs substantial computational overhead.

The other popular choice of distance metric is $(x - y)^\top M(x - y)$, which is cast from $\|Ax - Ay\|_2^2$ by replacing $A^\top A$ with a positive semidefinite (PSD) matrix.
Orthogonality-Promoting Distance Metric Learning: Convex Relaxation and Theoretical Analysis

M. This is known as the Mahalanobis distance (Xing et al., 2002). Correspondingly, the PDML formulation can be transformed into a Mahalanobis distance-based DML (MDML) problem: \( \min_{M \succeq 0} \frac{1}{2T} \sum_{(x,y) \in D} (x-y)^\top MM(x-y) + \frac{1}{2T} \sum_{(x,y) \in D} \max(0, \tau - (x-y)^\top M(x-y)) \), which is a convex problem where the global solution is guaranteed to be achievable. It also avoids tuning the number of projection vectors. However, the drawback of this approach is that, in order to satisfy the PSD constraint, one needs to perform eigen-decomposition of \( M \) in each iteration, which incurs \( O(D^3) \) complexity.

Orthogonality-Promoting Regularization Among the various orthogonality-promoting regularizers, we consider the BMD (Kulis et al., 2009) regularizer (Xie et al., 2017) in this study since it is amenable for convex relaxation and facilitates theoretical analysis.

To encourage orthogonality between two vectors \( a_i \) and \( a_j \), one can make their inner product \( a_i^\top a_j \) close to zero and their \( \ell_2 \) norm \( ||a_i||_2, ||a_j||_2 \) close to one. For a set of vectors \( \{a_i\}_{i \in 1} \), their near-orthogonality can be achieved by computing the Gram matrix \( G \) where \( G_{ij} = a_i^\top a_j \), then encouraging \( G \) to be close to an identity matrix. Off the diagonal of \( G \) and \( I \) are \( a_i^\top a_j \) and zero, respectively. On the diagonal of \( G \) and \( I \) are \( ||a_i||_2^2 \) and one, respectively. Making \( G \) close to \( I \) effectively encourages \( a_i^\top a_j \) to be close to zero and \( ||a_i||_2 \) close to one, which therefore encourages \( a_i \) and \( a_j \) to be close to orthogonal.

BMDs can be used to measure the “closeness” between two matrices. Let \( S^n \) denote real symmetric \( n \times n \) matrices. Given a strictly convex, differentiable function \( \phi : S^n \to \mathbb{R} \), a BMD is defined as \( \Omega(\mathbf{X}, \mathbf{Y}) = \phi(\mathbf{X}) - \phi(\mathbf{Y}) - \text{tr}(\mathbf{X}^\top \phi'(\mathbf{Y})^\top \mathbf{Y}) \), where \( \text{tr}(\mathbf{A}) \) denotes the trace of the matrix \( \mathbf{A} \). Different choices of \( \phi \) lead to different divergences. When \( \phi(\mathbf{Z}) = ||\mathbf{Z}||_F^2 \), the BMD is specialized to the squared Frobenius norm (SFN) \( ||\mathbf{X} - \mathbf{Y}||_F^2 \). If \( \phi(\mathbf{X}) = \text{tr}(\mathbf{X} \log \mathbf{X} - \mathbf{X}) \), where \( \log \mathbf{X} \) denotes the matrix logarithm of \( \mathbf{X} \), the divergence becomes \( \Gamma_{vnd}(\mathbf{X}, \mathbf{Y}) = \text{tr}(\mathbf{X} \log \mathbf{X} - \mathbf{X} \log \mathbf{Y} - \mathbf{X} + \mathbf{Y}) \), which is referred to as von Neumann divergence (VND) (Tsuda et al., 2005). If \( \phi(\mathbf{X}) = -\log \det \mathbf{X} \) where \( \det(\mathbf{X}) \) denotes the determinant of \( \mathbf{X} \), we get the log-determinant divergence (LDD) (Kulis et al., 2009): \( \Gamma_{ldd}(\mathbf{X}, \mathbf{Y}) = \text{tr}(\mathbf{X} \mathbf{Y}^{-1}) - \log \det(\mathbf{X} \mathbf{Y}^{-1}) - n \).

In PDML, to encourage orthogonality among the projection vectors (row vectors in \( \mathbf{A} \), Xie et al., 2017) define a family of regularizers \( \Omega(\mathbf{A}) = \Gamma_\phi(\mathbf{A}^\top \mathbf{A}) \) which encourage the BMD between the Gram matrix \( \mathbf{A}^\top \mathbf{A} \) and an identity matrix \( \mathbf{I} \) to be small. \( \Omega(\mathbf{A}) \) can be specialized to different instances, based on the choices of \( \Gamma_\phi(\cdot, \cdot) \). Under SFN, \( \Omega(\mathbf{A}) \) becomes \( \Omega_{sf}(\mathbf{A}) = ||\mathbf{A}^\top \mathbf{A} - \mathbf{I}||_F^2 \), which is used in (Wang et al., 2012; Fu et al., 2014; Ge et al., 2014b; Chen et al., 2017) to promote orthogonality. Under VND, \( \Omega(\mathbf{A}) \) becomes \( \Omega_{vnd}(\mathbf{A}) = \text{tr}(\mathbf{A}^\top \log(\mathbf{A}^\top \mathbf{A}) - \mathbf{A}^\top \mathbf{A}) + R \). Under LDD, \( \Omega(\mathbf{A}) \) becomes \( \Omega_{ldd}(\mathbf{A}) = \text{tr}(\mathbf{A}^\top \mathbf{A}) - \log \det(\mathbf{A})^2 - R \).

