Generalization Analysis of Multi-Modal Metric Learning

Yunwen Lei† and Yiming Ying‡

†State Key Lab of Software Engineering,
School of Computer, Wuhan University, Wuhan 430072, China
ywlei@whu.edu.cn
‡College of Engineering, Mathematics and Physical Sciences,
University of Exeter, Exeter, EX4 4QF, UK
mathying@gmail.com

Abstract

Multi-modal metric learning has recently received considerable attention since many real-world applications involve multi-modal data. However, there is relatively little study on the generalization analysis of the associated learning algorithms. In this paper, we bridge this theoretical gap by deriving its generalization bounds using Rademacher complexities. In particular, we establish a general Rademacher complexity result by systematically analyzing the behavior of the resulting models with various regularizers, e.g., ℓ_p-regularizer on the modality level with either a mixed (q, s)-norm or a Schatten norm on each modality. Our results and the discussion followed help to understand how the prior knowledge can be exploited by selecting an appropriate regularizer.

1 Introduction

Nowadays, many real-world applications often involve multi-modal data, where the information comes from multiple heterogeneous sources [12, 15]. For example, songs in music social networks can be described by acoustic features (e.g., rhythm and timbre), semantic features (e.g., tags, lyrics), and social features (e.g., collaborative filtering, biographies) [12]. The definition of distance metric in the multi-modal context becomes a key challenge since it forms the foundation for many machine learning algorithms such as k-nearest neighbor classification and k-means clustering. Simply applying single-modal metric learning methods in this case may lead to suboptimal performance since it fails to consider the dependency and complementarity relationships among different modalities [14, 16]. Also, the notion of distance metric may not be consistent on distinct modalities and thus there is generally no obvious approach to establishing a unified metric space which optimally integrates heterogeneous data [12]. Therefore, it is imperative to find a more involved strategy to tackle multi-modal data.

In view of this, McFee and Lanckriet [12] pioneered the work on multi-modal metric learning and they applied the multiple kernel learning technique for integrating heterogeneous data into a single unified similarity space. Xia et al. [15] proposed an online learning method to tackle the efficiency and scalability issues of the multi-modal learning framework in [12]. Xie and Xing [16] provided a principled methodology to embed data of arbitrary modalities into a single latent space where the distance metric can be learned under proper supervision. Wu et al. [14] used deep learning and online learning techniques to learn flexible nonlinear similarity functions for images with multi-modal representations. Empirical studies show that these methods work well in measuring dissimilarity/similarity for multi-modal data [12, 14, 15, 16].

Despite various multi-modal metric learning methods have been proposed, the theoretical result on their generalization performance remains largely un-explored. In this paper, we will initiate this exploration by presenting a novel generalization analysis for multi-modal metric learning. Specifically, we provide a novel Rademacher complexity in the context of multi-modal metric learning and show how it can be used to study the generalization performance for the resulting models. We also propose a general technique to estimate Rademacher complexities for matrix classes characterized by strongly convex functions, which allows us to systematically derive generalization error bounds under different regularization terms, e.g., ℓ_p-regularizer at the modality level with trace norm (or more generally Schatten norms) on each modality [12, 15], and ℓ_p-regularizer at the modality level with Frobenius norm (or more generally mixed (q,s)-norms) on each modality [14]. Our investigation also helps to illuminate the influence of the regularizer on the behavior of the resulting models. Our results generalize and refine the generalization analysis in [4] which was developed for the single-modal metric learning.

This paper is organized as follows. Section 2 formulates multi-modal metric learning problems. In Section 3, we establish the generalization bounds and present a general result on estimating Rademacher complexities. Section 4 provides the application of these results to some specific regularization learning schemes. Section 5 exhibits some discussions. Conclusions and possible directions for future research are presented in Section 6.

