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

Pengtao Xie[†*], Wei Wu[*], Yichen Zhu[§] and Eric P. Xing[†]
[†]Petuum Inc, USA
[*]School of Computer Science, Carnegie Mellon University, USA
[§]School of Mathematical Sciences, Peking University, China

Abstract

In this supplement material, we give more details on (1) related works, (2) the derivation of convex approximations, (3) proof of theorems, and (4) additional experimental settings and results.

1 Related Works

1.1 Distance Metric Learning

Many studies [1, 2, 3, 4, 5, 6, 7] have investigated DML. Please refer to [8, 9] for a detailed review. Xing et al. [1] learn a Mahalanobis distance by minimizing the sum of distances of all similar data pairs subject to the constraint that the sum of all dissimilar pairs is no less than 1. Weinberger et al. [2] propose large margin metric learning, which is applied for k-nearest neighbor classification. For each data example $\mathbf{x}_i$, they first obtain l nearest neighbors based on Euclidean distance. Then among the l neighbors, some (denoted by $\mathbf{S} = \{\mathbf{x}_j\}$) have the same class label with $\mathbf{x}_i$ and others (denoted by $D = \{\mathbf{x}_k\}$) do not. Then learn a projection matrix $\mathbf{L}$ such that $\|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_k)\|_2^2 - \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \geq 1$ where $\mathbf{x}_j \in \mathbf{S}$ and $\mathbf{x}_k \in \mathbf{D}$. Davis et al. [3] learn a Mahalanobis distance such that the distance between similar pairs is no more than a threshold s and the distance between dissimilar pairs is no greater than a threshold t. Guillaumin et al. [4] define a conditional probability of the similarity/dissimilarity label conditioned on the Mahalanobis distance: $p(y_{ij}|(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)) = 1/(1 + \exp((\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_i)))$, where the binary variable $y_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same class label. $\mathbf{M}$ is learned by maximizing the conditional likelihood of the training data. Kostinger et al. [6] learn a Mahalanobis distance metric from equivalence constraints based on likelihood ratio test. The Mahalanobis matrix is computed in one shot, without going through an iterative optimization procedure. Ying and Li [5] formulate DML as an eigenvalue optimization problem. Zadeh et al. [7] propose a geometric mean metric learning approach, based on the Riemannian geometry of positive definite matrices. Similar to [6], the Mahalanobis matrix has a closed form solution without iterative optimization.

To avoid overfitting in DML, various regularization approaches have been explored. Davis et al. [3] regularize the Mahalanobis matrix to be close to another matrix that encodes prior information, where the closeness is measured using log-determinant divergence. Qi et al. [10] use ℓ_1 regularization to learn sparse distance metrics for high-dimensional, small-sample problems. Ying et al. [11] use $\ell_{2,1}$ norm to simultaneously encourage low-rankness and sparsity. Trace norm is leveraged to encourage low-rankness in [12, 13]. Qian et al. [14] apply dropout to DML. Many works [15, 16, 17, 18, 19, 20] study diversity-promoting regularization in DML or hashing. They define regularizers based on squared Frobenius norm [15, 21, 16, 20] or angles [17, 18] to encourage the projection vectors to approach orthogonal. Several works [22, 23, 24, 25, 26] impose strict orthogonal constraint on the projection vectors. As observed in previous works [15, 21] and our experiments, strict orthogonality hurts performance. Isotropic hashing [27, 28] encourages the variances of different projected dimensions to be equal to achieve balance. Carreira-Perpinán and Raziperchikolaei [29] propose a diversity hashing method which first trains hash functions independently and then introduces diversity among them based on classifier ensembles.

1.2 Orthogonality-Promoting Regularization

Orthogonality-promoting regularization has been studied in other problems as well, including ensemble learning, latent variable modeling, classification and multitask learning. In ensemble learning, many studies [30, 31, 32, 33] promote orthogonality among the coefficient vectors of base classifiers or regressors, with the aim to improve generalization performance and reduce computational complexity. Recently, several works [34, 35, 36, 37] study orthogonality-promoting regularization of latent variable models (LVMs), which encourages the components in LVMs to be mutually orthogonal, for the sake of capturing infrequent patterns and reducing the number of components without sacrificing modeling power. In these works, various orthogonality-promoting regularizers have been proposed, based on Determinantal Point Process [38, 34] and cosine similarity [33, 35, 37]. In multi-way classification, Malkin and Bilmes [39] propose to use the determinant of a covariance matrix to encourage orthogonality among classifiers. Jalali et al. [40] propose a class of *variational Gram functions* (VGFs) to promote pairwise orthogonality among vectors. While these VGFs are convex, they can only be applied to non-convex DML formulations. As a result, the overall regularized DML is non-convex and is not amenable for convex relaxation.

In the sequel, we review two families of orthogonality-promoting regularizers.

Determinantal Point Process [34] employed the Determinantal Point Process (DPP) [38] as a prior to induce orthogonality in latent variable models. DPP is defined over K vectors: $p(\{\mathbf{a}_i\}_{i=1}^K) \propto \det(\mathbf{L})$, where $\mathbf{L}$ is a $K \times K$ kernel matrix with $L_{ij} = k(\mathbf{a}_i, \mathbf{a}_j)$ and $k(\cdot,\cdot)$ as a kernel function. $\det(\cdot)$ denotes the determinant of a matrix. A configuration of $\{\mathbf{a}_i\}_{i=1}^K$ with larger probability is deemed to be more orthogonal. The underlying intuition is that: $\det(\mathbf{L})$ represents the volume of the parallelepiped formed by vectors in the kernel-induced feature space. If these vectors are closer to being orthogonal, the volume is larger, which results in a larger $p(\{\mathbf{a}_i\}_{i=1}^K)$. The shortcoming of DPP is that it is sensitive to vector scaling. Enlarging the magnitudes of vectors results in larger volume, but does not essentially affects the orthogonality of vectors.

Pairwise Cosine Similarity Several works define orthogonality-promoting regularizers based on the pairwise cosine similarity among component vectors: if the cosine similarity scores are close to zero, then the components are closer to being orthogonal. Given K component vectors, the cosine similarity s_{ij} between each pair of components $\mathbf{a}_i$ and $\mathbf{a}_j$ is computed: $s_{ij} = \mathbf{a}_i \cdot \mathbf{a}_j / (\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2)$. Then these scores are aggregated as a single score. In [33], these scores are aggregated as $\sum_{1\leq i<j\leq K}(1 - s_{ij})$. In [35], the aggregation is performed as $-\log(\frac{1}{K(K-1)}\sum_{1\leq i<j\leq K}\beta|s_{ij}|)^{\frac{1}{\beta}}$ where $\beta > 0$. In [37], the aggregated score is defined as mean of $\arccos(|s_{ij}|)$ minus the variance of $\arccos(|s_{ij}|)$.

2 Convex Approximations of BMD Regularizers

Approximation of VND regularizer Given $\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, according to the property of matrix logarithm, $\log(\mathbf{A}\mathbf{A}^\top) = \mathbf{U}\widehat{\mathbf{\Lambda}}\mathbf{U}^\top$, where $\widehat{\Lambda}_{jj} = \log\lambda_j$. 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 $\{\lambda_j \log\lambda_j - \lambda_j\}_{j=1}^R$. Since $\mathrm{tr}(\mathbf{M}) = \sum_{j=1}^R \lambda_j$, we have $\Omega_{vnd}(\mathbf{A}) = \sum_{j=1}^R(\lambda_j\log\lambda_j - \lambda_j) + R$. Now we consider a matrix $\mathbf{A}^\top\mathbf{A} + \epsilon\mathbf{I}_D$, where $\epsilon > 0$ is a small scalar. The eigenvalues of this matrix are $\lambda_1 + \epsilon, \cdots, \lambda_R + \epsilon, \epsilon, \cdots, \epsilon$. Then we have

$$\begin{aligned}
&\Gamma_{vnd}(\mathbf{A}^\top\mathbf{A} + \epsilon\mathbf{I}_D, \mathbf{I}_D)\\
&= \mathrm{tr}((\mathbf{A}^\top\mathbf{A} + \epsilon\mathbf{I}_D)\log(\mathbf{A}^\top\mathbf{A} + \epsilon\mathbf{I}_D) - (\mathbf{A}^\top\mathbf{A} + \epsilon\mathbf{I}_D)) + D\\
&= \textstyle\sum_{j=1}^R((\lambda_j + \epsilon)\log(\lambda_j + \epsilon) - (\lambda_j + \epsilon)) + \sum_{j=R+1}^D(\epsilon\log\epsilon - \epsilon) + D\\
&= \textstyle\sum_{j=1}^R((\lambda_j + \epsilon)(\log\lambda_j + \log(1 + \frac{\epsilon}{\lambda_j})) - (\lambda_j + \epsilon)) + (D - R)(\epsilon\log\epsilon - \epsilon) + D\\
&= \textstyle\sum_{j=1}^R(\lambda_j\log\lambda_j - \lambda_j + \lambda_j\log(1 + \frac{\epsilon}{\lambda_j}) + \epsilon(\log\lambda_j + \log(1 + \frac{\epsilon}{\lambda_j})) - \epsilon) + (D - R)(\epsilon\log\epsilon - \epsilon) + D\\
&= \textstyle\Omega_{vnd}(\mathbf{A}) - R + \sum_{j=1}^R(\lambda_j\log(1 + \frac{\epsilon}{\lambda_j}) + \epsilon(\log\lambda_j + \log(1 + \frac{\epsilon}{\lambda_j})) - \epsilon) + (D - R)(\epsilon\log\epsilon - \epsilon) + D
\end{aligned} \tag{1}$$

Since ϵ is small, we have $\log(1+\frac{\epsilon}{\lambda_j}) \approx \frac{\epsilon}{\lambda_j}$. Then $\lambda_j \log(1+\frac{\epsilon}{\lambda_j}) \approx \epsilon$ and the last line in the above equation can be approximated with $\Omega_{vnd}(\mathbf{A}) - R + D + O(\epsilon)$, and therefore

$$\Omega_{vnd}(\mathbf{A}) \approx \Gamma_{vnd}(\mathbf{A}^\top \mathbf{A} + \epsilon \mathbf{I}_D, \mathbf{I}_D) + R - D \tag{2}$$

where $O(\epsilon)$ is small since ϵ is small, and is hence dropped.

Approximation of LDD regularizer

$$\begin{aligned}
&\Gamma_{ldd}(\mathbf{A}^\top \mathbf{A} + \epsilon \mathbf{I}_D, \mathbf{I}_D) \\
&= \sum_{j=1}^{R} \lambda_j + D\epsilon - (D-R)\log\epsilon - \sum_{j=1}^{R} \log(\lambda_j + \epsilon) \\
&= \sum_{j=1}^{R} \lambda_j + D\epsilon - (D-R)\log\epsilon - \sum_{j=1}^{R} (\log\lambda_j + \log(1+\tfrac{\epsilon}{\lambda_j})) \\
&\approx \sum_{j=1}^{R} (\lambda_j - \log\lambda_j) + R\log\epsilon - \epsilon \sum_{j=1}^{R} \tfrac{1}{\lambda_j} + D\epsilon - D\log\epsilon \\
&= \Omega_{ldd}(\mathbf{A}) + R + R\log\epsilon + O(\epsilon) - D\log\epsilon
\end{aligned} \tag{3}$$

Dropping $O(\epsilon)$, we obtain

$$\Omega_{ldd}(\mathbf{A}) = \Gamma_{ldd}(\mathbf{A}^\top \mathbf{A} + \epsilon \mathbf{I}_D, \mathbf{I}_D) - (\log\epsilon + 1)R + D\log\epsilon \tag{4}$$

3 Comments on Proximal SSD and Projected SSD

Note that one can also solve the MDML-CBMD problems using projected stochastic subgradient descent (SSD): (1) sampling a minibatch of data pairs and computing sub-gradient of the combined objective function which is the sum of the data-dependent loss defined over the minibatch and the regularizer; (2) updating $\mathbf{M}$ using subgradient descent; (3) projecting the updated $\mathbf{M}$ onto the positive semidefinite cone. We choose to use proximal SSD because its complexity only depends on the Lipschitz constant of the subgradient of the data-dependent loss, whereas that of projected SSD would depend on the sum of the Lipschitz constants of the data-dependent loss and the regularizer.