4. Convex Relaxation

The PDML-BMD problem is non-convex, where the global optimal solution of \( \mathbf{A} \) is very difficult to achieve. We seek a convex relaxation and solve the relaxed problem instead. The basic idea is to transform PDML into MDML and approximate the BMD regularizers with convex functions.

4.1. Convex Approximations of the BMD Regularizers

The approximations are based on the properties of eigenvalues. Given a full-rank matrix \( \mathbf{A} \in \mathbb{R}^{R \times D} (R < D) \), we know that \( \mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{R \times R} \) is a full-rank matrix with \( R \) positive eigenvalues \( \lambda_1, \ldots, \lambda_R \) and \( \mathbf{A}^\top \mathbf{A} \in \mathbb{R}^{D \times D} \) is a rank-deficient matrix with \( D - R \) zero eigenvalues and \( R \) positive eigenvalues that equal to \( \lambda_1, \ldots, \lambda_R \). For a general positive definite matrix \( \mathbf{Z} \in \mathbb{R}^{R \times R} \) whose eigenvalues are \( \gamma_1, \ldots, \gamma_R \), we have \( ||\mathbf{Z}||_F^2 = \sum_{j=1}^R \gamma_j^2 \), \( \text{tr}(\mathbf{Z}) = \sum_{j=1}^R \gamma_j \) and \( \log \det \mathbf{Z} = \sum_{j=1}^R \log \gamma_j \). Next, we leverage these facts to seek convex relaxations of the BMD regularizers.

A convex SFN regularizer The eigenvalues of \( \mathbf{A}^\top \mathbf{A} - \mathbf{I}_R \) are \( \lambda_1 - 1, \ldots, \lambda_R - 1 \) and those of \( \mathbf{A}^\top \mathbf{A} - \mathbf{I}_D \) are \( \lambda_1 - 1, \ldots, \lambda_R - 1, -1, \ldots, -1 \). Then \( \mathbf{A}^\top \mathbf{A} - \mathbf{I}_D \) is a Mahalanobis matrix and \( R \) is very difficult to achieve. We seek

\[
\hat{\Omega}_{w,fr}(\mathbf{M}) = ||\mathbf{M} - \mathbf{I}_D||_F^2 + \text{tr}(\mathbf{M}) \tag{1}
\]

A convex VND regularizer Given the eigen-decomposition \( \mathbf{A}^\top = \mathbf{U} \Lambda^\top \mathbf{U} \) where the eigenvalue \( \lambda_j \) equals to \( \lambda_j \), based on the property of the matrix logarithm, we have

\[
\log(\mathbf{A}^\top \mathbf{A}) = \mathbf{U} \log(\mathbf{A}^\top \mathbf{A}) \mathbf{U} \text{, where } \log(\mathbf{A}^\top \mathbf{A}) = \{\lambda_j \log(\mathbf{A}^\top \mathbf{A}) \} \text{.}
\]

Then \( \mathbf{A}^\top \mathbf{A} - \mathbf{I}_D \) is a Mahalanobis matrix and \( R \) is a small scalar. Using similar calculation, we have

\[
\Gamma_v = \text{tr}(\hat{\mathbf{A}}^\top \mathbf{A} + \epsilon \mathbf{I}_D) = \sum_{j=1}^R (\lambda_j + \epsilon)(\log(\lambda_j + \epsilon) - (\lambda_j + \epsilon)) + (D - R)(\epsilon \log \epsilon - \epsilon) + D.
\]

Performing certain algebra (see supplements), we get

\[
\Omega_v(\mathbf{A}) \approx \Gamma_v(\mathbf{A}^\top \mathbf{A} + \epsilon \mathbf{I}_D) + R + C \quad (C = \text{constant}).
\]

Replacing \( \mathbf{A}^\top \mathbf{A} \) with \( \mathbf{M} \), approximating \( R \) with \( \text{tr}(\mathbf{M}) \) and dropping
constant $D$, we get the convex VND (CVND) regularizer:

$$
\hat{\Omega}_{vnd}(M) = \Gamma_{vnd}(M + \epsilon I_D, I_D) + \text{tr}(M) \\
\propto \text{tr}((M + \epsilon I_D) \log(M + \epsilon I_D))
$$

whence convexity is shown in (Nielsen & Chuang, 2000).