Notations. Let $\mathbb{N}_n = \{1, 2, \dots, n\}$ for any $n \in \mathbb{N}$. The positive part of $x \in \mathbb{R}$ is denoted by $[x]_+ := \max(x, 0)$. For any matrices X, Y of the same size, the inner product is $\langle X, Y\rangle := \mathrm{Tr}(X^\top Y)$, where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix and $X^\top$ means the transpose of X. For any norm $\|\cdot\|$ on matrices, its dual norm is defined by $\|M\|_* = \sup\{\langle X, M\rangle : \|X\| \leq 1\}$. The ℓ_q-norm of a vector $x \in \mathbb{R}^n$ is given by $\|x\|_q = (\sum_{i\in\mathbb{N}_n} |x_i|^q)^{1/q}, q \geq 1$. Given a matrix M, we denote by $\sigma(M)$ the vector consisting of the singular values of M in a non-increasing order, and the q-Schatten norm $\|M\|_{S(q)}$ is defined as the ℓ_q-norm of $\sigma(M)$. For any $A = (\alpha^1, \alpha^2, \dots, \alpha^m) \in \mathbb{R}^{n\times m}$, the mixed (q,s)-norm of A is

$$\|A\|_{q,s} := \big\|\big(\|\alpha^1\|_q, \|\alpha^2\|_q \dots, \|\alpha^m\|_q\big)\big\|_s, \qquad \forall q, s \geq 1.$$

For any $d, m \in \mathbb{N}$, introduce the following class of matrices of size $d \times (md)$:

$$\mathbb{S}^{d\times(md)} := \{(M^1, \dots, M^m) : M^l \in \mathbb{R}^{d\times d}, (M^l)^\top = M^l, l \in \mathbb{N}_m\}.$$

We call (p, q) a dual pair (or p is the dual exponent of q) if $1/p + 1/q = 1$. A norm $\|\cdot\|$ is said to be absolutely symmetric if $\|x\|$ remains invariant under arbitrary permutations and sign changes of the components of x. All the norms considered in this paper are absolutely symmetric. For a convex function f, we denote by f^* its Fenchel conjugate, i.e.,

$$f^*(x) := \sup_y [\langle x, y\rangle - f(y)].$$

2 Formulation of multi-modal metric learning

In the context of multi-modal metric learning, we assume that each input data has m different modalities and each modality is encoded by a vector of length d (If the vector lengths of different modalities are not equal, one can simply add extra zeros to those modality vectors with smaller length). Therefore, every input data x has the following group structure:

$$x = ((x^1)^\top, (x^2)^\top, \dots, (x^m)^\top)^\top \in \mathcal{X} \subset \mathbb{R}^{dm}, \qquad x^l \in \mathbb{R}^d,$$

where $\mathcal{X}$ is called the input space. Suppose that the side-information available to us on x is its label $y \in \mathcal{Y} := \{0, 1\}$ and denote by $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ the sample space. Given a sequence of training data $\boldsymbol{z} := \{z_i = (x_i, y_i) \in \mathcal{Z}, i \in \mathbb{N}_n\}$ drawn independently from a measure ρ on $\mathcal{Z}$, our aim is to learn a distance metric so that any two samples with the same label admit a relatively small distance and those two lying in different classes admit a relatively large distance [17, 20]. For any two samples x_i, x_j, suppose that the distance on the l-th modality is the Mahalanobis distance captured by a positive semi-definite matrix $M^l \in \mathbb{S}^{d\times d}$:

$$d_{M^l}(x_i^l, x_j^l) := (x_i^l - x_j^l)^\top M^l (x_i^l - x_j^l).$$

We now catenate these sub-matrices $M^1, \dots, M^m$ into a large matrix $M \in \mathbb{S}^{d\times(md)}$ with the block structure $M := (M^1, M^2, \dots, M^m)$. Then the total distance between x_i and x_j is defined as the sum of the distances over all modalities:

$$d_M(x_i, x_j) := \sum_{l=1}^{m} d_{M^l}(x_i^l, x_j^l) = \sum_{l=1}^{m} (x_i^l - x_j^l)^\top M^l (x_i^l - x_j^l). \tag{1}$$

For any pair of examples (x_i, x_j), let $r(y_i, y_j) = 1$ if $y_i = y_j$ and $r(y_i, y_j) = -1$ otherwise. Similar to the work in [4], we use the following empirical error to quantify the empirical behavior of $(M, b) \in \mathbb{S}^{d\times(md)} \times \mathbb{R}$

$$\mathcal{E}_{\boldsymbol{z}}(M, b) := \frac{1}{n(n-1)} \sum_{i,j\in\mathbb{N}_n, i\neq j} [1 + r(y_i, y_j)(d_M(x_i, x_j) - b)]_+,$$

where the introduction of the offset term b is based on the intuition that those samples with distances smaller than an appropriately chosen threshold b are likely to lie in the same class and vice versa [6].