4 Proof of Theorem 1

4.1 Proof Sketch

We make the following two assumptions.

- The size of similar and dissimilar set $|\mathcal{S}|$ and $|\mathcal{D}|$ are fixed.
- $\mathbf{A}^*$ has full row rank R.

Denote the K classes as $\mathcal{C}_1, \mathcal{C}_2, \cdots \mathcal{C}_K$. The probability that a sample is drawn from the kth class is p_k, and $\sum_{k=1}^{K} p_k = 1$. Denote the class membership of example $\mathbf{x}$ as $c(\mathbf{x})$. Denote the probability that $\mathbf{x} \in \mathcal{C}_j, \mathbf{y} \in \mathcal{C}_k$ where $(\mathbf{x},\mathbf{y}) \in \mathcal{D}$ as $p_{jk} = p_j p_k/(1-\sum_{l=1}^{K} p_l^2)$. Define the SVD of matrix $\mathbf{A}^*$ as $\mathbf{U}\sqrt{\mathbf{\Lambda}}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{R\times R}$, $\mathbf{\Lambda} \in \mathbb{R}^{R\times R}$, and $\mathbf{V} \in \mathbb{R}^{D\times R}$. $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots \lambda_R)$. then $\mathbf{A}^{*\top}\mathbf{A}^* = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$. Denote $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_R]$. Then $\forall \mathbf{z} = \mathbf{x} - \mathbf{y}$, we have $\mathbf{z}^\top \mathbf{A}^{*\top}\mathbf{A}^*\mathbf{z} = \sum_{r=1}^{R} \lambda_r (\mathbf{v}_r^\top \mathbf{z})^2$. We see $\mathbf{z}^\top \mathbf{A}^{*\top}\mathbf{A}^*\mathbf{z}$ can be written as a sum of R terms. Inspired by this, we define a vector function $\alpha(\cdot)$ as $\alpha(\mathbf{u}) = \sum_{j,k=1}^{K} p_{jk}(\mathbf{u}^\top(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k))^2$. This function measures the weighted sum of $(\mathbf{u}^\top(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k))^2$ across all classes. Define $\mathcal{G} = \text{span}\{\boldsymbol{\mu}_j - \boldsymbol{\mu}_k : j \neq k\}$.

Definition 1 (feature values and feature vectors)**.** *For a linear space $\mathcal{G}$, define vectors $\mathbf{w}_1, \mathbf{w}_2, \cdots \mathbf{w}_{K-1}$ and positive real numbers $\beta_1, \beta_2, \cdots \beta_{K-1}$ as*

$$\mathbf{w}_1 = \arg\min_{\|\mathbf{u}\|=1, \mathbf{u}\in\mathcal{G}} \alpha(\mathbf{u}), \quad \beta_1 = \alpha(\mathbf{w}_1),$$

$$\mathbf{w}_r = \arg \min_{\substack{\|\mathbf{u}\|=1,\, \mathbf{u}\in\mathcal{G} \\ \mathbf{u}\perp\mathbf{w}_j, \forall j<r}} \alpha(\mathbf{u}), \quad \beta_r = \alpha(\mathbf{w}_r), \quad \forall r > 1$$

$\forall r > K-1$, define $\beta_r = 0$, and $\mathbf{w}_r$ as an arbitrary vector which has norm 1 and is orthogonal to $\mathbf{w}_1, \mathbf{w}_2, \cdots \mathbf{w}_{r-1}$. $\mathbf{w}_1, \mathbf{w}_2, \cdots$ are called feature vectors of $\mathcal{G}$, and $\beta_1, \beta_2, \cdots$ are called feature values of $\mathcal{G}$.

We give a condition for the regularizers.

Condition 1. *For a regularizer $\Omega_\phi(\cdot)$, there exists a unique matrix function $\varphi(\cdot)$ such that for any $\mathbf{A}^*$,*

$$\Omega_\phi(\mathbf{A}^*) = \varphi(\mathbf{A}^*\mathbf{A}^{*\top}) = \varphi(\mathbf{\Lambda}).$$

The VND and LDD regularizer satisfy this condition. For the VND regularizer, $\varphi(\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{\Lambda}\log\mathbf{\Lambda} - \mathbf{\Lambda}) + R$; for the LDD regularizer, $\varphi(\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{\Lambda}) - \log\det(\mathbf{\Lambda}) - R$. The SFN regularizer does not satisfy this condition.

Now we have enough preparation to give the following lemma. It shows that the linear space $\mathcal{G}$ can be recovered if the second moment of noise is smaller than a certain value.

Lemma 1. *Suppose $R \geq K-1$, $\max_{j\in k}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|_2 \leq B_0$, and the regularization parameter γ and distance margin τ satisfy $\gamma \geq \gamma_0, \tau \geq \tau_0$. If $\xi \leq \frac{-B_0+\sqrt{B_0^2+\gamma_{K-1}\beta_{K-1}/(2\mathrm{tr}(\mathbf{\Lambda}))}}{4}$, then*

$$\mathcal{G} \subset \mathrm{span}(\mathbf{A}^{*\top}). \tag{5}$$

Here $\mathrm{span}(\mathbf{A}^{*\top})$ denotes the column space of matrix $\mathbf{A}^{*\top}$. Both λ_0 and τ_0 depend on $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots \boldsymbol{\mu}_K$ and $p_1, p_2, \cdots p_K$.

The next lemma shows that if Eq.(5) holds, we can bound the imbalance factor η with the condition number of $\mathbf{A}^*\mathbf{A}^{*\top}$ (denoted by $\mathrm{cond}(\mathbf{A}^*\mathbf{A}^{*\top})$). Note that the BMD regularizers $\Omega_\phi(\mathbf{A}^*)$ encourage $\mathbf{A}^*\mathbf{A}^{*\top}$ to be close to an identity matrix, i.e., encouraging the condition number to be close to 1.

Lemma 2. *If Eq.(5) holds, and there exists a real function g such that*

$$\mathrm{cond}(\mathbf{A}^*\mathbf{A}^{*\top}) \leq g(\Omega_\phi(\mathbf{A}^*)),$$

then we have the following bound for the imbalance factor

$$\eta \leq g(\Omega_\phi(\mathbf{A}^*)) \frac{\max_{j\neq k}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|^2}{\min_{j\neq k}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|^2}.$$

Next, we derive the explicit forms of g for the VND and LDD regularizers.

Lemma 3. *For the VND regularizer $\Omega_{vnd}(\mathbf{A}^*)$, define $f(c) = c^{1/(c+1)}(1+1/c)$, then $f(c)$ is strictly increasing on $(0,1]$ and strictly decreasing on $[1,\infty)$. Define the inverse function of $f(\cdot)$ on $[1,\infty)$ as $f^{-1}(\cdot)$. Then if $\Omega_{vnd}(\mathbf{A}^*) < 1$, we have*

$$\mathrm{cond}(\mathbf{A}^*\mathbf{A}^{*\top}) \leq f^{-1}(2 - \Omega_{vnd}(\mathbf{A}^*)).$$

For the LDD regularizer $\Omega_{ldd}(\mathbf{A}^*)$, we have

$$\mathrm{cond}(\mathbf{A}^*\mathbf{A}^{*\top}) \leq 4e^{\Omega_{ldd}(\mathbf{A}^*)}.$$

Combining Lemma 1, 2 and 3, we finish the proof of Theorem 1 in the main paper.

4.2 Proof of Lemma 1

In order to prove Lemma 1, we first need some auxiliary lemmas on the properties of the function $\alpha(\cdot)$. Denote $\boldsymbol{\mu}_{jk} = \boldsymbol{\mu}_j - \boldsymbol{\mu}_k, \forall j \neq k$.

Lemma 4. *Suppose* $\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_r$ *and* $\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_r$ *are two sets of standard orthogonal vectors in* $\mathbb{R}^d$*, and* $\operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_r) = \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_r)$*, then we have*

$$\sum_{l=1}^{r} \alpha(\mathbf{u}_l) = \sum_{l=1}^{r} \alpha(\mathbf{v}_l).$$

Proof. By the definition of these two sets of vectors, there exists a $r \times r$ standard orthogonal matrix $\mathbf{B} = (b_{jk})$, such that $(\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_r) = (\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_r)\mathbf{B}$. Then we have

$$\begin{aligned}\sum_{l=1}^{r} \alpha(\mathbf{u}_l) =& \sum_{l=1}^{r} \sum_{j \neq k} p_{jk} ((\sum_{s=1}^{r} b_{ls} \mathbf{v}_s)^\top \boldsymbol{\mu}_{jk})^2 \\ =& \sum_{l=1}^{r} \sum_{j \neq k} p_{jk} \sum_{s,t=1}^{r} b_{ls} b_{lt} \mathbf{v}_s^\top \boldsymbol{\mu}_{jk} \mathbf{v}_t^\top \boldsymbol{\mu}_{jk} \\ =& \sum_{s=1}^{r} \sum_{j \neq k} p_{jk} (\mathbf{v}_s^\top \boldsymbol{\mu}_{jk})^2 \sum_{l=1}^{r} b_{ls}^2 + \sum_{j \neq k} p_{jk} \sum_{s,t=1}^{r} \mathbf{v}_s^\top \boldsymbol{\mu}_{jk} \mathbf{v}_t^\top \boldsymbol{\mu}_{jk} \sum_{l=1}^{r} b_{ls} b_{lt}\end{aligned}$$

Since $\mathbf{B}$ is a standard orthogonal matrix, we have $\forall s$, $\sum_{l=1}^{r} b_{ls}^2 = 1$ and $\forall s \neq t$, $\sum_{l=1}^{r} b_{ls} b_{lt} = 0$. Further, we have

$$\sum_{l=1}^{r} \alpha(\mathbf{u}_l) = \sum_{l=1}^{r} \alpha(\mathbf{v}_l).$$

□

Lemma 5. *For any positive integer r, any set of standard orthogonal vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_r \in \mathbb{R}^d$, and real numbers $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_r \geq 0$, we have*

$$\sum_{l=1}^{r} \gamma_l \alpha(\mathbf{u}_l) \leq \sum_{l=1}^{r} \gamma_l \beta_l, \tag{6}$$

where β_l is the l-th feature value.

Proof. We first prove the situation that $\gamma_1 = \gamma_2 = \cdots = \gamma_r = 1$, i.e.,

$$\sum_{l=1}^{r} \alpha(\mathbf{u}_l) \leq \sum_{l=1}^{r} \beta_l. \tag{7}$$

We prove it by induction on r. For $r = 1$, by the definition of feature values and feature vectors, Eq.(7) holds. Now supposing Eq.(7) holds for $r = s$, we prove it holds for $r = s + 1$ by contradiction. If Eq.(7) does not hold, then there exist standard orthogonal vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_{s+1} \in \mathbb{R}^d$, such that

$$\sum_{l=1}^{s+1} \alpha(\mathbf{u}_l) > \sum_{l=1}^{s+1} \alpha(\mathbf{w}_l), \tag{8}$$

where $\mathbf{w}_l$ are feature vectors. Since the dimension of $\operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_{s+1})$ is $s + 1$, there exists $\tilde{\mathbf{w}}_{s+1} \in \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_{s+1})$, such that $\tilde{\mathbf{w}}_{s+1} \perp \mathbf{w}_l, \forall 1 \leq l \leq s$. By the definition of feature vector $\mathbf{w}_{s+1}$, we have

$$\sum_{l=1}^{s+1} \alpha(\mathbf{w}_l) \geq \sum_{l=1}^{s} \alpha(\mathbf{w}_l) + \alpha(\tilde{\mathbf{w}}_{s+1}). \tag{9}$$

Let $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \cdots \tilde{\mathbf{w}}_{s+1}$ be a set of standard orthogonal basis of $\operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_{s+1})$, by Lemma 4, we have

$$\sum_{l=1}^{s+1} \alpha(\mathbf{u}_l) = \sum_{l=1}^{s+1} \alpha(\tilde{\mathbf{w}}_l). \tag{10}$$

Combine equation (8), (9) and (10) we get

$$\sum_{l=1}^{s+1} \alpha(\tilde{\mathbf{w}}_l) > \sum_{l=1}^{s} \alpha(\mathbf{w}_l) + \alpha(\tilde{\mathbf{w}}_{s+1}).$$

Thus we have

$$\sum_{l=1}^{s} \alpha(\tilde{\mathbf{w}}_l) > \sum_{l=1}^{s} \alpha(\mathbf{w}_l).$$

This contradicts with our induction assumption. The proof for the $\gamma_1 = \gamma_2 = \cdots = \gamma_r = 1$ case completes.