A convex LDD regularizer We have $\Omega_{ldd}(A) = \sum_{j=1}^{R} \lambda_j - \sum_{j=1}^{R} \log \lambda_j - R$ and $\Gamma_{ldd}(A^T A + \epsilon I_D, I_D) = \sum_{j=1}^{R} \lambda_j + D \epsilon - (D - R) \log \epsilon - \sum_{j=1}^{R} \log \lambda_j + \epsilon$. Certain algebra shows that $\Omega_{ldd}(A) \approx \Gamma_{ldd}(A^T A + \epsilon I_D, I_D) - (1 + \log \epsilon)R + D \log \epsilon$. After replacing $A^T A$ with $M$, approximating $R$ with $\text{tr}(M)$ and discarding constants, we obtain the convex LDD (CLDD) regularizer:

$$
\hat{\Omega}_{ldd}(M) = \Gamma_{ldd}(M + \epsilon I_D, I_D) - (1 + \log \epsilon)\text{tr}(M) \\
\propto - \log \det(M + \epsilon I_D) + (\log \frac{1}{\epsilon}) \text{tr}(M)
$$

where the convexity of $\log \det(M + \epsilon I_D)$ is proved in (Boyd & Vandenberghe, 2004). Note that in (Davis et al., 2007; Qi et al., 2009), an information theoretic regularizer based on log-determinant divergence $\Gamma_{ldd}(M, I) = - \log \det(M) + \text{tr}(M)$ is applied to encourage the Mahalanobis matrix to be close to the identity matrix. This regularizer requires $M$ to be full rank; in contrast, by associating a large weight $\epsilon$ close to the identity matrix. This regularizer requires $M$ to be low-rank. Since $M = A^T A$, reducing the rank of $M$ reduces the number of projection vectors in $A$.

We discuss the errors in convex approximation, which are from two sources: one is the approximation of $\Omega_{\phi}(A)$ using $\Gamma_{\phi}(A^T A + \epsilon I_D, I_D)$ where the error is controlled by $\epsilon$ and can be arbitrarily small (by setting $\epsilon$ to be very small); the other is the approximation of the matrix rank using the trace norm. Though the error of the second approximation can be large, it has been both empirically and theoretically (Candes & Recht, 2012) demonstrated that decreasing the trace norm can effectively reduce rank. We empirically verify that decreasing the convexified CSFN, CVND and CLDD regularizers can decrease the original non-convex counterparts SFN, VND and LDD (see supplements). A rigorous analysis is left for future study.

4.2. DML with a Convex BMD Regularization

Given these convex BMD (CBMD) regularizers (denoted by $\hat{\Omega}_\phi(M)$), we relax the non-convex PDML-BMD problems into convex MDML-CBMD formulations by replacing $\|Ax - Ay\|^2_2$ with $(x - y)^T M(x - y)$ and replacing the non-convex BMD regularizers $\Omega_{\phi}(A)$ with $\hat{\Omega}_{\phi}(M)$:

$$
\min_{M \geq 0} \frac{1}{2\eta} \sum_{(x,y) \in S} (x - y)^T M(x - y) + \gamma \hat{\Omega}_{\phi}(M) \\
+ \frac{1}{2\eta} \sum_{(x,y) \in D} \max(0, \tau - (x - y)^T M(x - y))
$$

5. Optimization

We use stochastic proximal subgradient descent algorithm (Parikh & Boyd, 2014) to solve the MDML-CBMD problems. The algorithm iteratively performs the following steps until convergence: (1) randomly sampling a mini-batch of data pairs, computing the subgradient $\Delta M$ of the data-dependent loss (the first and second term in the objective function) defined on the mini-batch, then performing a subgradient descent update: $\hat{M} = M - \eta \Delta M$, where $\eta$ is a small stepsize; and (2) applying proximal operators associated with the regularizers $\hat{\Omega}_{\phi}(M)$ to $\hat{M}$. The gradient of the CVND regularizer is $\log \det(M + \epsilon I_D) + I_D$. To compute $\log(M + \epsilon I_D)$, we first perform an eigen-decomposition: $M + \epsilon I_D = \mathbf{U} \Lambda \mathbf{U}^\top$, then take the log of every eigenvalue in $\Lambda$ which gets us a new diagonal matrix $\Lambda$, and finally compute $\log(M + \epsilon I_D)$ as $\mathbf{U} \Lambda \mathbf{U}^\top$. In the CLDD regularizer, the gradient of $\log \det(M + \epsilon I_D)$ is $(M + \epsilon I_D)^{-1}$, which can also be computed by eigen-decomposition. Next, we present the proximal operators.

5.1. Proximal Operators

Given the regularizer $\hat{\Omega}_\phi(M)$, the associated proximal operator $\text{prox}(\hat{\Omega}_\phi(M))$ is defined as: $\text{prox}(\hat{\Omega}_\phi(M)) = \arg \min_M \frac{1}{2\eta} \|M - \hat{M}\|^2_2 + \gamma \hat{\Omega}_\phi(M)$, subject to $M \succeq 0$.

Let $\{\lambda_j\}_{j=1}^D$ be the eigenvalues of $\hat{M}$ and $\{x_j\}_{j=1}^D$ be the eigenvalues of $M$, then the above problem can be equivalently written as:

$$
\min_{\{x_j\}_{j=1}^D} \frac{1}{2\eta} \sum_{j=1}^D (x_j - \hat{\lambda}_j)^2 + \gamma \sum_{j=1}^D h_\phi(x_j)
$$

s.t.
$$
\forall j = 1, \ldots, D, \quad x_j \geq 0
$$

where $h_\phi(x_j)$ is a regularizer-specific scalar function. This problem can be decomposed into $D$ independent ones: (P)

$$
\min_{x_j} f(x_j) = \frac{1}{2\eta} (x_j - \hat{\lambda}_j)^2 + \gamma h_\phi(x_j), \quad \text{subject to } x_j \geq 0,
$$

for $j = 1, \ldots, D$, which can be solved individually.