We consider here the regularization learning framework where a penalty $\|M\|^2$ is added to the empirical error to control the complexity of M:

$$(M_{\boldsymbol{z}}, b_{\boldsymbol{z}}) := \arg \min_{M\in\mathbb{S}^{d\times(md)}, b\in\mathbb{R}} [\mathcal{E}_{\boldsymbol{z}}(M, b) + \lambda\|M\|^2]. \tag{2}$$

This formulation is a natural extension of the single-modal metric learning algorithm considered in [4, 6]. Furthermore, equation (2) is also analogous to the multi-modal metric learning algorithms developed in [12, 14], with slight difference in the construction of the loss term: we use class labels as the side information to express distance constraints, while [12, 14] used the relative comparisons. The trade-off between the penalty and the empirical error in equation (2) is controlled by the regularization parameter $\lambda > 0$. Here restricting M in $\mathbb{S}^{d\times(md)}$ guarantees that the distance metric is symmetric, i.e.,

$$d_M(x_i, x_j) = d_M(x_j, x_i), \qquad \forall x_i, x_j \in \mathcal{X}.$$

When the model $(M_{\boldsymbol{z}}, b_{\boldsymbol{z}})$ is derived from equation (2), the true error (expected risk) to measure its quality is defined by

$$\mathcal{E}(M, b) := \iint [1 + r(y, y^{'})(d_M(x, x^{'}) - b)]_+ \mathrm{d}\rho(x, y)\mathrm{d}\rho(x^{'}, y^{'}).$$

3 Generalization errors and Rademacher complexities

In this section, we try to estimate the generalization error of $(M_{\boldsymbol{z}}, b_{\boldsymbol{z}})$ via a notion called Rademacher complexity [2]. We also provide a general result (Theorem 5) on Rademacher complexities when the involved class can be controlled by a strongly convex function. We begin our discussion with a lemma controlling the solution space of the regularization problem (2). This result was originally established in the single-modal case, however, it is not hard to show that this is also the case for multi-modal metric learning with exactly the same proof.

Lemma 1 ([4]). *Suppose that the sample $\boldsymbol{z}$ contains at least two examples with the same label and at least two examples with different labels, then any minimizer $(M_{\boldsymbol{z}}, b_{\boldsymbol{z}})$ of problem (2) would satisfy the inequality*

$$\|M_{\boldsymbol{z}}\| \leq 1/\sqrt{\lambda},$$
$$|b_{\boldsymbol{z}}| \leq 1 + \max_{i\neq j} d_{M_{\boldsymbol{z}}}(x_i, x_j).$$

It is clear that the assumptions of Lemma 1 are very mild. Indeed, the first part would be automatically satisfied if $n \geq 3$, while the violation of the second assumption means that all examples would belong to the same class, for which there is nothing interesting to learn as the side information accessible to us is the same for all examples. In the sequel, without loss of generality we will always assume that the assumptions of Lemma 1 hold true.

For any $x_i, x_j \in \mathcal{X}$ and any $M = (M^1, \ldots, M^m) \in \mathbb{R}^{d\times(md)}$, the distance metric $d_M(x_i, x_j)$ satisfies the following inequality

$$\begin{aligned} d_M(x_i, x_j) &= \sum_{l=1}^{m} \big\langle (x_i^l - x_j^l)(x_i^l - x_j^l)^\top, M^l \big\rangle \\ &= \big\langle \big((x_i^1 - x_j^1)(x_i^1 - x_j^1)^\top, \ldots, (x_i^m - x_j^m)(x_i^m - x_j^m)^\top\big), M \big\rangle \\ &\le \|\big((x_i^1 - x_j^1)(x_i^1 - x_j^1)^\top, \ldots, (x_i^m - x_j^m)(x_i^m - x_j^m)^\top\big)\|_* \|M\| \\ &\le X_* \|M\|, \end{aligned} \tag{3}$$

where

$$X_* := \sup_{x, x' \in \mathcal{X}} \|\big((x^1 - (x')^1)(x^1 - (x')^1)^\top, \ldots, (x^m - (x')^m)(x^m - (x')^m)^\top\big)\|_*. \tag{4}$$