Next, we prove the situation that γ_l are not all equal to 1, by utilizing Eq.(7).

$$\begin{aligned}
\sum_{l=1}^{r} \gamma_l \alpha(\mathbf{u}_l) &= \sum_{l=1}^{r-1} [(\gamma_l - \gamma_{l+1}) \sum_{t=1}^{l} \alpha(\mathbf{u}_t)] + \gamma_r \sum_{t=1}^{r} \alpha(\mathbf{u}_t) \\
&\leq \sum_{l=1}^{r-1} [(\gamma_l - \gamma_{l+1}) \sum_{t=1}^{l} \beta_t] + \gamma_r \sum_{t=1}^{r} \beta_t \\
&\leq \sum_{l=1}^{r} \gamma_l \beta_l
\end{aligned}$$

The proof completes. □

Note that in Lemma 5, r can be larger than the number of nonzero feature values $K-1$. This will be used in the proof of Lemma 1 later.

Another auxiliary lemma needed to prove Lemma 1 is given below.

Lemma 6. *Suppose $\mathbf{w}_0 \in \mathcal{G}$, define linear space $\mathcal{H} = \{\mathbf{v} \in \mathcal{G} : \mathbf{v} \perp \mathbf{w}_0\}$. Then there are $K-2$ nonzero feature values of $\mathcal{H}$. Denote them as $\beta'_1, \beta'_2, \cdots \beta'_{K-2}$, then $\forall\ r \leq K-2$, $\forall\ \gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_r \geq 0$,*

$$\sum_{l=1}^{r} \gamma_l \beta'_l \leq \sum_{l=1}^{r} \gamma_l \beta_l$$

Proof. Note that the dimension of $\mathcal{H}$ is $K-2$, then there are $K-2$ nonzero feature values. The feature vectors of $\mathcal{H}$ are also standard orthogonal vectors of the linear space $\mathcal{G}$. By Lemma 5, we have $\sum_{l=1}^{r} \gamma_l \beta'_l \leq \sum_{l=1}^{r} \gamma_l \beta_l$, $\forall\ r \leq K-2$. □

Now we are ready to prove Lemma 1.

Proof. (of Lemma 1) We conduct the proof by contradiction. Assuming Eq.(5) does not hold, we prove $\mathbf{A}^*$ can not be the global optimal solution of PDML. Let $\mathbf{U}\sqrt{\mathbf{\Lambda}}\mathbf{V}^\top$ be the SVD of $\mathbf{A}^*$. Define $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots \mathbf{w}_R)$ as a matrix whose columns contain the feature vectors. Let $\tilde{\mathbf{A}} = \mathbf{U}\sqrt{\mathbf{\Lambda}}\mathbf{W}^\top$. Then by Condition 1, we have $\Omega_\phi(\mathbf{A}^*) = \Omega_\phi(\tilde{\mathbf{A}})$. Define

$$L(\mathbf{A}) = \mathrm{E}\Big[\frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{S}} \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2^2 + \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}} \max(0, \tau - \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2^2)\Big].$$

Assuming Eq.(5) does not hold, we prove $L(\mathbf{A}^*) > L(\tilde{\mathbf{A}})$, i.e., $\mathbf{A}^*$ is not the optimal solution. We consider two cases: $\xi = 0$ and $\xi \neq 0$. Define $h(\mathbf{A}^*, \xi) = L(\mathbf{A}^*)$ and $h(\tilde{\mathbf{A}}, \xi) = L(\tilde{\mathbf{A}})$. When $\xi = 0$, we have:

$$\begin{aligned}
h(\mathbf{A}^*, 0) &= \mathrm{E}\Big[\frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{S}} \|\mathbf{A}^*\boldsymbol{\mu}_{c(\mathbf{x})} - \mathbf{A}^*\boldsymbol{\mu}_{c(\mathbf{y})}\|_2^2 + \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}} \max(0, \tau - \|\mathbf{A}^*\boldsymbol{\mu}_{c(\mathbf{x})} - \mathbf{A}^*\boldsymbol{\mu}_{c(\mathbf{y})}\|_2^2)\Big] \\
&= \sum_{j \neq k} p_{jk} \max(0, \tau - \|\mathbf{A}^*(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)\|_2^2),
\end{aligned}$$

and

$$h(\tilde{\mathbf{A}},0)=\sum_{j\neq k}p_{jk}\max(0,\tau-\|\tilde{\mathbf{A}}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2).$$

Since Eq.(5) does not hold by assumption, there exists $\mathbf{w}_0\in\mathcal{G}$, $\mathbf{w}_0\notin\mathrm{span}(\mathbf{A}^*)$. Denote $\mathcal{H}=\{\mathbf{v}\in\mathcal{G}:\mathbf{v}\perp\mathbf{w}_0\}$ and its $K-2$ nonzero feature values as $\beta_1',\beta_2',\cdots\beta_{K-2}'$. $\forall\mathbf{u}\in\mathrm{span}(\mathbf{A}^*)$, let $\mathbf{u}'$ be the projection of $\mathbf{u}$ to the space $\mathcal{H}$ and $\mathbf{u}'$ is rescaled to have norm 1. Then $\alpha(\mathbf{u}')\geq\alpha(\mathbf{u})$. Thus, $\forall r$, the r-th feature value of $\mathrm{span}(\mathbf{A}^*)$ is no larger than the r-th feature value of $\mathcal{G}$. By Lemma 6, we have $\sum_{l=1}^{r}\gamma_l\beta_l'\leq\sum_{l=1}^{r}\gamma_l\beta_l$. By the definition of feature values, we have

$$\sum_{j\neq k}p_{jk}\|\mathbf{A}^*(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2=\sum_{l=1}^{R}\gamma_l\alpha(\mathbf{a}_l)\leq\sum_{l=1}^{R}\gamma_l\beta_l'.$$

Since $\mathcal{H}$ has only $K-2$ nonzero feature values, we have

$$\sum_{l=1}^{R}\gamma_l\beta_l'=\sum_{l=1}^{K-2}\gamma_l\beta_l'\leq\sum_{l=1}^{K-2}\gamma_l\beta_l=\sum_{l=1}^{K-1}\gamma_l\alpha(\mathbf{w}_l)-\gamma_{K-1}\beta_{K-1}=\sum_{j\neq k}p_{jk}\|\tilde{\mathbf{A}}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2-\gamma_{K-1}\beta_{K-1}.$$

So we have

$$\sum_{j\neq k}p_{jk}\|\tilde{\mathbf{A}}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2\geq\sum_{j\neq k}p_{jk}\|\mathbf{A}^*(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2+\gamma_{K-1}\beta_{K-1}.$$

Next, we establish a relationship between $h(\mathbf{A}^*,0)$ and $h(\tilde{\mathbf{A}},0)$, which is given in the following lemma.

Lemma 7. *There exist constants τ_0,γ_0 which are determined by $p_1,p_2,\cdots p_K$ and $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_K$, such that if $\tau\geq\tau_0,\gamma\geq\gamma_0$, then we have*

$$h(\mathbf{A}^*,0)-h(\tilde{\mathbf{A}},0)>\frac{1}{2}\gamma_{K-1}\beta_{K-1}.$$

Proof. If $\|\tilde{\mathbf{A}}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2\leq\tau$ and $\|\mathbf{A}^*(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2\leq\tau$ for all $j\neq k$, we have $h(\mathbf{A}^*,0)-h(\tilde{\mathbf{A}},0)=\gamma_{K-1}\beta_{K-1}$. Since $\max_{j\neq k}\|\boldsymbol{\mu}_j-\boldsymbol{\mu}_k\|_2=B_0$, we have

$$\begin{aligned}&\|\mathbf{A}^*(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2\leq\mathrm{tr}(\boldsymbol{\Lambda})B_0^2,\\&\|\tilde{\mathbf{A}}(\boldsymbol{\mu}_j-\boldsymbol{\mu}_k)\|_2^2\leq\mathrm{tr}(\boldsymbol{\Lambda})B_0^2,\ \ \forall j\neq k.\end{aligned}\tag{11}$$

Select τ_0 such that $\tau_0\geq K(1+\epsilon_0)B_0^2$, where ϵ_0 is any positive constant. For the VND and LDD regularizers, as $\gamma\to\infty$, $\boldsymbol{\Lambda}\to\mathbf{I}_R$. Thereby, there exists γ_0, such that if $\gamma\geq\gamma_0$, $\forall j$, $|\lambda_j-1|\leq\epsilon$. Hence, if $\gamma\geq\gamma_0,\tau\geq\tau_0$,

$$\mathrm{tr}(\boldsymbol{\Lambda})B_0^2\leq K(1+\epsilon_0)B_0^2\leq\tau_0.$$

Combining this inequality with Eq.(11), we finish the proof. □

Now we continue to prove Lemma 1. In Lemma 7, we have already proved that $h(\mathbf{A}^*,0)$ is strictly larger than $h(\tilde{\mathbf{A}},0)$. We then prove that if the noise is smaller than a certain value, $h(\mathbf{A}^*,\xi)$ is strictly larger than $h(\tilde{\mathbf{A}},\xi)$. By the definition of ξ, we have

$$\begin{aligned}&|h(\mathbf{A}^*,\xi)-h(\mathbf{A}^*,0)|\\\leq&\mathrm{E}\frac{1}{|\mathcal{S}|}\sum_{(x,y)\in\mathcal{S}}\|\mathbf{A}^*(\mathbf{x}-\mathbf{y})\|_2^2+\mathrm{E}\frac{1}{|\mathcal{D}|}\sum_{(x,y)\in\mathcal{D}}[\|\mathbf{A}^*(\mathbf{x}-\mathbf{y})\|_2^2-\|\mathbf{A}^*(\boldsymbol{\mu}_{c(\mathbf{x})}-\boldsymbol{\mu}_{c(\mathbf{y})})\|_2^2]\\\leq&4\mathrm{tr}(\boldsymbol{\Lambda})\xi^2+(4B_0\xi+4\xi^2)\mathrm{tr}(\boldsymbol{\Lambda})\\=&8\xi^2\mathrm{tr}(\boldsymbol{\Lambda})+4B_0\xi\mathrm{tr}(\boldsymbol{\Lambda}).\end{aligned}\tag{12}$$

Similarly, we have

$$|h(\mathbf{A}^*,\xi)-h(\mathbf{A}^*,0)|\leq8\xi^2\mathrm{tr}(\boldsymbol{\Lambda})+4B_0\xi\mathrm{tr}(\boldsymbol{\Lambda}).\tag{13}$$

Combining Lemma 7 with Eq.(12) and Eq.(13), we have if $\xi\leq\frac{-B_0+\sqrt{B_0^2+\gamma_{K-1}\beta_{K-1}/(2\mathrm{tr}(\boldsymbol{\Lambda}))}}{4}$, then $L(\mathbf{A}^*)>L(\tilde{\mathbf{A}})$, i.e., $\mathbf{A}^*$ is not the global optimal solution. By contradiction, Eq.(5) holds. The proof completes. □