SFN For SFN where $\hat{\Omega}_\phi(M) = \|M - I_D\|^2_F + \text{tr}(M)$ and $h_{sfn}(x_j) = (x_j - 1)^2 + x_j$, the problem (P) is simply a quadratic programming problem. The optimal solution is

$$
x_j^* = \max(0, \frac{\lambda_j + \eta\gamma}{1 + 2\eta\gamma})
$$

VND For VND where $\hat{\Omega}_\phi(M) = \text{tr}((M + \epsilon I_D) \log(M + \epsilon I_D))$ and $h_{vnd}(x_j) = (x_j + \epsilon) \log(x_j + \epsilon)$, by taking the derivative of the objective function $f(x_j)$ in problem (P) w.r.t $x_j$ and setting the derivative to zero, we get $\eta \gamma \log(x_j + \epsilon) + x_j + \eta \gamma - \hat{\lambda}_j = 0$. The root of this equation is: $\eta \gamma \omega(\frac{-m \pm \hat{\lambda}_j}{\eta \gamma} - \log(\eta \gamma)) - \epsilon$, where $\omega(\cdot)$ is the Wright omega function (Gorenflo et al., 2007). If this root is negative, then the optimal $x_j$ is 0; if this root is positive, then the optimal $x_j$ could be either this root or 0. We pick the one that yields the lowest $f(x_j)$. Formally, $x_j^* = \arg \min_{x_j} f(x_j)$, where $x \in \{\max(\eta \gamma \omega(\frac{-m \pm \hat{\lambda}_j}{\eta \gamma} - \log(\eta \gamma)) - \epsilon, 0)\}$.
The optimal solution is achieved either at the positive roots by taking the derivative of \( f(x) \) w.r.t. \( x \) and setting the derivative to zero, we get a quadratic equation: 
\[
x^2 + ax + b = 0,
\]
by \( a = \epsilon - \lambda_j - \eta \gamma \log \epsilon \) and \( \eta \gamma (1 - \epsilon \log \epsilon) \).

The optimal solution is achieved either at the positive roots (if any) of this equation or \( 0 \). We pick the one that yields the lowest \( f(x_j) \). Formally, 
\[
x_j^* = \arg\min_{x_j} f(x_j),
\]
where \( x \in \{ \max(\frac{-b+\sqrt{b^2-4ac}}{2a},0) \}, \max(\frac{-b-\sqrt{b^2-4ac}}{2a},0) \} \).

Computational Complexity In this algorithm, the major computation workload is eigen-decomposition of \( D \)-by-\( D \) matrices, with a complexity of \( O(D^3) \). In our experiments, since \( D \) is no more than 1000, \( O(D^3) \) is not a big bottleneck. Besides, these matrices are symmetric, the structures of which can thus be leveraged to speed up eigen-decomposition. In implementation, we use the MAGMA\(^1\) library that supports the efficient eigen-decomposition of symmetric matrices on GPU. Note that the unregularized MDML also requires the eigen-decomposition (of \( M \)), hence adding these CBMD regularizers does not substantially increase additional computation cost.

6. Theoretical Analysis

In this section, we present theoretical analysis of balancedness and generalization error.

6.1. Analysis of Balancedness

In this section, we analyze how the nonconvex BMD regularizers that promote orthogonality affect the balancedness of the distance metrics learned by PDML-BMD\(^2\). Specifically, the analysis focuses on the following projection matrix: 
\[
A^* = \arg\min_{A} \mathbb{E}_{S,D} \left[ \frac{1}{|S|} \sum_{(x,y) \in S} ||Ax-Ay||_2^2 + \frac{1}{|D|} \sum_{(x,y) \in D} \max(0, \tau - ||Ax-Ay||_2^2) + \gamma \Omega(A) \right].
\]
We assume there are \( K \) classes, where class \( k \) has a distribution \( p_k \) and the corresponding expectation is \( \mu_k \). Each data sample in \( S \) and \( D \) is drawn from the distribution of one specific class. We define \( \xi_k = \mathbb{E}_{x \sim p_k} [\sup_{v \in \mathbb{R}^c} ||v||_2 = 1] \left| v^T (x - \mu_k) \right| \) and \( \xi = \max_k \xi_k \). Further, we assume \( A^* \) is of full rank \( R \) (which is the number of the projection vectors), and let \( UA^*U^T \) denote the eigen-decomposition of \( A^*A^{*\top} \), where \( A = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_R) \) with \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R \).

We define an imbalance factor (IF) to measure the (im)balancedness. For each class \( k \), we use the corresponding expectation \( \mu_k \) to characterize this class. We define the Mahalanobis distance between two classes \( j \) and \( k \) as: 
\[
d_{jk} = (\mu_j - \mu_k)^\top A^{*\top} A^* (\mu_j - \mu_k).
\]
We define the IF among all classes as:
\[
\eta = \frac{\max_{j \neq k} d_{jk}}{\min_{j \neq k} d_{jk}}.
\]

The motivation of such a definition is: for two frequent classes, since they have more training examples and hence contributing more in learning \( A^* \), DML intends to make their distance \( d_{jk} \) large; whereas for two infrequent classes, since they contribute less in learning (and DML is constrained by similar pairs which need to have small distances), their distance may end up being small. Consequently, if classes are imbalanced, some between-class distances can be large while others small, resulting in a large IF. The following theorem shows the upper bounds of IF:

