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

Pengtao Xie [1 2] **Wei Wu** [2] **Yichen Zhu** [3] **Eric P. Xing** [1]

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]Petuum Inc [2]School of Computer Science, Carnegie Mellon University [3]School of Mathematical Sciences, Peking University. Correspondence to: Pengtao Xie <pengtao.xie@petuum.com>, Eric P. Xing <eric.xing@petuum.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ is: $\|\mathbf{Ax} - \mathbf{Ay}\|_2$ (Weinberger et al., 2005; Xie, 2015; Chen et al., 2017), which is parameterized by R projection vectors (in $\mathbf{A} \in \mathbb{R}^{R\times D}$).

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 $\mathbf{A}$ grows linearly with R. It is therefore desirable to keep R small without hurting much ML performance. Regarding generalization performance, in the case where the sample size is small but the size of $\mathbf{A}$ is large, overfitting can easily happen.

To address these three issues, many studies (Wang et al., 2012; 2015; Xie, 2015; Carreira-Perpinán & Raziperchikolaei, 2016; Chen et al., 2017) propose to regularize the projection vectors to be close to being orthogonal. For balancedness, they argue that, without orthogonality-promoting regularization (OPR), the majority of projection vectors learn latent features for frequent classes since these classes have dominant signals in the dataset; through OPR, the projection vectors uniformly "spread out", giving both infrequent and frequent classes a fair treatment and thus leading to a more balanced distance metric (see (Xie et al., 2017) for details). For compactness, they claim that: "diversified" projection vectors bear less redundancy and are mutually complementary; as a result, a small number of such vectors are sufficient to achieve a "good" distance metric. For generalization performance, they posit that OPR imposes a structured constraint on the function class of DML, hence reduces model complexity.

While these orthogonality-promoting DML methods have shown promising results, they have three problems. First, they involve solving non-convex optimization problems where the global solution is extremely difficult, if not impossible, to obtain. Second, no formal analysis is conducted regarding why OPR can promote balancedness. Third, while the generalization error (GE) analysis of OPR has been studied in (Xie et al., 2017), it is incomplete. In this analysis, they first show that the upper bound of GE is a function of cosine similarity (CS), then show that CS and the regularizer are somewhat aligned in shape. They did not establish a direct relationship between the GE bound and the regularizer.

In this paper, we aim at addressing these problems by making the following contributions:

- We relax the nonconvex, orthogonality-promoting DML problems into convex problems and develop efficient proximal gradient descent algorithms. The algorithms only run once with a single initialization, and hence are much more efficient than existing non-convex methods.
- We perform theoretical analysis which formally reveals the relationship between OPR and balancedness: stronger OPR leads to more balancedness.
- We perform generalization error (GE) analysis which shows that reducing the convex orthogonality-promoting regularizers can reduce the upper bound of GE.
- 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), ℓ_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" $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{S}|}$ or "dissimilar" $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}|}$, 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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$, a linear projection matrix $\mathbf{A} \in \mathbb{R}^{R\times D}$ 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: $\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2^2$. $\mathbf{A}$ can be learned by minimizing (Xing et al., 2002): $\frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2^2 + \frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \max(0, \tau - \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2^2)$, which aims at making the distances between similar examples as small as possible while separating dissimilar examples with a margin τ 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 $(\mathbf{x} - \mathbf{y})^\top \mathbf{M}(\mathbf{x} - \mathbf{y})$, which is cast from $\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2^2$ by replacing $\mathbf{A}^\top\mathbf{A}$ with a positive semidefinite (PSD) matrix

$\mathbf{M}$. 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: $\min_{\mathbf{M}\succeq 0} \frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}}(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})+\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}\max(0,\tau-(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y}))$. 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 $\mathbf{M}$ in each iteration, which incurs $O(D^3)$ complexity.

Orthogonality-Promoting Regularization Among the various orthogonality-promoting regularizers, we choose 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 $\mathbf{a}_i$ and $\mathbf{a}_j$, one can make their inner product $\mathbf{a}_i^\top\mathbf{a}_j$ close to zero and their ℓ_2 norm $\|\mathbf{a}_i\|_2$, $\|\mathbf{a}_j\|_2$ close to one. For a set of vectors $\{\mathbf{a}_i\}_{i=1}^R$, their near-orthogonality can be achieved by computing the Gram matrix $\mathbf{G}$ where $G_{ij}=\mathbf{a}_i^\top\mathbf{a}_j$, then encouraging $\mathbf{G}$ to be close to an identity matrix. Off the diagonal of $\mathbf{G}$ and $\mathbf{I}$ are $\mathbf{a}_i^\top\mathbf{a}_j$ and zero, respectively. On the diagonal of $\mathbf{G}$ and $\mathbf{I}$ are $\|\mathbf{a}_i\|_2^2$ and one, respectively. Making $\mathbf{G}$ close to $\mathbf{I}$ effectively encourages $\mathbf{a}_i^\top\mathbf{a}_j$ to be close to zero and $\|\mathbf{a}_i\|_2$ close to one, which therefore encourages $\mathbf{a}_i$ and $\mathbf{a}_j$ to be close to orthogonal.