4.3 Proof of Lemma 2

Proof. For any vector $\mathbf{u} \in \mathcal{G}$, since the condition of Lemma 1 is satisfied, we have $\mathbf{u} \in \text{span}(\mathbf{A}^*)$. Recall $\mathbf{A}^{*\top}\mathbf{A}^* = \mathbf{V}\mathbf{\Gamma}\mathbf{V}^\top$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_R]$. We can denote $\mathbf{u}$ as $\mathbf{u} = \|\mathbf{u}\| \sum_{j=1}^R t_j \mathbf{v}_j$, where $\sum_{j=1}^R t_j^2 = 1$. Then we have $\forall \mathbf{u} \in \mathcal{G}$,

$$\mathbf{u}^\top \mathbf{A}^{*\top}\mathbf{A}^*\mathbf{u} = \sum_{j=1}^R \langle \mathbf{v}_j, \mathbf{u}\rangle^2 \lambda_j = \sum_{j=1}^R \|\mathbf{u}\|^2 t_j^2 \lambda_j \leq \|\mathbf{u}\|^2 \lambda_1.$$

Similarly, we have $\mathbf{u}^\top \mathbf{A}^{*\top}\mathbf{A}^*\mathbf{u} \geq \|\mathbf{u}\|^2 \lambda_R$. Noting $\forall j \neq k$, $\boldsymbol{\mu}_j - \boldsymbol{\mu}_k \in \mathcal{G}$, we have

$$\eta \leq \text{cond}(\mathbf{A}^*\mathbf{A}^{*\top}) \frac{\max_{j\neq k} \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|^2}{\min_{j\neq k} \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|^2}.$$

Combining this inequality with $\text{cond}(\mathbf{A}^*\mathbf{A}^{*\top}) \leq g(\Omega_\phi(\mathbf{A}^*))$, we complete the proof. □

4.4 Proof of Lemma 3

Proof. We first prove the result about the VND regularizer. Define scalar function $s(x) = x \log x - x + 1$ and denote $\text{cond}(\mathbf{A}^*\mathbf{A}^{*\top}) = c$. Since $s'(x) = \log x$, and $s(1) = 0$, we have

$$\begin{aligned} \Omega_{vnd}(\mathbf{A}^*) &= \sum_{j=1}^R s(\lambda_j) \\ &\geq s(\lambda_1) + s(\lambda_R) \\ &= \lambda_1 \log \lambda_1 - \lambda_1 + \frac{\lambda_1}{c} \log \frac{\lambda_1}{c} - \frac{\lambda_1}{c} + 2 \end{aligned}$$

Define $F(\lambda_1, c) = \lambda_1 \log \lambda_1 - \lambda_1 + \frac{\lambda_1}{c} \log \frac{\lambda_1}{c} - \frac{\lambda_1}{c} + 2$. We aim to maximize c, so

$$\frac{\partial}{\partial \lambda_1} F(\lambda_1, c) = 0.$$

This equation has a unique solution: $\log \lambda_1 = \frac{\log c}{c+1}$. Therefore we have

$$c^{1/(c+1)}(1 + \frac{1}{c}) \geq 2 - \Omega_{vnd}(\mathbf{A}^*).$$

Define $f(c) = c^{1/(c+1)}(1 + \frac{1}{c})$. Its derivative is: $f'(c) = -\frac{\log c}{c(c+1)} c^{1/(c+1)}$. Analyzing $f'(c)$, we know that $f(c)$ increases on $(0, 1]$, decreases on $[1, \infty)$, and $f(1) = 2$. Also we have the following limits:

$$\lim_{c\to 0} f(c) = 0, \ \lim_{c\to\infty} f(c) = 1.$$

We denote the inverse function of $f(\cdot)$ on $[1, \infty)$ as $f^{-1}(\cdot)$. Then for any $\Omega_{vnd}(\mathbf{A}^*) < 1$, we have

$$\text{cond}(\mathbf{A}^*\mathbf{A}^{*\top}) \leq f^{-1}(2 - \Omega_{vnd}(\mathbf{A}^*)).$$

Next we prove the result for the LDD regularizer $\Omega_{ldd}(\mathbf{A}^*)$. Define scalar function $s(x) = x - \log x - 1$ and denote $\text{cond}(\mathbf{A}^*\mathbf{A}^{*\top}) = c$. Since $s'(x) = 1 - \frac{1}{x}$ and $s(1) = 0$, we have

$$\begin{aligned} \Omega_{ldd}(\mathbf{A}^*) &= \sum_{j=1}^R s(\lambda_j) \\ &\geq s(\lambda_1) + s(\lambda_R) \\ &= \lambda_1 - \log \lambda_1 + \frac{\lambda_1}{c} - \log \frac{\lambda_1}{c} - 2 \end{aligned}$$

Therefore we have

$$\begin{aligned}\log c &\le \Omega_{ldd}(\mathbf{A}^*) + 2\log\lambda_1 - \lambda_1(1+\frac{1}{c}) + 2\\ &\le \Omega_{ldd}(\mathbf{A}^*) + 2\log\lambda_1 - \lambda_1 + 2\\ &\le \Omega_{ldd}(\mathbf{A}^*) + 2\log 2 - 2 + 2\\ &= \Omega_{ldd}(\mathbf{A}^*) + 2\log 2\end{aligned}$$

The third inequality is obtained from the following fact: the scalar function $\log x - x$ gets its maximum when $x = 2$. Further, we have

$$c \le 4e^{\Omega_{ldd}(\mathbf{A}^*)}.$$

The proof completes.

□

5 Proof of Theorem 2

5.1 Proof Sketch

Part of the proof is tailored to the CVND regularizer. Extensions to CSFN and CLDD are given later. The proof is based on Rademacher complexity (RC) [41], which measures the complexity of a hypothesis class. In MDML, the Rademacher complexity $\mathcal{R}(\mathcal{M})$ of the function class $\mathcal{M}$ is defined as:

$$\mathcal{R}(\mathcal{M}) = \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\mathbf{M}\in\mathcal{M}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i(\mathbf{x}_i - \mathbf{y}_i)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{y}_i)$$

where m is the number of data pairs in the training data ($m = |\mathcal{S}| + |\mathcal{D}|$), $\sigma_i \in \{-1, 1\}$ is the Rademacher variable and $\sigma = (\sigma_1, \sigma_2, \cdots \sigma_m)$.

We first establish a upper bound of the generalization error based on RC. Intuitively, a less-complicated hypothesis class generalizes better on unseen data. Then we upper bound the RC based on the CBMD regularizers. Combining the two steps together, we establish upper bounds of the generalization error based on CBMD regularizers.

The following lemma presents the RC-based upper bound of the generalization error. Its proof is adapted from [41].

Lemma 8. *With probability at least* $1-\delta$*, we have*

$$\sup_{\mathbf{M}\in\mathcal{M}} (L(\mathbf{M}) - \hat{L}(\mathbf{M})) \le 2\mathcal{R}(\mathcal{M}) + \max(\tau, \sup_{\substack{(\mathbf{x},\mathbf{y})\in\mathcal{S}\\ \mathbf{M}\in\mathcal{M}}} (\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y}))\sqrt{\frac{2\log(1/\delta)}{m}}. \tag{14}$$

For the second term in the bound, it is easy to verify

$$\sup_{\substack{(\mathbf{x},\mathbf{y})\in\mathcal{S}\\ \mathbf{M}\in\mathcal{M}}} (x-y)^\top\mathbf{M}(\mathbf{x}-\mathbf{y}) \le \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}) \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \|\mathbf{x}-\mathbf{y}\|_2^2. \tag{15}$$

Now we focus on the first term. We denote $\mathbf{z} = \mathbf{x} - \mathbf{y}$, $\mathbf{z}_i = \mathbf{x}_i - \mathbf{y}_i$.

Lemma 9. *Suppose* $\sup_{\|\mathbf{v}\|_2\le 1,\mathbf{z}} |\mathbf{v}^\top\mathbf{z}| \le B$*, then we have*

$$\mathcal{R}(\mathcal{M}) \le \frac{2B^2}{\sqrt{m}} \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}). \tag{16}$$

We next show that $\mathrm{tr}(\mathbf{M})$ can be bounded by the CVND regularizer $\hat{\Omega}_{vnd}(\mathbf{M})$.

Lemma 10. *For the convex VND regularizer* $\hat{\Omega}_{vnd}(\mathbf{M})$*, for any positive semidefinite matrix* $\mathbf{M}$*, we have*

$$\mathrm{tr}(\mathbf{M}) \le \hat{\Omega}_{vnd}(\mathbf{M}).$$

Combining Lemma 8, 9, 10 and Eq.(15) and noting that $\mathcal{E} = L(\hat{\mathbf{M}}^*) - \hat{L}(\hat{\mathbf{M}}^*) \leq \sup_{\mathbf{M}\in\mathcal{M}}(L(\mathbf{M}) - \hat{L}(\mathbf{M}))$ and $\hat{\Omega}_{vnd}(\mathbf{M}) \leq C$ (C is the upper bound in the hypothesis class $\mathcal{M}$), we complete the proof of the first bound in Theorem 2 (Eq.(6) in the main paper).

In the sequel, we present detailed proofs of these lemmas and the extension to CSNF and CLDD.

5.2 Proof of Lemma 9

Proof. For any $\mathbf{M} \in \mathcal{M}$, denote its spectral decomposition as $\mathbf{M} = \mathbf{V}\mathbf{\Pi}\mathbf{V}^\top$, where $\mathbf{V}$ is standard orthogonal matrix and $\mathbf{\Pi}$ is diagonal matrix. Denote $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_D)$, $\mathbf{\Pi} = \mathrm{diag}(\pi_1, \pi_2, \cdots \pi_D)$, then we have

$$\begin{aligned}
\mathcal{R}(\mathcal{M}) &= \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\mathbf{M}\in\mathcal{M}} \left[\frac{1}{m} \sum_{i=1}^{m} \sigma_i \mathbf{z}_i^T \mathbf{M} \mathbf{z}_i \right] \\
&= \frac{1}{m} \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\mathbf{M}\in\mathcal{M}} \left[\sum_{i=1}^{m} \sigma_i \sum_{j=1}^{D} \pi_j (\mathbf{v}_j^\top \mathbf{z}_i)^2 \right] \\
&= \frac{1}{m} \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\mathbf{M}\in\mathcal{M}} \left[\sum_{j=1}^{D} \pi_j \sum_{i=1}^{m} \sigma_i (\mathbf{v}_j^\top \mathbf{z}_i)^2 \right] \\
&= \frac{1}{m} \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\mathbf{M}\in\mathcal{M}} \left[\sum_{j=1}^{D} \pi_j \sup_{\|\mathbf{v}\|_2 \leq 1} \sum_{i=1}^{m} \sigma_i (\mathbf{v}^\top \mathbf{z}_i)^2 \right] \\
&= \frac{1}{m} \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\mathbf{\Pi}} \sum_{j=1}^{D} \pi_j \sup_{\|\mathbf{v}\|_2 \leq 1} \sum_{i=1}^{m} \sigma_i (\mathbf{v}^\top \mathbf{z}_i)^2 \\
&\leq \frac{1}{m} \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}) \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\|\mathbf{v}\|_2 \leq 1} \sum_{i=1}^{m} \sigma_i (\mathbf{v}^\top \mathbf{z}_i)^2.
\end{aligned}$$