**Theorem 1** Let \( C \) denote the ratio between \( \max_{j \neq k} ||\mu_j - \mu_k||_2 \) and \( \min_{j \neq k} ||\mu_j - \mu_k||_2 \) and assume \( \max_{j,k} ||\mu_j - \mu_k||_2 \leq B_0 \). Suppose the regularization parameter \( \gamma \) and distance margin \( \tau \) are sufficiently large: \( \gamma \geq \gamma_0 \) and \( \tau \geq \tau_0 \), where \( \gamma_0 \) and \( \tau_0 \) depend on \( \{ p_k \}_{k=1}^K \) and \( \{ \mu_k \}_{k=1}^K \). If \( R \geq K - 1 \) and \( \xi \leq (-B_0 + \sqrt{B_0^2 + \lambda_{K-1} \beta_{K-1}}/(2\tau \Omega(A)))/4 \), then we have the following bounds for the IF:

- For the VND regularizer \( \Omega_{\text{vnd}}(A^*) \), if \( \Omega_{\text{vnd}}(A^*) \leq 1 \), the following bound of the IF \( \eta \) holds:
  \[
  \eta \leq C g(\Omega_{\text{vnd}}(A^*))
  \]
where \( g(\cdot) \) is an increasing function defined in the following way. Let \( f(c) = e^{1/(c+1)}(1 + 1/c) \), which is strictly increasing on \( [0,1] \) and strictly decreasing on \( [1, \infty] \) and let \( f^{-1}(c) \) be the inverse function of \( f(c) \) on \( [1, \infty] \), then \( g(c) = f^{-1}(2 - c) \) for \( c < 1 \).

- For the LDD regularizer \( \Omega_{\text{ldd}}(A^*) \), we have
  \[
  \eta \leq 4C e^{\Omega_{\text{ldd}}(A^*)}
  \]

As can be seen, the bounds are increasing functions of the BMD regularizers \( \Omega_{\text{vnd}}(A^*) \) and \( \Omega_{\text{ldd}}(A^*) \). Decreasing these regularizers would reduce the upper bounds of the imbalance factor, hence leading to more balancedness. For SFN, such a bound cannot be derived.

6.2. Analysis of Generalization Error

In this section, we analyze how the convex BMD regularizers affect the generalization error in MDML-CBMD problems. Following (Verma & Branson, 2015), we use distance-based error to measure the quality of a Mahalanobis distance matrix \( M \). Given the sample \( S \) and \( D \) where the total number of data pairs is \( m = |S| + |D| \), the empirical error is defined as 
\[
L(M) = \frac{1}{|S|} \sum_{(x,y) \in S} (x - \text{SPN}, \text{DML})
\]

\(^1\)http://icl.cs.utk.edu/magma/
\(^2\)The analysis of convex BMD regularizers in MDML-CBMD will be left for future work.
Theorem 2 Suppose the generalization error bound. Let $M^*$ be optimal matrix learned by minimizing the empirical error: $\hat{M}^* = \arg\min_M \hat{L}(M)$. We are interested in how well $\hat{M}^*$ performs on unseen data. The performance is measured using generalization error: $\mathcal{E} = L(M^*) - \hat{L}(M^*)$. To incorporate the impact of the CBMD regularizers $\Omega_\phi(M)$, we define the hypothesis class of $M$ to be $\mathcal{M} = \{M \geq 0 : \Omega_\phi(M) \leq C\}$. The upper bound $C$ controls the strength of regularization. A smaller $C$ entails stronger promotion of orthogonality. $C$ is controlled by the regularization parameter $\gamma$ in Eq.(4). Increasing $\gamma$ reduces $C$. For different CBMD regularizers, we have the following generalization error bound.

**Theorem 2** Suppose $\sup_{|v| \leq 1}(x,y) \in S |v^\top(x-y)| \leq B$, then with probability at least $1 - \delta$, we have:

- For the CVND regularizer,
  $$\mathcal{E} \leq (4B^2C + \max(\tau, B^2C)\sqrt{2\log(1/\delta)}) \frac{1}{\sqrt{m}}.$$

- For the CLDD regularizer,
  $$\mathcal{E} \leq \left( \frac{4B^2C}{\log(1/\epsilon)-1} + \max(\tau, C - D_\epsilon)\sqrt{2\log(1/\delta)}) \frac{1}{\sqrt{m}}.\right.$$

- For the CSFN regularizer,
  $$\mathcal{E} \leq (2B^2 \min(2C, \sqrt{C}) + \max(\tau, C)\sqrt{2\log(1/\delta)}) \frac{1}{\sqrt{m}}.$$

From these generalization error bounds (GBEs), we can see two major implications. First, CBMD regularizers can effectively control the GEBs. Increasing the strength of CBMD regularization (by enlarging $\gamma$) reduces $C$, which decreases the GEBs since they are all increasing functions of $C$. Second, the GEBs converge with rate $O(1/\sqrt{m})$, where $m$ is the number of training data pairs. This rate matches with that in (Bellet & Habrard, 2015; Verma & Branson, 2015).