BMDs can be used to measure the "closeness" between two matrices. Let $\mathbf{S}^n$ denote real symmetric $n\times n$ matrices. Given a strictly convex, differentiable function $\phi:\mathbf{S}^n\rightarrow\mathbb{R}$, a BMD is defined as $\Gamma_\phi(\mathbf{X},\mathbf{Y})=\phi(\mathbf{X})-\phi(\mathbf{Y})-\mathrm{tr}((\bigtriangledown\phi(\mathbf{Y}))^\top(\mathbf{X}-\mathbf{Y}))$, where $\mathrm{tr}(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. Different choices of $\phi(\mathbf{X})$ lead to different divergences. When $\phi(\mathbf{X})=\|\mathbf{X}\|_F^2$, the BMD is specialized to the *squared Frobenius norm* (SFN) $\|\mathbf{X}-\mathbf{Y}\|_F^2$. If $\phi(\mathbf{X})=\mathrm{tr}(\mathbf{X}\log\mathbf{X}-\mathbf{X})$, where $\log\mathbf{X}$ denotes the matrix logarithm of $\mathbf{X}$, the divergence becomes $\Gamma_{vnd}(\mathbf{X},\mathbf{Y})=\mathrm{tr}(\mathbf{X}\log\mathbf{X}-\mathbf{X}\log\mathbf{Y}-\mathbf{X}+\mathbf{Y})$, 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): $\Gamma_{ldd}(\mathbf{X},\mathbf{Y})=\mathrm{tr}(\mathbf{X}\mathbf{Y}^{-1})-\log\det(\mathbf{X}\mathbf{Y}^{-1})-n$.

In PDML, to encourage orthogonality among the projection vectors (row vectors in $\mathbf{A}$), Xie et al. (2017) define a family of regularizers $\Omega_\phi(\mathbf{A})=\Gamma_\phi(\mathbf{A}\mathbf{A}^\top,\mathbf{I})$ which encourage the BMD between the Gram matrix $\mathbf{A}\mathbf{A}^\top$ and an identity matrix $\mathbf{I}$ to be small. $\Omega_\phi(\mathbf{A})$ can be specialized to different instances, based on the choices of $\Gamma_\phi(\cdot,\cdot)$. Under SFN, $\Omega_\phi(\mathbf{A})$ becomes $\Omega_{sfn}(\mathbf{A})=\|\mathbf{A}\mathbf{A}^\top-\mathbf{I}\|_F^2$, 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, $\Omega_\phi(\mathbf{A})$ becomes $\Omega_{vnd}(\mathbf{A})=\mathrm{tr}(\mathbf{A}\mathbf{A}^\top\log(\mathbf{A}\mathbf{A}^\top)-\mathbf{A}\mathbf{A}^\top)+R$. Under LDD, $\Omega_\phi(\mathbf{A})$ becomes $\Omega_{ldd}(\mathbf{A})=\mathrm{tr}(\mathbf{A}\mathbf{A}^\top)-\log\det(\mathbf{A}\mathbf{A}^\top)-R$.

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 $\mathbf{A}\mathbf{A}^\top\in\mathbb{R}^{R\times R}$ is a full-rank matrix with R positive eigenvalues $\lambda_1,\cdots,\lambda_R$ 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 $\lambda_1,\cdots,\lambda_R$. For a general positive definite matrix $\mathbf{Z}\in\mathbb{R}^{R\times R}$ whose eigenvalues are $\gamma_1,\cdots,\gamma_R$, we have $\|\mathbf{Z}\|_F^2=\sum_{j=1}^R\gamma_j^2$, $\mathrm{tr}(\mathbf{Z})=\sum_{j=1}^R\gamma_j$ 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 $\mathbf{A}\mathbf{A}^\top-\mathbf{I}_R$ are $\lambda_1-1,\cdots,\lambda_R-1$ and those of $\mathbf{A}^\top\mathbf{A}-\mathbf{I}_D$ are $\lambda_1-1,\cdots,\lambda_R-1,-1,\cdots,-1$. Then $\|\mathbf{A}^\top\mathbf{A}-\mathbf{I}_D\|_F^2=\sum_{j=1}^R(\lambda_j-1)^2+\sum_{j=R+1}^D(-1)^2=\|\mathbf{A}\mathbf{A}^\top-\mathbf{I}_R\|_F^2+D-R$. Therefore, the SFN regularizer $\|\mathbf{A}\mathbf{A}^\top-\mathbf{I}_R\|_F^2$ equals to $\|\mathbf{A}^\top\mathbf{A}-\mathbf{I}_D\|_F^2-D+R=\|\mathbf{M}-\mathbf{I}_D\|_F^2-D+R$, where $\mathbf{M}=\mathbf{A}^\top\mathbf{A}$ is a Mahalanobis matrix and $R=\mathrm{rank}(\mathbf{A}^\top\mathbf{A})=\mathrm{rank}(\mathbf{M})$. It is well-known that the trace norm of a matrix is a convex envelope of its rank (Srebro & Shraibman, 2005). We use $\mathrm{tr}(\mathbf{M})$ to approximate $\mathrm{rank}(\mathbf{M})$ and get $\|\mathbf{A}\mathbf{A}^\top-\mathbf{I}_R\|_F^2\approx\|\mathbf{M}-\mathbf{I}_D\|_F^2+\mathrm{tr}(\mathbf{M})-D$, where the right hand side is a convex function. Dropping the constant, we get the convex SFN (CSFN) regularizer defined over $\mathbf{M}$:

$$\widehat{\Omega}_{sfn}(\mathbf{M})=\|\mathbf{M}-\mathbf{I}_D\|_F^2+\mathrm{tr}(\mathbf{M}) \tag{1}$$