Since $(\mathbf{v}^\top \mathbf{z})^2$ is Lipschitz continuous w.r.t $\mathbf{v}^\top \mathbf{z}$ with constant $2 \sup_{\|\mathbf{v}\|_2 \leq 1, \mathbf{z}} \mathbf{v}^\top \mathbf{z}$, according to the composition property [41] of Rademacher complexity on Lipschitz continuous functions, we have

$$\begin{aligned}
\mathcal{R}(\mathcal{M}) &\leq \frac{1}{m} 2 \sup_{\|\mathbf{v}\|_2 \leq 1, \mathbf{z}} (\mathbf{v}^\top \mathbf{z}) \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}) \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\|\mathbf{v}\|_2 \leq 1} \sum_{i=1}^{m} \sigma_i \mathbf{v}^\top \mathbf{z}_i \\
&= 2 \frac{B}{m} \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}) \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\|\mathbf{v}\|_2 \leq 1} \sum_{i=1}^{m} \sigma_i \mathbf{v}^\top \mathbf{z}_i \\
&\leq 2 \frac{B}{m} \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}) \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sup_{\|\mathbf{v}\|_2 \leq 1} \|\mathbf{v}\|_2 \| \sum_{i=1}^{m} \sigma_i \mathbf{z}_i \|_2 \\
&= 2 \frac{B}{m} \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}) \mathrm{E}_{\mathcal{S},\mathcal{D},\sigma} \sqrt{ (\sum_{i=1}^{m} \sigma_i \mathbf{z}_i)^2 }.
\end{aligned}$$

By Jensen's inequality, we have

$$\begin{aligned}
\mathcal{R}(\mathcal{M}) &\leq 2 \frac{B}{m} \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}) \mathrm{E}_{\mathcal{S},\mathcal{D}} \sqrt{ \mathrm{E}_{\sigma} (\sum_{i=1}^{m} \sigma_i \mathbf{z}_i)^2 } \\
&=\leq 2 \frac{B}{m} \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}) \mathrm{E}_{\mathcal{S},\mathcal{D}} \sqrt{ \sum_{i=1}^{m} \mathbf{z}_i^2 } \\
&\leq \frac{2B^2}{\sqrt{m}} \sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M}).
\end{aligned}$$

□

5.3 Proof of lemma 10

Proof. For any positive semidefinite matrix $\mathbf{M}$, we use notations $\mathbf{V}, \mathbf{\Pi}, \pi_j, 1 \leq j \leq D$ as they are defined in Section 5.2. By the definition of the convex VND regularizer, we have

$$\begin{aligned}\hat{\Omega}_{vnd}(\mathbf{M}) =&\Gamma_{vnd}(\mathbf{M} + \epsilon\mathbf{I}_D, \mathbf{I}_D) + \mathrm{tr}(\mathbf{M})\\ =&\mathrm{tr}[(\mathbf{M} + \epsilon\mathbf{I}_D)\log(\mathbf{M} + \epsilon\mathbf{I}_D) - (\mathbf{M} + \epsilon\mathbf{I}_D)\log\mathbf{I}_D - (\mathbf{M} + \epsilon) + \mathbf{I}_D] + \mathrm{tr}(\mathbf{M})\\ =&\sum_{j=1}^{D}[(\pi_j + \epsilon)\log(\pi_j + \epsilon) - (\pi_j + \epsilon) + 1] + \sum_{j=1}^{D}\pi_j\\ =&\sum_{j=1}^{D}[(\lambda_j + \epsilon)\log(\lambda_j + \epsilon) - \epsilon + 1]\end{aligned}$$

Denote $\bar{\pi} = (\sum_{j=1}^{D}\pi_j)/D = \mathrm{tr}(\mathbf{M})/D$, then by Jensen's inequality, we have

$$\sum_{j=1}^{D}(\lambda_j + \epsilon)\log(\lambda_j + \epsilon) \geq D(\bar{\pi} + \epsilon)\log(\bar{\pi} + \epsilon).$$

Since $\forall x \in \mathbb{R}_+, x - 1 \leq x\log x$, so we have

$$\begin{aligned}\bar{\pi} + \epsilon - 1 &\leq (\bar{\pi} + \epsilon)\log(\bar{\pi} + \epsilon)\\ &\leq \frac{1}{D}\sum_{j=1}^{D}(\lambda_j + \epsilon)\log(\lambda_j + \epsilon)\\ &\leq \frac{1}{D}\hat{\Omega}_{vnd}(\mathbf{M}) + \epsilon - 1.\end{aligned}$$

Therefore we have

$$\mathrm{tr}(\mathbf{M}) \leq \hat{\Omega}_{vnd}(\mathbf{M}).$$

□

5.4 Generalization error bound for the convex SFN regularizer

In this section we prove generalization error bounds for the convex SFN regularizer. The CSFN is composed of two parts. One is the squared Frobenius norm of $\mathbf{M} - \mathbf{I}_D$ and the other is the trace of $\mathbf{M}$. We have already established a relationship between $\mathrm{tr}(\mathbf{M})$ and $\mathcal{R}(\mathcal{M})$. Now we analyze the relationship between $\|\mathbf{M} - \mathbf{I}_D\|_F$ and $\mathcal{R}(\mathcal{M})$, which is given in the following lemma.

Lemma 11. *Suppose* $\sup_{\|\mathbf{v}\|_2 \leq 1, \mathbf{z}} |\mathbf{v}^\top\mathbf{z}| \leq B$*, then we have*

$$\mathcal{R}(\mathcal{M}) \leq \frac{B^2}{\sqrt{m}}\sup_{\mathbf{M}\in\mathcal{M}}\|\mathbf{M} - \mathbf{I}_D\|_F \tag{17}$$

Proof. Denote $M(j,k) = a_{jk}$, and $\delta_{jk} = \mathrm{I}_{\{j=k\}}$, $\mathbf{z}_i = (z_{i1}, z_{i2}, \cdots z_{id})$, then we have

$$\begin{aligned}\mathcal{R}(\mathcal{M}) =&\frac{1}{m}\mathrm{E}_{\mathcal{S},\mathcal{D},\sigma}\sup_{\mathbf{M}\in\mathcal{M}}\left[\sum_{j,k}a_{jk}\sum_{i=1}^{m}\sigma_i z_{ij}z_{ik}\right]\\ =&\frac{1}{m}\mathrm{E}_{\mathcal{S},\mathcal{D},\sigma}\sup_{\mathbf{M}\in\mathcal{M}}\left[\sum_{j,k}(a_{jk} - \delta_{jk})\sum_{i=1}^{m}\sigma_i z_{ij}z_{ik} + \sum_{j,k}\delta_{jk}\sum_{i=1}^{m}\sigma_i z_{ij}z_{ik}\right]\\ \leq&\frac{1}{m}\mathrm{E}_{\mathcal{S},\mathcal{D},\sigma}\sup_{\mathbf{M}\in\mathcal{M}}\left[\|\mathbf{M} - \mathbf{I}_D\|_F\sqrt{\sum_{j,k}(\sum_{i=1}^{m}\sigma_i z_{ij}z_{ik})^2}\right]\end{aligned}$$

Here the inequality is attained by Cauchy's inequality. Applying Jensen's inequality, we have

$$\begin{aligned}\mathcal{R}(\mathcal{M}) &\leq \frac{1}{m} \sup_{\mathbf{M}\in\mathcal{M}} \|\mathbf{M}-\mathbf{I}_D\|_F \ \mathrm{E}_{\mathcal{S},\mathcal{D}} \left[\sqrt{\mathrm{E}_\sigma \sum_{j,k} (\sum_{i=1}^{m} \sigma_i z_{ij} z_{ik})^2}\right] \\ &= \frac{1}{\sqrt{m}} \sup_{\mathbf{M}\in\mathcal{M}} \|\mathbf{M}-\mathbf{I}_D\|_F \ \mathrm{E}_{\mathcal{S},\mathcal{D}} \left[\sqrt{\sum_{j,k} z_{ij}^2 z_{ik}^2}\right]\end{aligned}$$

Recalling the definition of B, we have

$$\mathcal{R}(\mathcal{M}) \leq \frac{B^2}{\sqrt{m}} \sup_{\mathbf{M}\in\mathcal{M}} \|\mathbf{M}-\mathbf{I}_D\|_F.$$

□

We now bound the generalization error with the convex SFN regularizer, which is given in the following lemma.

Lemma 12. *Suppose* $\sup_{\|\mathbf{v}\|_2\leq 1,\mathbf{z}} |\mathbf{v}^\top \mathbf{z}| \leq B$*, then with probability at least* $1-\delta$*, we have*

$$\sup_{\mathbf{M}\in\mathcal{M}} (L(\mathbf{M}) - \hat{L}(\mathbf{M})) \leq \frac{2B^2}{\sqrt{m}} \min(2\hat{\Omega}_{sfn}(\mathbf{M}), \sqrt{\hat{\Omega}_{sfn}(\mathbf{M})}) + \max(\tau, \hat{\Omega}_{sfn}(\mathbf{M}))\sqrt{\frac{2\log(1/\delta)}{m}}.$$

Proof. For the convex SFN regularizer $\hat{\Omega}_{sfn}(\mathbf{M})$, we have $\mathrm{tr}(\mathbf{M}) \leq \hat{\Omega}_{sfn}(\mathbf{M})$ and $\|\mathbf{M}-\mathbf{I}_D\| \leq \hat{\Omega}_{sfn}(\mathbf{M})$. By Eq.(15), we have

$$\sup_{\substack{(\mathbf{x},\mathbf{y}) \in \mathcal{S} \\ \mathbf{M} \in \mathcal{M}}} (x-y)^\top \mathbf{M}(\mathbf{x}-\mathbf{y}) \leq \sup_{\mathbf{M}\in\mathcal{M}} \hat{\Omega}_{sfn}(\mathbf{M}) B^2. \tag{18}$$

By Lemma 9 and 11, we have

$$\mathcal{R}(\mathcal{M}) \leq \frac{B^2}{\sqrt{m}} \min(2\hat{\Omega}_{sfn}(\mathbf{M}), \sqrt{\hat{\Omega}_{sfn}(\mathbf{M})}). \tag{19}$$

Substituting Eq.(19) and Eq.(18) into Lemma 8, we have

$$\sup_{\mathbf{M}\in\mathcal{M}} (L(\mathbf{M}) - \hat{L}(\mathbf{M})) \leq \frac{2B^2}{\sqrt{m}} \min(2\hat{\Omega}_{sfn}(\mathbf{M}), \sqrt{\hat{\Omega}_{sfn}(\mathbf{M})}) + \max(\tau, \hat{\Omega}_{sfn}(\mathbf{M}))\sqrt{\frac{2\log(1/\delta)}{m}}.$$

□

Noting that $\mathcal{E} = L(\hat{\mathbf{M}}^*) - \hat{L}(\hat{\mathbf{M}}^*) \leq \sup_{\mathbf{M}\in\mathcal{M}}(L(\mathbf{M}) - \hat{L}(\mathbf{M}))$ and $\hat{\Omega}_{sfn}(\mathbf{M}) \leq C$, we conclude $\mathcal{E} \leq \frac{2B^2}{\sqrt{m}} \min(2C, \sqrt{C}) + \max(\tau, C)\sqrt{\frac{2\log(1/\delta)}{m}}$.

5.5 Generalization error bound for the convex LDD regularizer

Starting from Lemma 8, we bound $\mathcal{R}(\mathcal{M})$ and $\sup_{\mathbf{M}\in\mathcal{M}} \mathrm{tr}(\mathbf{M})$ which are given in the following two lemmas.