### 7. Experiments

**Datasets** We used 7 datasets in the experiments: two electronic health record datasets MIMIC (version III) (Johnson et al., 2016) and EICU (version 1.1) (Goldberger et al., 2000); two text datasets Reuters and 20-Newsgroups (News); two image datasets Stanford-Cars (Cars) (Krause et al., 2013) and Caltech-UCSD-Birds (Birds) (Welinder et al., 2010); and one sensory dataset 6-Activities (Act) (Anguita et al., 2012). The class labels in MIMIC and EICU are the primary diagnoses of patients. In Reuters, documents belong to more than one classes are removed. Since there is no standard split of the training/test set, we perform five random splits and average the results of the five runs. The details of the datasets and feature extraction are deferred to the supplements.

**Experimental Settings** Two examples are considered as similar if they belong to the same class and dissimilar if otherwise. The learned distance metrics are applied for retrieval (using each test example to query the rest of the test examples) whose performance is evaluated using the Area Under precision-recall Curve (AUC) (Manning et al., 2008). We apply the proposed convex regularizers CSFN, CVND, CLDD to MDML. We compare them with two sets of baseline regularizers. The first set aims at promoting orthogonality, which are based on determinant of covariance (DC) (Malkin & Bilmes, 2008), cosine similarity (CS) (Yu et al., 2011), determinantal point process (DPP) (Kulesza et al., 2012; Zou & Adams, 2012), InCoherence (IC) (Bao et al., 2013), variational Gram function (VGF) (Zhou et al., 2011; Jalali et al., 2015), decorrelation (DeC) (Cogswell et al., 2015), mutual angles (MA) (Xie et al., 2015), squared Frobenius norm (SNF) (Wang et al., 2012; Fu et al., 2014; Ge et al., 2014b; Chen et al., 2017), von Neumann divergence (VND) (Xie et al., 2017), log-determinant divergence (LDD) (Xie et al., 2017), and orthogonal constraint (OC) $\textbf{A}\textbf{A}^\top = \textbf{I}$ (Liu et al., 2008; Wang et al., 2015). All these regularizers are applied to PDML. The other set of regularizers are not designed particularly for promoting orthogonality but are commonly used, including $\ell_2$ norm, $\ell_1$ norm (Qi et al., 2009), $\ell_{2,1}$ norm (Ying et al., 2009), trace norm (Tr) (Liu et al., 2015), information theoretic (IT) regularizer $-\log\det(M) + \text{tr}(M)$ (Davis et al., 2007), and Dropout (Drop) (Srivastava et al., 2014). All these regularizers are applied to MDML. We compare with a common approach for dealing with class-imbalance: oversampling (OS) (Galar et al., 2012). In addition, we compare with other DML methods including LMNN (Weinberger et al., 2005), ITML (Davis et al., 2007), LDML (Guillaumin et al., 2009), MLEC (Kostinger et al., 2012), GMML (Zadeh et al., 2016), and ILHD (Carreira-Perpinán & Raziperchikolaei, 2016).

**Results** The training time taken by different methods to reach convergence is shown in Table 2. For the non-convex, PDML-based methods, we report the total time taken by the following computation: tuning the regularization parameter (4 choices) and the number of projection vectors (NPVs, 6 choices) on a two-dimensional grid via 3-fold cross validation ($4 \times 6 \times 3 = 72$ experiments in total); for each of the 72 experiments, the algorithm restarts 5 times\(^6\), each

\(^6\)Our experiments show that for non-convex methods, multiple re-starts are of great necessity to improve performance. For example, for PDML-VND on MIMIC with 100 projection vectors, the AUC is non-decreasing with the number of re-starts: the AUC after 1, 2, ..., 5 re-starts are 0.651, 0.651, 0.658, 0.667, 0.667.
Orthogonality-Promoting Distance Metric Learning: Convex Relaxation and Theoretical Analysis

Table 1. On the three imbalanced datasets – MIMIC, EICU, Reuters, we show the mean AUC (averaged on 5 random train/test splits) on all classes (A-All) and infrequent classes (A-IF) and the balance score. On the rest 4 balanced datasets, A-All is shown. The AUC on frequent classes and the standard errors are in the supplements.