A convex VND regularizer Given the eigen-decomposition $\mathbf{A}\mathbf{A}^\top=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ where the eigenvalue Λ_{jj} equals to λ_j, based on the property of the matrix logarithm, we have $\log(\mathbf{A}\mathbf{A}^\top)=\mathbf{U}\widehat{\mathbf{\Lambda}}\mathbf{U}^\top$ where $\widehat{\Lambda}_{jj}=\log\Lambda_{jj}$. Then $(\mathbf{A}\mathbf{A}^\top)\log(\mathbf{A}\mathbf{A}^\top)-(\mathbf{A}\mathbf{A}^\top)=\mathbf{U}(\mathbf{\Lambda}\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda})\mathbf{U}^\top$, where the eigenvalues are $\{\lambda_j\log\lambda_j-\lambda_j\}_{j=1}^R$. Then $\Omega_{vnd}(\mathbf{A})=\sum_{j=1}^R(\lambda_j\log\lambda_j-\lambda_j)+R$. Now we consider a matrix $\mathbf{A}^\top\mathbf{A}+\epsilon\mathbf{I}_D$, where $\epsilon>0$ is a small scalar. Using similar calculation, we have $\Gamma_{vnd}(\mathbf{A}^\top\mathbf{A}+\epsilon\mathbf{I}_D,\mathbf{I}_D)=\sum_{j=1}^R((\lambda_j+\epsilon)\log(\lambda_j+\epsilon)-(\lambda_j+\epsilon))+(D-R)(\epsilon\log\epsilon-\epsilon)+D$. Performing certain algebra (see supplements), we get $\Omega_{vnd}(\mathbf{A})\approx\Gamma_{vnd}(\mathbf{A}^\top\mathbf{A}+\epsilon\mathbf{I}_D,\mathbf{I}_D)+R-D$. Replacing $\mathbf{A}^\top\mathbf{A}$ with $\mathbf{M}$, approximating R with $\mathrm{tr}(\mathbf{M})$ and dropping constant D, we get the convex VND (CVND) regularizer:

$$\begin{aligned}\widehat{\Omega}_{vnd}(\mathbf{M}) &= \Gamma_{vnd}(\mathbf{M}+\epsilon\mathbf{I}_D,\mathbf{I}_D)+\mathrm{tr}(\mathbf{M}) \\ &\propto \mathrm{tr}((\mathbf{M}+\epsilon\mathbf{I}_D)\log(\mathbf{M}+\epsilon\mathbf{I}_D))\end{aligned} \quad (2)$$

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

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

$$\begin{aligned}\widehat{\Omega}_{ldd}(\mathbf{M}) &= \Gamma_{ldd}(\mathbf{M}+\epsilon\mathbf{I}_D,\mathbf{I}_D)-(1+\log\epsilon)\mathrm{tr}(\mathbf{M}) \\ &\propto -\mathrm{logdet}(\mathbf{M}+\epsilon\mathbf{I}_D)+(\log\tfrac{1}{\epsilon})\mathrm{tr}(\mathbf{M})\end{aligned} \quad (3)$$

where the convexity of logdet$(\mathbf{M}+\epsilon\mathbf{I}_D)$ 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 $\Gamma_{ldd}(\mathbf{M},\mathbf{I}) = -\mathrm{logdet}(\mathbf{M}) + \mathrm{tr}(\mathbf{M})$ is applied to encourage the Mahalanobis matrix to be close to the identity matrix. This regularizer requires $\mathbf{M}$ to be full rank; in contrast, by associating a large weight $\log\frac{1}{\epsilon}$ to the trace norm tr($\mathbf{M}$), our CLDD regularizer encourages $\mathbf{M}$ to be low-rank. Since $\mathbf{M}=\mathbf{A}^\top\mathbf{A}$, reducing the rank of $\mathbf{M}$ reduces the number of projection vectors in $\mathbf{A}$.

We discuss the errors in convex approximation, which are from two sources: one is the approximation of $\Omega_\phi(\mathbf{A})$ using $\Gamma_\phi(\mathbf{A}^\top\mathbf{A}+\epsilon\mathbf{I}_D,\mathbf{I}_D)$ where the error is controlled by ϵ and can be arbitrarily small (by setting ϵ 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 $\widehat{\Omega}_\phi(\mathbf{M})$), we relax the non-convex PDML-BMD problems into convex MDML-CBMD formulations by replacing $\|\mathbf{Ax}-\mathbf{Ay}\|_2^2$ with $(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})$ and replacing the non-convex BMD regularizers $\Omega_\phi(\mathbf{A})$ with $\widehat{\Omega}_\phi(\mathbf{M})$:

$$\begin{aligned}\min_{\mathbf{M}\succeq 0}\ & \frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}}(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})+\gamma\widehat{\Omega}_\phi(\mathbf{M}) \\ & +\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}\max(0,\tau-(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y}))\end{aligned} \quad (4)$$

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 $\triangle\mathbf{M}$ 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: $\widetilde{\mathbf{M}} = \mathbf{M} - \eta\triangle\mathbf{M}$, where η is a small stepsize; and (2) applying proximal operators associated with the regularizers $\widetilde{\Omega}_\phi(\mathbf{M})$ to $\widetilde{\mathbf{M}}$. The gradient of the CVND regularizer is $\log(\mathbf{M}+\epsilon\mathbf{I}_D)+\mathbf{I}_D$. To compute $\log(\mathbf{M}+\epsilon\mathbf{I}_D)$, we first perform an eigen-decomposition: $\mathbf{M}+\epsilon\mathbf{I}_D = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, then take the log of every eigenvalue in $\mathbf{\Lambda}$ which gets us a new diagonal matrix $\widetilde{\mathbf{\Lambda}}$, and finally compute $\log(\mathbf{M}+\epsilon\mathbf{I}_D)$ as $\mathbf{U}\widetilde{\mathbf{\Lambda}}\mathbf{U}^\top$. In the CLDD regularizer, the gradient of logdet$(\mathbf{M}+\epsilon\mathbf{I}_D)$ is $(\mathbf{M}+\epsilon\mathbf{I}_D)^{-1}$, which can also be computed by eigen-decomposition. Next, we present the proximal operators.