Lemma 13. *Suppose* $\sup_{\|\mathbf{v}\|_2\leq 1,\mathbf{z}} |\mathbf{v}^\top \mathbf{z}| \leq B$*, then we have*

$$\mathcal{R}(\mathcal{M}) \leq \frac{B}{\sqrt{m}} \frac{\hat{\Omega}_{ldd}(\mathbf{M})}{\log(1/\epsilon)-1}.$$

Proof. We first perform some calculation on the convex LDD regularizer.

$$\begin{aligned}\hat{\Omega}_{ldd}(\mathbf{M}) &= \Gamma_{ldd}(\mathbf{M}+\epsilon\mathbf{I}_D, \mathbf{I}_D) - (1+\log\epsilon)\mathrm{tr}(\mathbf{M}) \\ &= \mathrm{tr}((\mathbf{M}+\epsilon\mathbf{I}_D)\mathbf{I}_D^{-1}) - \log\det((\mathbf{M}+\epsilon\mathbf{I}_D)\mathbf{I}_D^{-1}) - D - (1+\log\epsilon)\mathrm{tr}(\mathbf{M}) \\ &= \textstyle\sum_{j=1}^{D}(\pi_j+\epsilon) - \sum_{j=1}^{D}\log(\pi_j+\epsilon) - D - (1+\log\epsilon)\sum_{j=1}^{D}\pi_j \\ &= \textstyle\log(\frac{1}{\epsilon})\sum_{j=1}^{D}\pi_j - \sum_{j=1}^{D}\log(\pi_j+\epsilon) - D(1-\epsilon).\end{aligned} \tag{20}$$

Now we upper bound the Rademacher complexity using the CLDD regularizer.

$$\begin{aligned}\log(\frac{1}{\epsilon})\mathcal{R}(\mathcal{M}) =&\frac{\log(\frac{1}{\epsilon})}{m}\mathrm{E}_{\mathcal{S},\mathcal{D},\sigma}\sup_{\mathbf{M}\in\mathcal{M}}\left[\sum_{j=1}^{D}\pi_j\sum_{i=1}^{m}\sigma_i(\mathbf{v}_j^\top\mathbf{z}_i)^2\right]\\ \le&\frac{1}{m}\mathrm{E}_{\mathcal{S},\mathcal{D},\sigma}\sup_{\mathbf{\Pi}}\sum_{j=1}^{D}[(\log(\frac{1}{\epsilon})\pi_j - \log(\pi_j+\epsilon)) + \log(\pi_j+\epsilon)]\sup_{\|\mathbf{v}\|_2\le 1}\sum_{i=1}^{m}\sigma_i(\mathbf{v}^\top\mathbf{z}_i)^2\end{aligned}$$

Similar to the proof of Lemma 9, we have

$$\begin{aligned}\log(\tfrac{1}{\epsilon})\mathcal{R}(\mathcal{M}) &\le \tfrac{2B^2}{\sqrt{m}}\sup_{\mathbf{\Pi}}\textstyle\sum_{j=1}^{D}[(\log(\frac{1}{\epsilon})\pi_j - \log(\pi_j+\epsilon)) + \log(\pi_j+\epsilon)]\\ &\le \tfrac{2B^2}{\sqrt{m}}[\sup_{\mathbf{M}\in\mathcal{M}}\hat{\Omega}_{ldd}(\mathbf{M}) + \sup_{\mathbf{M}\in\mathcal{M}}\textstyle\sum_{j=1}^{D}\log(\pi_j+\epsilon)].\end{aligned} \tag{21}$$

Denoting $A = \sum_{j=1}^{D}\log(\pi_j+\epsilon)$, we bound A with $\hat{\Omega}_{ldd}(\mathbf{M})$. Denoting $\bar{\pi} = (\sum_{j=1}^{D}\pi_j)/D = \mathrm{tr}(\mathbf{M})/D$, by Jensen's inequality, we have

$$A \le D\log(\bar{\pi}+\epsilon), \tag{22}$$

then $\bar{\pi} \ge e^{A/D} - \epsilon$. Replacing $\bar{\pi}$ with A in Eq.(20), we have

$$\begin{aligned}\hat{\Omega}_{ldd}(\mathbf{M}) \ge& D\log(1/\epsilon)(e^{A/D} - \epsilon) - A - D(1-\epsilon)\\ \ge& D\log(1/\epsilon)(\frac{A}{D} + 1 - \epsilon) - A - D(1-\epsilon)\\ =&(\log(1/\epsilon) - 1)A + [\log(\frac{1}{\epsilon}) - 1]D(1-\epsilon).\end{aligned}$$

Further,

$$A \le \frac{\hat{\Omega}_{ldd}(\mathbf{M})}{\log(\frac{1}{\epsilon}) - 1} - D(1-\epsilon). \tag{23}$$

Substituting this upper bound of A into Eq.(21), we have

$$\mathcal{R}(\mathcal{M}) \le \frac{2B^2}{\sqrt{m}}\frac{\sup_{\mathbf{M}\in\mathcal{M}}\hat{\Omega}_{ldd}(\mathbf{M})}{\log(1/\epsilon) - 1}.$$

□

The next lemma shows the bound of $\mathrm{tr}(\mathbf{M})$.

Lemma 14. *For any positive semidefinite matrix* $\mathbf{M}$, *we have*

$$\mathrm{tr}(\mathbf{M}) \le \frac{\hat{\Omega}_{ldd}(\mathbf{M}) - D\epsilon}{\log(\frac{1}{\epsilon}) - 1}.$$

Proof.

$$\begin{aligned}\hat{\Omega}_{ldd}(\mathbf{M}) \ge& D\log(1/\epsilon)\bar{\pi} - D\log(\bar{\pi}+\epsilon) - D(1-\epsilon)\\ \ge& D\log(1/\epsilon)\bar{\pi} + D(1-\bar{\pi}) - D(1-\epsilon)\\ =&D[\log(1/\epsilon) - 1]\bar{\pi} + D\epsilon.\end{aligned}$$

Then

$$\mathrm{tr}(\mathbf{M}) = D\bar{\pi} \le \frac{\hat{\Omega}_{ldd}(\mathbf{M}) - D\epsilon}{\log(\frac{1}{\epsilon}) - 1}.$$

□

Combining Lemma 13, 14, and 8, we get the following generalization error bound w.r.t the convex LDD regularizer.

Lemma 15. *Suppose* $\sup_{\|\mathbf{v}\|_2\le 1,\mathbf{z}}|\mathbf{v}^\top\mathbf{z}| \le B$, *then with probability at least* $1-\delta$, *we have*

$$\sup_{\mathbf{M}\in\mathcal{M}}(L(\mathbf{M}) - \hat{L}(\mathbf{M})) \le \frac{4B^2}{\sqrt{m}}\frac{\hat{\Omega}_{ldd}(\mathbf{M})}{[\log(1/\epsilon) - 1]} + \max\Big(\tau, \frac{\hat{\Omega}_{ldd}(\mathbf{M}) - D\epsilon}{\log(\frac{1}{\epsilon}) - 1}\Big)\sqrt{\frac{2\log(1/\delta)}{m}}.$$

6 Experiments

6.1 Details of Datasets and Feature Extraction

MIMIC-III MIMIC-III contains 58K hospital admissions of patients who stayed within the intensive care units at Beth Israel Deaconess Medical Center between 2001 and 2012. Each admission has a primary diagnosis (a disease), which acts as the class label of this admission. There are 2833 unique diseases. We extract 7207-dimensional features: (1) 2 dimensions from demographics, including age and gender; (2) 5300 dimensions from clinical notes, including 5000-dimensional bag-of-words (weighted using *tf-idf*) and 300-dimensional Word2Vec [42]; (3) 1905-dimensions from lab tests where the zero-order, first-order and second-order temporal features are extracted for each of the 635 lab items. In the extraction of bag-of-words features from clinical notes, we remove stop words, then count the document frequency (DF) of the remaining words. Then we select the largest 5000 words to form the dictionary. Based on this dictionary, we extract *tfidf* features. In the extraction of word2vec features, we train 300-dimensional embedding vector for each word using an open source word2vec tool[1]. To represent a clinical note, we average the embeddings of all words in this note. In lab tests, there are 635 test items in total. An item is tested at different time points for each admission. For an item, we extract three types of temporal features: (1) *zero-order*: averaging the values of this item measured at different time points; (2) *first-order*: taking the difference of values at every two consecutive time points t and $t-1$, and averaging these differences; (3) *second-order*: for the sequence of first-order differences generated in (2), taking the difference (called second-order difference) of values at every two consecutive time points t and $t-1$, and averaging these second-order differences. If an item is missing in an admission, we set the zero-order, first-order and second-order feature values to 0. The features are normalized using min-max normalization along each dimension. We use PCA to reduce the feature dimension to 1000.

EICU The EICU dataset contains hospital admissions of patients who were treated as part of the Philips eICU program across intensive care units in the United States between 2014 and 2015. Each admission has a primary diagnosis (a disease), which acts as the class label of this admission. There are 2175 unique diseases. There are 474 lab test items and 48 vital sign items. Each admission has a past medical history, which is a collection of diseases. There are 2644 unique past diseases. We extract the following features: (1) age and gender; (2) zero, first and second order temporal features of lab test and vital signs; (3) past medical history: we use a binary vector to encode them; if an element in the vector is 1, then the patient had the corresponding disease in the past. The features are normalized using min-max normalization along each dimension. We use PCA to reduce the feature dimension to 1000.

Reuters and News The original Reuters-21578 dataset contains 21578 documents in 135 classes. We remove documents that have more than one labels, and remove classes that have less than 3 documents, which leaves us 5931 documents and 48 classes. Documents in Reuters and News are represented with *tfidf* vectors where the vocabulary size is 5000. The features are normalized using min-max normalization along each dimension. We use PCA to reduce the feature dimension to 1000.

Birds and Cars For the two image datasets, we use the VGG16 [43] convolutional neural network trained on the ImageNet [44] dataset to extract features, which are the 4096-dimensional outputs of the second fully-connected layer. The features are normalized using min-max normalization along each dimension. We use PCA to reduce the feature dimension to 1000.

6-Activities The 6-Activities dataset contains sensory recordings of 30 subjects performing 6 activities (which are the class labels). The features are 561-dimensional sensory signals.

6.2 Additional 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)

[1] https://code.google.com/archive/p/word2vec/

	MIMIC	EICU	Reuters	News	Cars	Birds	Act
MDML-ℓ_2	0.01	0.01	0.001	0.1	0.01	0.001	0.1
MDML-ℓ_1 [10]	0.01	0.01	0.1	0.01	0.01	0.01	0.001
MDML-$\ell_{2,1}$ [11]	0.001	0.01	0.001	0.001	0.1	0.1	0.1
MDML-Tr [13]	0.01	0.01	0.01	0.001	0.1	0.01	0.1
MDML-IT [3]	0.001	0.1	0.1	0.01	0.1	0.001	0.01
MDML-Drop [14]	0.01	0.01	0.1	0.001	0.1	0.1	0.01
PDML-DC [39]	0.01	0.1	0.01	0.01	0.1	0.1	0.01
PDML-CS [33]	0.01	0.1	0.01	1	0.001	0.001	0.1
PDML-DPP [34]	0.1	0.01	0.001	0.1	0.1	0.01	0.1
PDML-IC [35]	0.01	0.001	0.01	0.1	0.01	0.1	0.01
PDML-DeC [36]	0.1	0.001	0.01	0.1	0.01	0.1	1
PDML-VGF [40]	0.01	0.01	0.1	0.1	0.1	0.001	0.01
PDML-MA [17]	0.001	1	0.01	0.01	0.1	0.01	0.01
PDML-SFN [15, 20]	0.01	0.01	0.01	0.1	0.1	0.01	0.1
PDML-VND [47]	0.01	0.1	0.01	0.001	0.001	0.1	0.01
PDML-LDD [47]	0.001	0.01	0.1	0.001	0.01	0.01	0.01
MDML-CSFN	0.01	0.001	0.01	0.001	0.1	0.01	0.1
MDML-CVND	0.01	0.01	0.1	0.1	0.1	0.01	0.01
MDML-CLDD	0.01	0.1	0.01	0.001	0.01	0.001	0.1

Table 1: Best tuned regularization parameters via cross validation.