<table>
<thead>
<tr>
<th></th>
<th>MIMIC</th>
<th>EICU</th>
<th>Reuters</th>
<th>News</th>
<th>Cars</th>
<th>Birds</th>
<th>Act</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A-All</td>
<td>A-IF</td>
<td>BS</td>
<td>A-All</td>
<td>A-IF</td>
<td>BS</td>
<td>A-All</td>
</tr>
<tr>
<td>PDML</td>
<td>0.634</td>
<td>0.608</td>
<td>0.070</td>
<td>0.671</td>
<td>0.637</td>
<td>0.077</td>
<td>0.949</td>
</tr>
<tr>
<td></td>
<td>0.641</td>
<td>0.617</td>
<td>0.069</td>
<td>0.677</td>
<td>0.652</td>
<td>0.055</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td>0.777</td>
<td>0.722</td>
<td>0.085</td>
<td>0.816</td>
<td>0.855</td>
<td>0.052</td>
<td>0.769</td>
</tr>
<tr>
<td>LMINN</td>
<td>0.628</td>
<td>0.609</td>
<td>0.054</td>
<td>0.662</td>
<td>0.633</td>
<td>0.066</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td>0.934</td>
<td>0.906</td>
<td>0.042</td>
<td>0.748</td>
<td>0.706</td>
<td>0.047</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>0.916</td>
<td>0.921</td>
<td>0.067</td>
<td>0.738</td>
<td>0.707</td>
<td>0.017</td>
<td>0.781</td>
</tr>
<tr>
<td></td>
<td>0.676</td>
<td>0.647</td>
<td>0.046</td>
<td>0.934</td>
<td>0.906</td>
<td>0.042</td>
<td>0.748</td>
</tr>
<tr>
<td></td>
<td>0.701</td>
<td>0.677</td>
<td>0.053</td>
<td>0.794</td>
<td>0.728</td>
<td>0.082</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>0.703</td>
<td>0.661</td>
<td>0.091</td>
<td>0.785</td>
<td>0.731</td>
<td>0.087</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>0.676</td>
<td>0.707</td>
<td>0.067</td>
<td>0.771</td>
<td>0.724</td>
<td>0.058</td>
<td>0.956</td>
</tr>
<tr>
<td></td>
<td>0.689</td>
<td>0.679</td>
<td>0.045</td>
<td>0.779</td>
<td>0.732</td>
<td>0.069</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>0.701</td>
<td>0.670</td>
<td>0.057</td>
<td>0.780</td>
<td>0.733</td>
<td>0.089</td>
<td>0.951</td>
</tr>
<tr>
<td></td>
<td>0.676</td>
<td>0.730</td>
<td>0.095</td>
<td>0.816</td>
<td>0.751</td>
<td>0.017</td>
<td>0.956</td>
</tr>
<tr>
<td></td>
<td>0.689</td>
<td>0.679</td>
<td>0.045</td>
<td>0.779</td>
<td>0.732</td>
<td>0.069</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>0.676</td>
<td>0.707</td>
<td>0.067</td>
<td>0.771</td>
<td>0.724</td>
<td>0.058</td>
<td>0.956</td>
</tr>
<tr>
<td></td>
<td>0.689</td>
<td>0.679</td>
<td>0.045</td>
<td>0.779</td>
<td>0.732</td>
<td>0.069</td>
<td>0.963</td>
</tr>
</tbody>
</table>

Next, we verify whether CSFN, CVND and CLDD are able to learn more balanced distance metrics. On three datasets MIMIC, EICU and Reuters where the classes are imbalanced, we consider a class as “frequent” if it contains more than 1000 examples, and “infrequent” if otherwise. We measure AUCs on all classes (A-All), infrequent classes (A-IF) and frequent classes (A-F), then define a balance score (BS) as $\frac{\text{AUC}_{\text{A-All}} - \text{AUC}_{\text{A-IF}}}{\text{AUC}_{\text{A-All}} + \text{AUC}_{\text{A-IF}} - 1}$. A smaller BS indicates more balancedness. As shown in Table 1, MDML-(CSFN,CVND,CLDD) achieve the highest A-All on 6 datasets and the highest A-IF on all 3 imbalanced datasets. In terms of BS, our convex methods outperform all baseline DML methods. These results demonstrate our methods can learn more balanced metrics. By encouraging the projection vectors to be close to being orthogonal, our methods can reduce the redundancy among vectors. Mutually complementary vectors can achieve a broader coverage of latent features, including those associated with infrequent classes.

Table 2. Training time (hours) on seven datasets. The training time of other baseline methods are deferred to the supplements.