5.1. Proximal Operators

Given the regularizer $\widetilde{\Omega}_\phi(\mathbf{M})$, the associated proximal operator prox$(\widetilde{\mathbf{M}})$ is defined as: prox$(\widetilde{\mathbf{M}})$ = argmin$_{\mathbf{M}}$ $\frac{1}{2\eta}\|\mathbf{M}-\widetilde{\mathbf{M}}\|_2^2+\gamma\widetilde{\Omega}_\phi(\mathbf{M})$, subject to $\mathbf{M}\succeq 0$. Let $\{\tilde{\lambda}_j\}_{j=1}^D$ be the eigenvalues of $\widetilde{\mathbf{M}}$ and $\{x_j\}_{j=1}^D$ be the eigenvalues of $\mathbf{M}$, then the above problem can be equivalently written as:

$$\begin{aligned}\min_{\{x_j\}_{j=1}^D}\ & \frac{1}{2\eta}\sum_{j=1}^D(x_j-\tilde{\lambda}_j)^2+\gamma\sum_{j=1}^D h_\phi(x_j) \\ s.t.\ & \forall j=1,\cdots,D,\quad x_j\ge 0\end{aligned} \quad (5)$$

where $h_\phi(x_j)$ is a regularizer-specific scalar function. This problem can be decomposed into D independent ones: (P) $\min_{x_j} f(x_j) = \frac{1}{2\eta}(x_j-\tilde{\lambda}_j)^2+\gamma h_\phi(x_j)$, subject to $x_j\ge 0$, for $j=1,\cdots,D$, which can be solved individually.

SFN For SFN where $\widetilde{\Omega}_\phi(\mathbf{M}) = \|\mathbf{M}-\mathbf{I}_D\|_F^2+\mathrm{tr}(\mathbf{M})$ and $h_{sfn}(x_j) = (x_j-1)^2+x_j$, the problem (P) is simply a quadratic programming problem. The optimal solution is $x_j^* = \max(0, \frac{\tilde{\lambda}_j+\eta\gamma}{1+2\eta\gamma})$

VND For VND where $\widetilde{\Omega}_\phi(\mathbf{M}) = \mathrm{tr}((\mathbf{M}+\epsilon\mathbf{I}_D)\log(\mathbf{M}+\epsilon\mathbf{I}_D))$ and $h_\phi(x_j) = (x_j+\epsilon)\log(x_j+\epsilon)$, 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 $\eta\gamma\log(x_j+\epsilon)+x_j+\eta\gamma-\tilde{\lambda}_j = 0$. The root of this equation is: $\eta\gamma\omega(\frac{\epsilon-\eta\gamma+\tilde{\lambda}_j}{\eta\gamma}-\log(\eta\gamma))-\epsilon$, 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, $x_j^* = \mathrm{argmin}_{x_j} f(x_j)$, where $x\in\{\max(\eta\gamma\omega(\frac{\epsilon-\eta\gamma+\tilde{\lambda}_j}{\eta\gamma}-\log(\eta\gamma))-\epsilon,0),0\}$.

LDD For LDD where $\widetilde{\Omega}_\phi(\mathbf{M}) = -\text{logdet}(\mathbf{M}+\epsilon\mathbf{I}_D) + (\log\frac{1}{\epsilon})\text{tr}(\mathbf{M})$ and $h_\phi(x_j) = -\log(x_j+\epsilon)+x_j\log\frac{1}{\epsilon}$, by taking the derivative of $f(x_j)$ w.r.t x_j and setting the derivative to zero, we get a quadratic equation: $x_j^2+ax_j+b=0$, where $a=\epsilon-\tilde{\lambda}_j-\eta\gamma\log\epsilon$ 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, $x_j^* = \text{argmin}_{x_j} f(x_j)$, where $x\in\{\max(\frac{-b+\sqrt{b^2-4ac}}{2a},0), \max(\frac{-b-\sqrt{b^2-4ac}}{2a},0),0\}$.

Computational Complexity In this algorithm, the major computation workload is eigen-decomposion 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 $\mathbf{M}$), hence adding these CBMD regularizes 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: $\mathbf{A}^* = \arg\min_{\mathbf{A}} \ \mathbb{E}_{\mathcal{S},\mathcal{D}}[\frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}}\|\mathbf{Ax}-\mathbf{Ay}\|_2^2 + \frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}\max(0,\tau-\|\mathbf{Ax}-\mathbf{Ay}\|_2^2)+\gamma\Omega_\phi(\mathbf{A})]$. We assume there are K classes, where class k has a distribution p_k and the corresponding expectation is $\boldsymbol{\mu}_k$. Each data sample in $\mathcal{S}$ and $\mathcal{D}$ is drawn from the distribution of one specific class. We define $\xi_k = \mathbb{E}_{\mathbf{x}\sim p_k}[\sup_{\|\mathbf{v}\|_2=1}|\mathbf{v}^\top(\mathbf{x}-\boldsymbol{\mu}_k)|]$ and $\xi=\max_k\xi_k$. Further, we assume $\mathbf{A}^*$ has full rank R (which is the number of the projection vectors), and let $\mathbf{U\Lambda U}^\top$ denote the eigen-decomposition of $\mathbf{A}^*\mathbf{A}^{*\top}$, where $\mathbf{\Lambda}=\text{diag}(\lambda_1,\lambda_2,\cdots\lambda_R)$ 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 $\boldsymbol{\mu}_k$ to characterize this class. We define the Mahalanobis distance between two classes j and k as: $d_{jk} = (\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)^\top\mathbf{A}^{*\top}\mathbf{A}^*(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)$. We define the IF among all classes as:

$$\eta = \frac{\max_{j\neq k} d_{jk}}{\min_{j\neq k} d_{jk}}. \tag{6}$$

The motivation of such a definition is: for two frequent classes, since they have more training examples and hence contributing more in learning $\mathbf{A}^*$, 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 $\max_{j\neq k}\|\boldsymbol{\mu}_j-\boldsymbol{\mu}_k\|_2^2$ and $\min_{j\neq k}\|\boldsymbol{\mu}_j-\boldsymbol{\mu}_k\|_2^2$ and assume $\max_{j,k}\|\boldsymbol{\mu}_j-\boldsymbol{\mu}_k\|_2\leq B_0$. Suppose the regularization parameter γ and distance margin τ are sufficiently large: $\gamma\geq\gamma_0$ and $\tau\geq\tau_0$, where γ_0 and τ_0 depend on $\{p_k\}_{k=1}^K$ and $\{\boldsymbol{\mu}_k\}_{k=1}^K$. If $R\geq K-1$ and $\xi\leq(-B_0+\sqrt{B_0^2+\lambda_{K-1}\beta_{K-1}/(2tr(\mathbf{\Lambda}))})/4$, then we have the following bounds for the IF*[3].

- *For the VND regularizer $\Omega_{vnd}(\mathbf{A}^*)$, if $\Omega_{vnd}(\mathbf{A}^*)\leq 1$, the following bound of the IF η holds:*

  $$\eta \leq Cg(\Omega_{vnd}(\mathbf{A}^*))$$

  where $g(\cdot)$ is an increasing function defined in the following way. Let $f(c)=c^{1/(c+1)}(1+1/c)$, 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 $\Omega_{ldd}(\mathbf{A}^*)$, we have*

  $$\eta \leq 4Ce^{\Omega_{ldd}(\mathbf{A}^*)}$$

As can be seen, the bounds are increasing functions of the BMD regularizers $\Omega_{vnd}(\mathbf{A}^*)$ and $\Omega_{ldd}(\mathbf{A}^*)$. 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 $\mathbf{M}$. Given the sample $\mathcal{S}$ and $\mathcal{D}$ where the total number of data pairs is $m=|S|+|D|$, the empirical error is defined as $\widehat{L}(\mathbf{M}) = \frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}}(\mathbf{x}-$

[1] http://icl.cs.utk.edu/magma/

[2] The analysis of convex BMD regularizers in MDML-CBMD will be left for future work.

[3] Please refer to the supplements for the definition of β_{K-1} and the detailed proof.

$\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})+\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}\max(0,\tau-(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y}))$ and the expected error is $L(\mathbf{M})=\mathbb{E}_{\mathcal{S},\mathcal{D}}[\widehat{L}(\mathbf{M})]$. Let $\widehat{\mathbf{M}}^*$ be optimal matrix learned by minimizing the empirical error: $\widehat{\mathbf{M}}^*=\text{argmin}_{\mathbf{M}}\widehat{L}(\mathbf{M})$. We are interested in how well $\widehat{\mathbf{M}}^*$ performs on unseen data. The performance is measured using generalization error: $\mathcal{E}=L(\widehat{\mathbf{M}}^*)-\widehat{L}(\widehat{\mathbf{M}}^*)$. To incorporate the impact of the CBMD regularizers $\Omega_\phi(\mathbf{M})$, we define the hypothesis class of $\mathbf{M}$ to be $\mathcal{M}=\{\mathbf{M}\succeq 0:\Omega_\phi(\mathbf{M})\leq C\}$. The upper bound C controls the strength of regularization. A smaller C entails stronger promotion of orthogonality. C is controlled by the regularization parameter γ in Eq.(4). Increasing γ reduces C. For different CBMD regularizers, we have the following generalization error bound.

Theorem 2 *Suppose* $\sup_{\|\mathbf{v}\|_2\leq 1,(\mathbf{x},\mathbf{y})\in\mathcal{S}}|\mathbf{v}^\top(\mathbf{x}-\mathbf{y})|\leq B$, *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,*

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

- *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 (GEBs), we can see two major implications. First, CBMD regularizers can effectively control the GEBs. Increasing the strength of CBMD regularization (by enlarging γ) 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[4] and 20-Newsgroups (News)[5]; 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 (SFN) (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) $\mathbf{A}\mathbf{A}^\top=\mathbf{I}$ (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 ℓ_2 norm, ℓ_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 $-\text{logdet}(\mathbf{M})+\text{tr}(\mathbf{M})$ (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

[4] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[5] http://qwone.com/~jason/20Newsgroups/

[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.

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.