whose performance is evaluated using the Area Under precision-recall Curve (AUC) [45] which is the higher, the better. Note that the learned distance metrics can also be applied to other tasks such as clustering and classification. Due to the space limit, we focus on retrieval. 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) [39], cosine similarity (CS) [33], determinantal point process (DPP) [38, 34], InCoherence (IC) [35], variational Gram function (VGF) [46, 40], decorrelation (DeC) [36], mutual angles (MA) [37], squared Frobenius norm (SFN) [15, 21, 16, 20], von Neumann divergence (VND) [47], log-determinant divergence (LDD) [47], and orthogonal constraint (OC) $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$ [22, 26]. 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 [10], $\ell_{2,1}$ norm [11], trace norm (Tr) [13], information theoretic (IT) regularizer $-\text{logdet}(\mathbf{M}) + \text{tr}(\mathbf{M})$ [3], and Dropout (Drop) [48]. All these regularizers are applied to MDML. One common way of dealing with class-imbalance is *over-sampling* (OS) [49], which repetitively draws samples from the empirical distributions of infrequent classes until all classes have the same number of samples. We apply this technique to PDML and MDML. In addition, we compare with vanilla Euclidean distance (EUC) and other distance learning methods including large margin nearest neighbor (LMNN) metric learning, information theoretic metric learning (ITML) [3], logistic discriminant metric learning (LDML) [4], metric learning from equivalence constraints (MLEC) [6], geometric mean metric learning (GMML) [7], and independent Laplacian hashing with diversity (ILHD) [29]. The PDML-based methods except PDML-OC are solved with stochastic subgradient descent (SSD). PDML-OC is solved using the algorithm proposed in [50]. The MDML-based methods are solved with proximal SSD. The learning rate is set to 0.001. The mini-batch size is set to 100 (50 similar pairs and 50 dissimilar pairs). We use 5-fold cross validation to tune the regularization parameter among $\{10^{-3}, \cdots, 10^0\}$ and the number of projection vectors (of the PDML methods) among $\{50, 100, 200, \cdots, 500\}$. In CVND and CLDD, ϵ is set to be $1e-5$. The margin t is set to be 1. In the MDML-based methods, after the Mahalanobis matrix $\mathbf{M}$ (rank R) is learned, we factorize it into $\mathbf{M} = \mathbf{L}^\top\mathbf{L}$ where $\mathbf{L} \in \mathbb{R}^{R\times D}$ (see supplements), then perform retrieval based on $\|\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{y}\|_2^2$, which is more efficient than that based on $(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})$. Each method is implemented on top of GPU using the MAGMA library. The experiments are conducted on a GPU-cluster with 40 machines.

Retrieval settings For each test example, we use it to query the rest of test examples based on the learned distance metric. If the distance between $\mathbf{x}$ and $\mathbf{y}$ is smaller than a threshold s and they have the same class label, then this is a true positive. By choosing different values of s, we obtain a receiver operating

	MIMIC	EICU	Reuters
PDML	0.654± 0.015	0.690± 0.009	0.963± 0.012
MDML	0.659± 0.014	0.691± 0.005	0.962± 0.008
EUC	0.558± 0.007	0.584± 0.008	0.887± 0.009
LMNN [3]	0.643± 0.011	0.678± 0.007	0.951± 0.020
LDML [4]	0.638± 0.017	0.678± 0.020	0.946± 0.009
MLEC [6]	0.633± 0.018	0.692± 0.008	0.936± 0.007
GMML [7]	0.621± 0.017	0.679± 0.006	0.938± 0.011
ILHD [29]	0.590± 0.006	0.652± 0.018	0.919± 0.014
MDML-ℓ_2	0.664± 0.019	0.706± 0.006	0.966± 0.012
MDML-ℓ_1 [10]	0.664± 0.017	0.715± 0.015	0.967± 0.005
MDML-$\ell_{2,1}$ [11]	0.658± 0.008	0.727± 0.016	0.970± 0.008
MDML-Tr [13]	0.672± 0.011	0.709± 0.004	0.969± 0.015
MDML-IT [3]	0.673± 0.009	0.705± 0.007	0.964± 0.007
MDML-Drop [14]	0.660± 0.016	0.718± 0.006	0.968± 0.010
MDML-OS	0.665± 0.009	0.711± 0.007	0.968± 0.012
PDML-DC [39]	0.662± 0.005	0.717± 0.012	0.976± 0.007
PDML-CS [33]	0.676± 0.019	0.736± 0.007	0.973± 0.011
PDML-DPP [34]	**0.679**± 0.008	0.725± 0.010	0.972± 0.015
PDML-IC [35]	0.674± 0.010	0.726± 0.005	0.984± 0.019
PDML-DeC [36]	0.666± 0.007	0.711± 0.015	0.977± 0.011
PDML-VGF [40]	0.674± 0.007	0.730± 0.011	0.988± 0.008
PDML-MA [17]	0.670± 0.009	0.731± 0.006	0.983± 0.007
PDML-SFN [15, 21, 16, 20]	0.677± 0.011	0.736± 0.013	0.984± 0.009
PDML-OC [22, 26]	0.663± 0.005	0.716± 0.010	0.966± 0.017
PDML-OS	0.658± 0.006	0.691± 0.004	0.965± 0.009
PDML-VND [47]	0.676± 0.013	0.748± 0.020	0.983± 0.007
PDML-LDD [47]	0.674± 0.012	0.743± 0.006	0.981± 0.009
MDML-CSFN	**0.679**± 0.009	0.741± 0.011	0.991± 0.010
MDML-CVND	0.678± 0.007	0.744± 0.005	**0.994**± 0.008
MDML-CLDD	0.678± 0.012	**0.750**± 0.006	0.991± 0.006

Table 2: Mean AUC and standard errors on frequent classes.

	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	0.008	0.019	0.014	0.007	0.009	0.010	0.005	0.022	0.017	0.005	0.021	0.006	0.016
MDML	0.020	0.006	0.024	0.009	0.016	0.009	0.011	0.015	0.012	0.008	0.017	0.013	0.021
EUC	0.008	0.005	0.012	0.010	0.006	0.015	0.017	0.006	0.008	0.024	0.016	0.021	0.010
LMNN [3]	0.013	0.022	0.009	0.011	0.016	0.009	0.014	0.018	0.022	0.020	0.011	0.017	0.008
LDML [4]	0.025	0.014	0.023	0.008	0.005	0.012	0.024	0.007	0.011	0.010	0.008	0.011	0.005
MLEC [6]	0.012	0.018	0.016	0.011	0.017	0.020	0.005	0.021	0.007	0.019	0.007	0.023	0.013
GMML [7]	0.008	0.011	0.020	0.021	0.024	0.013	0.016	0.011	0.009	0.008	0.013	0.007	0.010
ILHD [29]	0.013	0.017	0.007	0.010	0.022	0.004	0.013	0.020	0.006	0.018	0.011	0.015	0.012
MDML-ℓ_2	0.016	0.011	0.013	0.021	0.005	0.013	0.007	0.023	0.016	0.022	0.007	0.021	0.025
MDML-ℓ_1 [10]	0.018	0.020	0.006	0.013	0.018	0.014	0.023	0.006	0.013	0.017	0.022	0.018	0.009
MDML-$\ell_{2,1}$ [11]	0.012	0.008	0.017	0.016	0.012	0.022	0.015	0.014	0.020	0.012	0.024	0.019	0.015
MDML-Tr [13]	0.011	0.024	0.009	0.022	0.007	0.011	0.012	0.007	0.015	0.013	0.009	0.018	0.010
MDML-IT [3]	0.013	0.009	0.017	0.020	0.016	0.021	0.015	0.017	0.013	0.019	0.011	0.008	0.016
MDML-Drop [14]	0.005	0.014	0.008	0.027	0.013	0.016	0.005	0.023	0.009	0.008	0.006	0.024	0.025
PDML-DC [39]	0.008	0.017	0.019	0.006	0.015	0.009	0.011	0.012	0.018	0.014	0.017	0.023	0.008
PDML-CS [33]	0.019	0.022	0.017	0.021	0.023	0.010	0.007	0.020	0.016	0.012	0.013	0.014	0.022
PDML-DPP [34]	0.014	0.006	0.011	0.009	0.008	0.017	0.018	0.007	0.013	0.011	0.006	0.022	0.005
PDML-IC [35]	0.007	0.009	0.011	0.006	0.014	0.015	0.006	0.017	0.023	0.007	0.005	0.019	0.008
PDML-DeC [36]	0.019	0.024	0.021	0.008	0.006	0.009	0.015	0.018	0.006	0.014	0.008	0.012	0.018
PDML-VGF [40]	0.009	0.008	0.017	0.013	0.019	0.010	0.015	0.009	0.014	0.008	0.022	0.021	0.008
PDML-MA [17]	0.021	0.014	0.009	0.005	0.019	0.021	0.011	0.014	0.016	0.013	0.011	0.007	0.009
PDML-SFN [15, 21, 16, 20]	0.015	0.021	0.006	0.022	0.007	0.017	0.013	0.010	0.008	0.023	0.016	0.024	0.012
PDML-OC [22, 26]	0.016	0.010	0.011	0.007	0.018	0.008	0.019	0.023	0.016	0.015	0.011	0.005	0.009
PDML-VND [47]	0.009	0.018	0.007	0.024	0.011	0.019	0.021	0.017	0.022	0.014	0.006	0.012	0.025
PDML-LDD [47]	0.021	0.012	0.008	0.018	0.017	0.013	0.011	0.007	0.009	0.007	0.012	0.006	0.016
MDML-CSFN	0.011	0.009	0.013	0.007	0.008	0.014	0.009	0.012	0.008	0.025	0.007	0.004	0.011
MDML-CVND	0.006	0.007	0.011	0.012	0.014	0.009	0.012	0.013	0.006	0.009	0.011	0.014	0.013
MDML-CLDD	0.009	0.012	0.011	0.010	0.005	0.013	0.018	0.005	0.012	0.011	0.015	0.008	0.010

Table 3: Standard errors.

characteristic (ROC) curve. For AUC on infrequent classes, we use examples belonging to infrequent classes to query the entire test set (excluding the query). AUC on frequent classes is measured in a similar way.

For computational efficiency, in MDML-based methods, we do not use $(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})$ to compute distance directly. Given the learned matrix $\mathbf{M}$ (which is of rank k), we can decompose it into $\mathbf{L}^\top\mathbf{L}$ where $\mathbf{L}\in\mathbb{R}^{k\times d}$. Let $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be the eigen-decomposition of $\mathbf{M}$. Let $\lambda_1,\cdots,\lambda_k$ denote the k nonzero eigenvalues and $\mathbf{u}_i,\cdots,\mathbf{u}_k$ denote the corresponding eigenvectors. Then $\mathbf{L}$ is the transpose of $[\sqrt{\sigma_1}\mathbf{u}_1,\cdots,\sqrt{\sigma_k}\mathbf{u}_k]$. Given $\mathbf{L}$, we can use it to transform each input d-dimensional feature vector $\mathbf{x}$ into a new k-dimensional vector $\mathbf{Lx}$, then perform retrieval on the new vectors based on Euclidean distance. Note that only when computing Euclidean distance between $\mathbf{Lx}$ and $\mathbf{Ly}$, we have that $\|\mathbf{Lx}-\mathbf{Ly}\|_2^2$ is equivalent to $(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})$. For other distances or similarity measures between $\mathbf{Lx}$ and $\mathbf{Ly}$, such as L1 distance and cosine similarity, this does not hold. Performing retrieval based on $\|\mathbf{Lx}-\mathbf{Ly}\|_2^2$ is more efficient than that based on $(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})$ when k is smaller than d. Given m test examples, the computation complexity of $\|\mathbf{Lx}-\mathbf{Ly}\|_2^2$ based retrieval is $O(mkd+m^2k)$, while that of $(\mathbf{x}-\mathbf{y})^\top\mathbf{M}(\mathbf{x}-\mathbf{y})$ based retrieval is $O(m^2d^2)$.