<table>
<thead>
<tr>
<th></th>
<th>MIMIC</th>
<th>EICU</th>
<th>Reuters</th>
<th>News</th>
<th>Cars</th>
<th>Birds</th>
<th>Act</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDML-DC</td>
<td>424.7</td>
<td>499.2</td>
<td>35.2</td>
<td>85.6</td>
<td>61.8</td>
<td>66.2</td>
<td>17.2</td>
</tr>
<tr>
<td>PDML-CS</td>
<td>263.2</td>
<td>284.8</td>
<td>22.6</td>
<td>47.2</td>
<td>34.5</td>
<td>42.8</td>
<td>14.4</td>
</tr>
<tr>
<td>PDML-DPP</td>
<td>411.8</td>
<td>479.1</td>
<td>36.9</td>
<td>61.9</td>
<td>64.2</td>
<td>70.5</td>
<td>16.5</td>
</tr>
<tr>
<td>PDML-RC</td>
<td>263.9</td>
<td>281.2</td>
<td>25.3</td>
<td>46.1</td>
<td>37.5</td>
<td>45.2</td>
<td>15.3</td>
</tr>
<tr>
<td>PDML-DCT</td>
<td>458.5</td>
<td>497.5</td>
<td>41.8</td>
<td>78.2</td>
<td>79.8</td>
<td>80.7</td>
<td>19.9</td>
</tr>
<tr>
<td>PDML-VGF</td>
<td>267.3</td>
<td>284.1</td>
<td>22.3</td>
<td>48.9</td>
<td>35.8</td>
<td>38.7</td>
<td>15.4</td>
</tr>
<tr>
<td>PDML-MA</td>
<td>271.4</td>
<td>282.3</td>
<td>25.6</td>
<td>50.2</td>
<td>30.9</td>
<td>39.6</td>
<td>17.5</td>
</tr>
<tr>
<td>PDML-OS</td>
<td>404.9</td>
<td>418.9</td>
<td>45.7</td>
<td>143.8</td>
<td>143.8</td>
<td>143.8</td>
<td>59.9</td>
</tr>
<tr>
<td>PDML-SFN</td>
<td>261.7</td>
<td>277.6</td>
<td>22.9</td>
<td>46.5</td>
<td>36.2</td>
<td>38.2</td>
<td>15.9</td>
</tr>
<tr>
<td>PDML-VND</td>
<td>403.8</td>
<td>488.3</td>
<td>33.8</td>
<td>62.5</td>
<td>67.5</td>
<td>73.4</td>
<td>17.1</td>
</tr>
<tr>
<td>PDML-LDD</td>
<td>407.5</td>
<td>483.5</td>
<td>34.3</td>
<td>60.1</td>
<td>61.8</td>
<td>72.6</td>
<td>17.9</td>
</tr>
<tr>
<td>PDML-CSFN</td>
<td>41.1</td>
<td>43.9</td>
<td>3.3</td>
<td>7.3</td>
<td>6.5</td>
<td>6.9</td>
<td>1.8</td>
</tr>
<tr>
<td>PDML-CVND</td>
<td>43.8</td>
<td>46.7</td>
<td>3.6</td>
<td>8.1</td>
<td>6.9</td>
<td>7.8</td>
<td>2.0</td>
</tr>
<tr>
<td>PDML-CLDD</td>
<td>41.7</td>
<td>44.5</td>
<td>3.4</td>
<td>7.5</td>
<td>6.6</td>
<td>7.2</td>
<td>1.8</td>
</tr>
</tbody>
</table>
In general, the orthogonality-promoting (OP) regularizers outperform the non-OP regularizers, suggesting the effectiveness of promoting orthogonality. The orthogonal constraint (OC) (Liu et al., 2008; Wang et al., 2015) imposes strict orthogonality, which may be too restrictive that hurts performance. ILHD (Carreira-Perpinán & Raziperchikolaei, 2016) learns binary hash codes, which makes retrieval more efficient, however, it achieves lower AUCs due to the quantization errors. MDML-(CSFN,CVND,CLDD) are compact. Table 3 shows the numbers of projection vectors (NPVs) and compactness score (CS, × 10^-3).

Next we verify whether the learned distance metrics by MDML-(CSFN,CVND,CLDD) are compact. Table 3 shows the numbers of the projection vectors (NPVs) that achieve the AUCs in Table 1. For MDML-based methods, the NPV equals to the rank of the Mahalanobis matrix since \( \mathbf{M} = \mathbf{A}^\top \mathbf{A} \). We define a compactness score (CS) which is the ratio between \( \mathbf{A} \)-All (given in Table 1) and NPV. A higher CS indicates achieving higher AUC by using fewer projection vectors. From Table 3, we can see that on 5 datasets, MDML-(CSFN,CVND,CLDD) achieve larger CSs than the baseline methods, demonstrating their better capability in learning compact distance metrics. Similar to the observations in Table 1, CSFN, CVND and CLDD perform better than non-convex regularizers, and CVND, CLDD perform better than CSFN. The reduction of NPV improves the efficiency of retrieval since the computational complexity grows linearly with this number. Together, these results demonstrate that MDML-(CSFN,CVND,CLDD) outperform other methods in terms of learning both compact and balanced distance metrics.

As can be seen from Table 1, our methods MDML-(CVND,CLDD) achieve the best AUC-All. In Table 5 in the supplements, it is shown that MDML-(CVND,CLDD) have the smallest gap between training and testing AUC. This indicates that our methods are better capable of reducing overfitting and improving generalization performance.

### 8. Conclusions

In this paper, we have attempted to address three issues of existing orthogonality-promoting DML methods, which include computational inefficiency and lacking theoretical analysis in balancedness and generalization. To address the computation issue, we perform a convex relaxation of these regularizers and develop a proximal gradient descent algorithm to solve the convex problems. To address the analysis issue, we define an imbalance factor (IF) to measure (im)balancedness and prove that decreasing the Bregman matrix divergence regularizers (which promote orthogonality) can reduce the upper bound of the IF, hence leading to more balancedness. We provide a generalization error (GE) analysis showing that decreasing the convex regularizers can reduce the GE upper bound. Experiments on datasets from different domains demonstrate that our methods are computationally more efficient and are more capable of learning balanced, compact and generalizable distance metrics than other approaches.
Acknowledgements

We would like to thank the anonymous reviewers for their very constructive and helpful comments and suggestions. Pengtao Xie, Wei Wu, and Eric P. Xing are supported by National Institutes of Health P30DA035778, R01GM114311, Pennsylvania Department of Health BD4BH4100070287, and National Science Foundation IIS1617583.

References


Ge, T., He, K., and Sun, J. Graph cuts for supervised binary coding. In ECCV, 2014b.


Ji, J., Yan, S., Li, J., Gao, G., Tian, Q., and Zhang, B. Batch-orthogonal locality-sensitive hashing for angular similarity. TPAMI, 2014.


Orthogonality-Promoting Distance Metric Learning: Convex Relaxation and Theoretical Analysis


Xie, P. Learning compact and effective distance metrics with diversity regularization. In ECML, 2015.


Ying, Y. and Li, P. Distance metric learning with eigenvalue optimization. JMLR, 2012.