	MIMIC			EICU			Reuters			News	Cars	Birds	Act
	A-All	A-IF	BS	A-All	A-IF	BS	A-All	A-IF	BS	A-All	A-All	A-All	A-All
PDML	0.634	0.608	0.070	0.671	0.637	0.077	0.949	0.916	0.049	0.757	0.714	0.851	0.949
MDML	0.641	0.617	0.064	0.677	0.652	0.055	0.952	0.929	0.034	0.769	0.722	0.855	0.952
LMNN	0.628	0.609	0.054	0.662	0.633	0.066	0.943	0.913	0.040	0.731	0.728	0.832	0.912
LDML	0.619	0.594	0.068	0.667	0.647	0.046	0.934	0.906	0.042	0.748	0.706	0.847	0.937
MLEC	0.621	0.605	0.045	0.679	0.656	0.053	0.927	0.916	0.021	0.761	0.725	0.814	0.917
GMML	0.607	0.588	0.053	0.668	0.648	0.045	0.931	0.905	0.035	0.738	0.707	0.817	0.925
ILHD	0.577	0.560	0.051	0.637	0.610	0.064	0.905	0.893	0.028	0.711	0.686	0.793	0.898
MDML-ℓ_2	0.648	0.627	0.055	0.695	0.676	0.042	0.955	0.930	0.037	0.774	0.728	0.872	0.958
MDML-ℓ_1	0.643	0.615	0.074	0.701	0.677	0.053	0.953	0.948	0.020	0.791	0.725	0.868	0.961
MDML-$\ell_{2,1}$	0.646	0.630	0.043	0.703	0.661	0.091	0.963	0.936	0.035	0.783	0.728	0.861	0.964
MDML-Tr	0.659	0.642	0.044	0.696	0.673	0.051	0.961	0.934	0.036	0.785	0.731	0.875	0.955
MDML-IT	0.653	0.626	0.070	0.692	0.668	0.053	0.954	0.920	0.046	0.771	0.724	0.858	0.967
MDML-Drop	0.647	0.630	0.045	0.701	0.670	0.067	0.959	0.937	0.032	0.787	0.729	0.864	0.962
MDML-OS	0.649	0.626	0.059	0.689	0.679	0.045	0.957	0.938	0.031	0.779	0.732	0.869	0.963
PDML-DC	0.652	0.639	0.035	0.706	0.686	0.044	0.962	0.943	0.034	0.773	0.736	0.882	0.964
PDML-CS	0.661	0.641	0.053	0.712	0.670	0.089	0.967	0.954	0.020	0.803	0.742	0.895	0.971
PDML-DPP	0.659	0.632	0.069	0.714	0.695	0.041	0.958	0.937	0.036	0.797	0.751	0.891	0.969
PDML-IC	0.660	0.642	0.047	0.711	0.685	0.057	0.972	0.954	0.030	0.801	0.740	0.887	0.967
PDML-DeC	0.648	0.625	0.063	0.698	0.675	0.050	0.965	0.960	0.017	0.786	0.728	0.860	0.958
PDML-VGF	0.657	0.634	0.059	0.718	0.697	0.045	0.974	0.952	0.036	0.806	0.747	0.894	**0.974**
PDML-MA	0.659	0.644	0.040	0.721	0.703	0.038	0.975	0.959	0.024	0.815	0.743	0.898	0.968
PDML-OC	0.651	0.636	0.041	0.705	0.685	0.043	0.955	0.931	0.036	0.779	0.727	0.875	0.956
PDML-OS	0.639	0.614	0.067	0.675	0.641	0.072	0.951	0.928	0.038	0.764	0.716	0.855	0.950
PDML-SFN	0.662	0.642	0.051	0.724	0.701	0.045	0.973	0.947	0.038	0.808	0.749	0.896	0.970
PDML-VND	0.667	0.655	0.032	0.733	0.706	0.057	0.976	0.971	0.012	0.814	0.754	0.902	0.972
PDML-LDD	0.664	0.651	0.035	0.731	0.711	0.043	0.973	0.964	0.017	0.816	0.751	0.904	0.971
MDML-CSFN	0.668	0.653	0.039	0.728	0.705	0.049	0.978	0.968	0.023	0.813	0.753	0.905	0.972
MDML-CVND	**0.672**	**0.664**	**0.022**	0.735	0.718	**0.035**	**0.984**	**0.982**	0.012	**0.822**	0.755	0.908	0.973
MDML-CLDD	0.669	0.658	0.029	**0.739**	**0.719**	0.042	0.981	0.980	**0.011**	0.819	**0.759**	**0.913**	0.971

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

	MIMIC	EICU	Reuters	News	Cars	Birds	Act
PDML	62.1	66.6	5.2	11.0	8.4	10.1	3.4
MDML	3.4	3.7	0.3	0.6	0.5	0.6	0.2
PDML-DC	424.7	499.2	35.2	65.6	61.8	66.2	17.2
PDML-CS	263.2	284.8	22.6	47.2	34.5	42.8	14.4
PDML-DPP	411.8	479.1	36.9	61.9	64.2	70.5	16.5
PDML-IC	265.9	281.2	23.4	46.1	37.5	45.2	15.3
PDML-DeC	458.5	497.5	41.8	78.2	78.9	80.7	19.9
PDML-VGF	267.3	284.1	22.3	48.9	35.8	38.7	15.4
PDML-MA	271.4	282.9	23.6	50.2	30.9	39.6	17.5
PDML-OC	104.9	118.2	9.6	14.3	14.8	17.0	3.9
PDML-SFN	261.7	277.6	22.9	46.3	36.2	38.2	15.9
PDML-VND	401.8	488.3	33.8	62.5	67.5	73.4	17.1
PDML-LDD	407.5	483.5	34.3	60.1	61.8	72.6	17.9
MDML-CSFN	41.1	43.9	3.3	7.3	6.5	6.9	1.8
MDML-CVND	43.8	46.2	3.6	8.1	6.9	7.8	2.0
MDML-CLDD	41.7	44.5	3.4	7.5	6.6	7.2	1.8