Additional details of baselines In the Large Margin Nearest Neighbor (LMNN) DML method [2], there is a nonconvex formulation and a convex formulation. We used the convex one. Though the variational Gram function (VGF) [40] is convex, when it is used to regularize PDML, the overall problem is non-convex and it is unclear how to seek a convex relaxation. In Geometric Mean Metric Learning (GMML) [7], the prior matrix was set to an identity matrix. In Independent Laplacian Hashing with Diversity (ILHD) [29], we use the ILTitf variant. The hash function is kernel SVM with a radial basis function (RB) kernel. We did not compare with unsupervised hashing methods [23, 27, 24, 28, 25].

Hyperparameters Table 1 shows the best tuned regularization parameters on different datasets. In Orthogonal Constraints (OR) [22, 26], there is no regularization parameter. In dropout [14], the regularization parameter designates the probability of dropping elements in the Mahalanobis matrix. In LMNN, the weighting parameter μ was set to 0.5. In GMML [7], the regularization parameter λ was set to 0.1. The step length t of geodesic was set to 0.3. In ILHD [29], the scale parameter of the RBF kernel was set to 0.1.

	MIMIC	EICU	Reuters	News	Cars	Birds	Act
PDML	0.175	0.145	0.043	0.095	0.149	0.075	0.045
MDML	0.187	0.142	0.045	0.087	0.124	0.066	0.042
LMNN	0.183	0.153	0.031	0.093	0.153	0.073	0.013
LDML	0.159	0.139	0.034	0.079	0.131	0.072	0.068
MLEC	0.162	0.131	0.042	0.088	0.151	0.039	0.043
GMML	0.197	0.157	0.051	0.063	0.118	0.067	0.036
ILHD	0.164	0.162	0.048	0.077	0.117	0.045	0.059
MDML-ℓ_2	0.184	0.136	0.037	0.072	0.105	0.053	0.041
MDML-ℓ_1	0.173	0.131	0.042	0.064	0.113	0.061	0.026
MDML-$\ell_{2,1}$	0.181	0.129	0.034	0.073	0.121	0.044	0.024
MDML-Tr	0.166	0.138	0.024	0.076	0.111	0.058	0.037
MDML-IT	0.174	0.134	0.033	0.061	0.109	0.036	0.013
MDML-Drop	0.182	0.140	0.021	0.076	0.114	0.063	0.024
MDML-OS	0.166	0.133	0.032	0.063	0.108	0.057	0.031
PDML-DC	0.159	0.131	0.035	0.069	0.127	0.064	0.035
PDML-CS	0.163	0.135	0.031	0.083	0.103	0.045	0.033
PDML-DPP	0.147	0.140	0.038	0.067	0.117	0.072	0.041
PDML-IC	0.155	0.127	0.018	0.075	0.116	0.074	0.029
PDML-DeC	0.164	0.123	0.023	0.082	0.125	0.051	0.033
PDML-VGF	0.158	0.136	0.014	0.064	0.136	0.035	0.028
PDML-MA	0.143	0.128	0.023	0.078	0.102	0.031	0.042
PDML-OC	0.161	0.142	0.032	0.061	0.111	0.063	0.034
PDML-OS	0.169	0.137	0.015	0.083	0.119	0.058	0.042
PDML-SFN	0.153	0.126	0.022	0.069	0.127	0.043	0.028
PDML-VND	0.148	0.135	0.019	0.078	0.116	0.067	0.035
PDML-LDD	0.146	0.121	0.017	0.054	0.111	0.036	0.021
MDML-CSFN	0.142	0.124	0.019	0.062	0.092	0.043	0.019
MDML-CVND	0.137	0.115	0.008	0.055	0.094	0.038	0.013
MDML-CLDD	0.131	0.118	0.012	0.058	0.089	0.026	0.016

Table 4: The gap of training AUC and testing AUC (training-AUC minus testing-AUC)

	MIMIC	EICU	Reuters	News	Cars	Birds	Act
LMNN	3.8	4.0	0.4	0.7	0.6	0.7	0.3
ITML	12.6	11.4	1.2	3.2	3.0	2.7	0.8
LDML	3.7	3.4	0.3	0.6	0.5	0.6	0.2
MLEC	0.4	0.4	0.026	0.049	0.043	0.044	0.018
GMML	0.5	0.4	0.035	0.056	0.052	0.049	0.022
MDML-ℓ_2	3.4	3.5	0.3	0.6	0.5	0.6	0.2
MDML-ℓ_1	3.4	3.6	0.5	0.6	0.5	0.6	0.2
MDML-$\ell_{2,1}$	3.5	3.7	0.3	0.5	0.5	0.6	0.1
MDML-Tr	3.4	3.7	0.3	0.6	0.6	0.4	0.3
MDML-IT	5.2	5.5	0.5	0.9	0.8	1.0	0.4
MDML-Drop	9.5	10.4	1.2	1.7	1.9	1.7	0.6

Table 5: Training time (hours) of additional baselines.

Reduced dimension	1000	2000	3000
PDML	0.634	0.637	0.629
MDML	0.641	0.638	0.646
PDML-VND	0.667	0.669	0.671
PDML-CVND	0.672	0.675	0.673

Table 6: Mean AUC under different reduced dimensions.

6.3 Additional Experimental Results

Training time of other baselines Table 5 shows the training time of additional baselines.

AUC on frequent classes Table 2 shows the mean AUC and standard errors on frequent classes. MDML-(CSFN,CVND,CLDD) achieve better mean AUC than the baselines.

Standard errors Table 3 shows the standard errors of AUC on all classes and infrequent classes and standard errors of balance scores.

Gap between training AUC and testing AUC Table 4 shows the gap between training AUC and testing AUC (training-AUC minus testing-AUC).

Additional experimental analysis

- **Training time** Unregularized PDML runs faster that regularized PDML methods because it has no need to tune the regularization parameter, which reduces the number of experimental runs by 4 times. Unregularized MDML runs faster than regularized MDML methods because it has no need to tune the regularization parameter or the number of projection vectors, which reduces the number of experimental runs by 12 times. PDML-(DC,DPP,VND,LDD) takes longer time than other regularized PDML methods since they need eigendecomposition to compute the gradients. PDML-OC has no regularization parameter to tune, hence its number of experimental runs is 4 times fewer than other regularized PDML methods.
- **Balance** In most DML methods, the AUC on infrequent classes is worse than that on frequent classes, showing that DML is sensitive to the imbalance of pattern-frequency, tends to be biased towards frequent patterns and is less capable to capture infrequent patterns. This is in accordance with previous study [37].

Dimension reduction We study whether using PCA to reduce dimensionality of features would hurt performance. We set the reduced dimension to 1000, 2000, 3000 and measure the performance of four methods: PDML, MDML, PDML-VND, PDML-CVND. Table 6 shows the mean AUC on all classes of the MIMIC dataset. As can be seen, the AUCs under different dimensions have no significant difference, suggesting that 1000 dimensions are enough to retain the information of data.

References

[1] Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *NIPS*, 2002.

[2] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005.

[3] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[4] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[5] Yiming Ying and Peng Li. Distance metric learning with eigenvalue optimization. *JMLR*, 2012.

[6] M Kostinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[7] Pourya Habib Zadeh, Reshad Hosseini, and Suvrit Sra. Geometric mean metric learning. In *ICML*, 2016.

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

[9] Fei Wang and Jimeng Sun. Survey on distance metric learning and dimensionality reduction in data mining. *Data Mining and Knowledge Discovery*, 2015.

[10] Guo-Jun Qi, Jinhui Tang, Zheng-Jun Zha, Tat-Seng Chua, and Hong-Jiang Zhang. An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization. In *ICML*, 2009.

[11] Yiming Ying, Kaizhu Huang, and Colin Campbell. Sparse metric learning via smooth optimization. In *NIPS*, 2009.

[12] Gang Niu, Bo Dai, Makoto Yamada, and Masashi Sugiyama. Information-theoretic semisupervised metric learning via entropy regularization. *Neural Computation*, 2012.

[13] Wei Liu, Cun Mu, Rongrong Ji, Shiqian Ma, John R Smith, and Shih-Fu Chang. Low-rank similarity metric learning in high dimensions. In *AAAI*, 2015.

[14] Qi Qian, Juhua Hu, Rong Jin, Jian Pei, and Shenghuo Zhu. Distance metric learning using dropout: a structured regularization approach. In *KDD*, 2014.

[15] Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Semi-supervised hashing for large-scale search. *TPAMI*, 2012.

[16] Tiezheng Ge, Kaiming He, and Jian Sun. Graph cuts for supervised binary coding. In *ECCV*, 2014.

[17] Pengtao Xie. Learning compact and effective distance metrics with diversity regularization. In *ECML*, 2015.

[18] Wenbin Yao, Zhenyu Weng, and Yuesheng Zhu. Diversity regularized metric learning for person re-identification. In *ICIP*, 2016.

[19] Ramin Raziperchikolaei and Miguel A Carreira-Perpinán. Learning independent, diverse binary hash functions: Pruning and locality. *ICDM*, 2016.

[20] Yong Chen, Hui Zhang, Yongxin Tong, and Ming Lu. Diversity regularized latent semantic match for hashing. *Neurocomputing*, 2017.

[21] Xiping Fu, Brendan McCane, Steven Mills, and Michael Albert. Nokmeans: Non-orthogonal k-means hashing. In *ACCV*, 2014.

[22] Wei Liu, Steven Hoi, and Jianzhuang Liu. Output regularized metric learning with side information. *ECCV*, 2008.

[23] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *NIPS*, 2009.

[24] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *TPAMI*, pages 2916–2929, 2013.

[25] Jianqiu Ji, Shuicheng Yan, Jianmin Li, Guangyu Gao, Qi Tian, and Bo Zhang. Batch-orthogonal locality-sensitive hashing for angular similarity. *TPAMI*, 2014.

[26] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In *IJCAI*, 2015.

[27] Weihao Kong and Wu-Jun Li. Isotropic hashing. In *NIPS*, 2012.

[28] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization. *TPAMI*, 2014.

[29] Miguel A Carreira-Perpinán and Ramin Raziperchikolaei. An ensemble diversity approach to supervised binary hashing. In *NIPS*, 2016.

[30] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 2003.

[31] Robert E Banfield, Lawrence O Hall, Kevin W Bowyer, and W Philip Kegelmeyer. Ensemble diversity measures and their application to thinning. *Information Fusion*, 2005.

[32] Ioannis Partalas, Grigorios Tsoumakas, and Ioannis P Vlahavas. Focused ensemble selection: A diversity-based method for greedy ensemble selection. In *European Conference on Artificial Intelligence*, 2008.

[33] Yang Yu, Yu-Feng Li, and Zhi-Hua Zhou. Diversity regularized machine. In *IJCAI*, 2011.

[34] James Y Zou and Ryan P Adams. Priors for diversity in generative latent variable models. In *NIPS*, 2012.

[35] Yebo Bao, Hui Jiang, Lirong Dai, and Cong Liu. Incoherent training of deep neural networks to decorrelate bottleneck features for speech recognition. In *ICASSP*, 2013.

[36] Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. *ICLR*, 2015.

[37] Pengtao Xie, Yuntian Deng, and Eric P. Xing. Diversifying restricted boltzmann machine for document modeling. In *KDD*, 2015.

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

[39] Jonathan Malkin and Jeff Bilmes. Ratio semi-definite classifiers. In *ICASSP*, 2008.

[40] Amin Jalali, Lin Xiao, and Maryam Fazel. Variational gram functions: Convex analysis and optimization. *arXiv preprint arXiv:1507.04734*, 2015.

[41] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

[42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[45] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[46] Dengyong Zhou, Lin Xiao, and Mingrui Wu. Hierarchical classification via orthogonal transfer. *ICML*, 2011.

[47] Pengtao Xie, Barnabas Poczos, and Eric P Xing. Near-orthogonality regularization in kernel methods. *UAI*, 2017.

[48] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[49] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics*, 2012.

[50] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.