with a different initialization, and picks the one yielding the lowest objective value. In total, the number of runs is $72 \times 5 = 360$. For the MDML-based methods, there is no need to restart multiple times or tune the NPVs. The total number of runs is $4 \times 3 = 12$. As can be seen from the table, the proposed convex methods are much faster than the non-convex ones, due to the greatly reduced number of experimental runs, although for each single run the convex methods are less efficient than the non-convex methods due to the overhead of eigen-decomposition. The unregularized MDML takes the least time to train since it has no parameters to tune and runs only once. On average, the time of each single run in MDML-(CSFN,CVND,CLDD) is close to that in the unregularized MDML, since an eigen-decomposition is required anyway regardless of the presence of the regularizers.

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 $|\frac{\text{A-IF}}{\text{A-F}} - 1|$. 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. As a result, these vectors improve the performance on infrequent classes and lead to better balancedness. Thanks to their convexity nature, our methods can achieve the global optimal solution and outperform the non-convex ones that can only achieve a local optimal and hence a sub-optimal solution. Comparing (PDML,MDML)-OS with the unregularized PDLM/MDML, we can see that over-sampling indeed improves balancedness. However, this improvement is less significant than that achieved by our meth-

Table 3. Number of projection vectors (NPV) and compactness score (CS, $\times 10^{-3}$).

	MIMIC		EICU		Reuters		News		Cars		Birds		Act	
	NPV	CS	NPV	CS	NPV	CS	NPV	CS	NPV	CS	NPV	CS	NPV	CS
PDML	300	2.1	400	1.7	300	3.2	300	2.5	300	2.4	500	1.7	200	4.7
MDML	247	2.6	318	2.1	406	2.3	336	2.3	376	1.9	411	2.1	168	5.7
LMNN	200	3.1	400	1.7	400	2.4	300	2.4	400	1.8	500	1.7	300	3.0
LDML	300	2.1	400	1.7	400	2.3	200	3.7	300	2.4	400	2.1	300	3.1
MLEC	487	1.3	493	1.4	276	3.4	549	1.4	624	1.2	438	1.9	327	2.8
GMML	1000	0.6	1000	0.7	1000	0.9	1000	0.7	1000	0.7	1000	0.8	1000	0.9
ILHD	100	5.8	100	6.4	**50**	18.1	100	7.1	100	6.9	100	7.9	**50**	18.0
MDML-ℓ_2	269	2.4	369	1.9	374	2.6	325	2.4	332	2.2	459	1.9	179	5.4
MDML-ℓ_1	341	1.9	353	2.0	417	2.3	317	2.5	278	2.6	535	1.6	161	6.0
MDML-$\ell_{2,1}$	196	3.3	251	2.8	288	3.3	316	2.5	293	2.5	326	2.6	135	7.1
MDML-Tr	148	4.5	233	3.0	217	4.4	254	3.1	114	6.4	286	3.1	129	7.4
MDML-IT	1000	0.7	1000	0.7	1000	1.0	1000	0.8	1000	0.7	1000	0.9	1000	1.0
MDML-Drop	183	3.5	284	2.5	315	3.0	251	3.1	238	3.1	304	2.8	147	6.5
PDML-DC	100	6.5	300	2.4	100	9.6	200	3.9	200	3.7	300	2.9	100	9.6
PDML-CS	200	3.3	200	3.6	200	4.8	100	8.0	100	7.4	200	4.5	**50**	**19.4**
PDML-DPP	100	6.6	200	3.6	100	9.6	100	8.0	200	3.8	200	4.5	100	9.7
PDML-IC	200	3.3	200	3.6	200	4.9	100	8.0	200	3.7	100	8.9	100	9.7
PDML-DeC	200	3.2	300	2.3	200	4.8	200	3.9	200	3.6	200	4.3	100	9.6
PDML-VGF	200	3.3	200	3.6	200	4.9	100	8.1	200	3.7	200	4.5	100	9.7
PDML-MA	200	3.3	200	3.6	100	9.8	100	8.2	100	7.4	200	4.5	**50**	**19.4**
PDML-SFN	100	6.6	200	3.6	100	9.7	100	8.1	100	7.5	200	4.5	**50**	**19.4**
PDML-OC	100	6.5	100	7.1	**50**	19.1	50	15.6	100	7.3	100	8.8	**50**	19.1
PDML-VND	100	6.7	100	7.3	**50**	**19.5**	100	8.1	100	7.5	100	9.0	**50**	**19.4**
PDML-LDD	100	6.6	200	3.7	100	9.7	100	8.2	100	7.5	100	9.0	**50**	**19.4**
MDML-CSFN	143	4.7	209	3.5	174	5.6	87	9.3	**62**	**12.1**	139	6.5	64	15.2
MDML-CVND	**53**	**12.7**	**65**	**11.3**	61	16.0	63	13.0	127	5.9	92	9.9	68	14.3
MDML-CLDD	76	8.8	128	5.8	85	11.5	**48**	**17.1**	91	8.3	**71**	**12.9**	55	17.7

ods. 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) outperform popular DML approaches including LMNN, LDML, MLEC and GMML, demonstrating their competitive standing in the DML literature.

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 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

Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *Ambient assisted living and home care*. Springer, 2012.

Bao, Y., Jiang, H., Dai, L., and Liu, C. Incoherent training of deep neural networks to decorrelate bottleneck features for speech recognition. In *ICASSP*, 2013.

Bellet, A. and Habrard, A. Robustness and generalization for metric learning. *Neurocomputing*, 2015.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Candes, E. and Recht, B. Exact matrix completion via convex optimization. *CACM*, 2012.

Carreira-Perpinán, M. A. and Raziperchikolaei, R. An ensemble diversity approach to supervised binary hashing. In *NIPS*, 2016.

Chen, Y., Zhang, H., Tong, Y., and Lu, M. Diversity regularized latent semantic match for hashing. *Neurocomputing*, 2017.

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. Reducing overfitting in deep networks by decorrelating representations. *ICLR*, 2015.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. Information-theoretic metric learning. In *ICML*, 2007.

Fu, X., McCane, B., Mills, S., and Albert, M. Nokmeans: Non-orthogonal k-means hashing. In *ACCV*, 2014.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics*, 2012.

Ge, T., He, K., Ke, Q., and Sun, J. Optimized product quantization. *TPAMI*, 2014a.

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

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000.

Gong, Y., Lazebnik, S., Gordo, A., and Perronnin, F. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *TPAMI*, pp. 2916–2929, 2013.

Gorenflo, R., Luchko, Y., and Mainardi, F. Analytical properties and applications of the wright function. *arXiv preprint math-ph/0701069*, 2007.

Guillaumin, M., Verbeek, J., and Schmid, C. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

Jalali, A., Xiao, L., and Fazel, M. Variational gram functions: Convex analysis and optimization. *arXiv preprint arXiv:1507.04734*, 2015.

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

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

Kong, W. and Li, W.-J. Isotropic hashing. In *NIPS*, 2012.

Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013.

Kulesza, A., Taskar, B., et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Kulis, B., Sustik, M. A., and Dhillon, I. S. Low-rank kernel learning with bregman matrix divergences. *JMLR*, 2009.

Kulis, B. et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

Liu, W., Hoi, S., and Liu, J. Output regularized metric learning with side information. *ECCV*, 2008.

Liu, W., Mu, C., Ji, R., Ma, S., Smith, J. R., and Chang, S.-F. Low-rank similarity metric learning in high dimensions. In *AAAI*, 2015.

Malkin, J. and Bilmes, J. Ratio semi-definite classifiers. In *ICASSP*, 2008.

Manning, C. D., Raghavan, P., Schütze, H., et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

Nielsen, M. A. and Chuang, I. L. *Quantum computation and Quantum information*. 2000.

Niu, G., Dai, B., Yamada, M., and Sugiyama, M. Information-theoretic semisupervised metric learning via entropy regularization. *Neural Computation*, 2012.

Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

Qi, G.-J., Tang, J., Zha, Z.-J., Chua, T.-S., and Zhang, H.-J. An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization. In *ICML*, 2009.

Qian, Q., Hu, J., Jin, R., Pei, J., and Zhu, S. Distance metric learning using dropout: a structured regularization approach. In *KDD*, 2014.

Raziperchikolaei, R. and Carreira-Perpinán, M. A. Learning independent, diverse binary hash functions: Pruning and locality. *ICDM*, 2016.

Srebro, N. and Shraibman, A. Rank, trace-norm and max-norm. In *COLT*, 2005.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

Tsuda, K., Rätsch, G., and Warmuth, M. K. Matrix exponentiated gradient updates for on-line learning and bregman projection. *JMLR*, 2005.

Verma, N. and Branson, K. Sample complexity of learning mahalanobis distance metrics. *arXiv preprint arXiv:1505.02729*, 2015.

Wang, D., Cui, P., Ou, M., and Zhu, W. Deep multimodal hashing with orthogonal regularization. In *IJCAI*, 2015.

Wang, F. and Sun, J. Survey on distance metric learning and dimensionality reduction in data mining. *Data Mining and Knowledge Discovery*, 2015.

Wang, J., Kumar, S., and Chang, S.-F. Semi-supervised hashing for large-scale search. *TPAMI*, 2012.

Weinberger, K. Q., Blitzer, J., and Saul, L. K. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005.

Weiss, Y., Torralba, A., and Fergus, R. Spectral hashing. In *NIPS*, 2009.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-ucsd birds 200. 2010.

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

Xie, P., Deng, Y., and Xing, E. P. Diversifying restricted boltzmann machine for document modeling. In *KDD*, 2015.

Xie, P., Poczos, B., and Xing, E. P. Near-orthogonality regularization in kernel methods. *UAI*, 2017.

Xing, E. P., Jordan, M. I., Russell, S., and Ng, A. Y. Distance metric learning with application to clustering with side-information. In *NIPS*, 2002.

Yao, W., Weng, Z., and Zhu, Y. Diversity regularized metric learning for person re-identification. In *ICIP*, 2016.

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

Ying, Y., Huang, K., and Campbell, C. Sparse metric learning via smooth optimization. In *NIPS*, 2009.

Yu, Y., Li, Y.-F., and Zhou, Z.-H. Diversity regularized machine. In *IJCAI*, 2011.

Zadeh, P. H., Hosseini, R., and Sra, S. Geometric mean metric learning. In *ICML*, 2016.

Zhou, D., Xiao, L., and Wu, M. Hierarchical classification via orthogonal transfer. *ICML*, 2011.

Zou, J. Y. and Adams, R. P. Priors for diversity in generative latent variable models. In *NIPS*, 2012.