Combining the above inequality with Lemma 1 together, it is straightforward to see that the regularization minimizer $(M_{\boldsymbol{z}}, b_{\boldsymbol{z}})$ always falls into the class:

$$\mathcal{F} := \{(M, b) \in \mathbb{S}^{d\times(md)} \times \mathbb{R} : \|M\| \le 1/\sqrt{\lambda}, |b| \le 1 + X_*\|M\|\}.$$

The generalization performance of $(M_{\boldsymbol{z}}, b_{\boldsymbol{z}})$ defined in equation (2) relies heavily on the uniform deviation between the empirical risk and expected risk with (M, b) from the class $\mathcal{F}$. As we will see soon, this deviation can be further controlled by the Rademacher complexity of the class $\{M \in \mathbb{S}^{d\times(md)} : \|M\| \le 1/\sqrt{\lambda}\}$.

Definition 1 (Rademacher Complexity [2])**.** Let $\mathcal{M} \subset \mathbb{S}^{d\times(md)}$ be a class of matrices and let $\{\sigma_i : i = 1, \ldots, \lfloor \frac{n}{2} \rfloor\}$ be a sequence of independent Rademacher random variables, that is, $\Pr\{\sigma_i = +1\} = \Pr\{\sigma_i = -1\} = 1/2$. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\} \subset \mathcal{X} \subset \mathbb{R}^{md}$ be an i.i.d. sequence of examples. Then, define

$$\hat{R}_n(\mathcal{M}) := \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E}_\sigma \sup_{M \in \mathcal{M}} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i d_M(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}).$$

The Rademacher complexity for multi-modal metric learning is defined as its expectation: $R_n(\mathcal{M}) = \mathbb{E}\hat{R}_n(\mathcal{M})$.

The following lemma, attributed to Ledoux and Talagrand [10], provides a contraction property of Rademacher averages. The version we present here can be found in Theorem A.6 of [1].

Lemma 2. *Let $T \subset \mathbb{R}^n$ and let $\psi_i : \mathbb{R} \to \mathbb{R}, i = 1, \ldots, n$ be functions such that*

$$|\psi_i(\mu) - \psi_i(v)| \le |\mu - v|, \qquad \forall \mu, v \in \mathbb{R}.$$

Assume that $\{\sigma_i\}_{i\in\mathbb{N}_n}$ is a sequence of i.i.d. Rademacher variables. Then we have

$$\mathbb{E} \sup_{t\in T} \sum_{i=1}^{n} \psi_i(t_i)\sigma_i \le \mathbb{E} \sup_{t\in T} \sum_{i=1}^{n} t_i \sigma_i.$$

We are now ready to present a result showing how the Rademacher complexity defined in Definition 1 is related to the generalization analysis of multi-modal metric learning. Theorem 3 is an extension of Theorem 3 in [4] from single-modal metric learning to the multi-modal case and it can be derived using a similar strategy. For completeness, we provide here a sketched proof only highlighting the distinction from the original one developed in the single-modal case. It should be noted that the Rademacher complexity used here differs from the following complexity introduced in [4]

$$R_n^{'} = \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E} \Big\| \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i (x_i - x_{\lfloor \frac{n}{2} \rfloor + i})(x_i - x_{\lfloor \frac{n}{2} \rfloor + i})^\top \Big\|_*.$$

Also, the constant factor in the inequality (5) is better than the corresponding one in [4] as we use here a better structural result (Lemma 2) on Rademacher complexities.

Theorem 3. *Let $(M_{\boldsymbol{z}}, b_{\boldsymbol{z}})$ be the solution of the regularization formulation* (2). *Then, for any $0 < \delta < 1$, there holds, with probability at least $1-\delta$, that*

$$\mathcal{E}(M_{\boldsymbol{z}}, b_{\boldsymbol{z}}) - \mathcal{E}_{\boldsymbol{z}}(M_{\boldsymbol{z}}, b_{\boldsymbol{z}}) \leq 2R_n(\{M \in \mathbb{S}^{d\times(md)} : \|M\| \leq \lambda^{-1/2}\}) + 2(1 + X_*/\sqrt{\lambda})\frac{1 + 2\sqrt{\log\frac{1}{\delta}}}{\sqrt{\lfloor\frac{n}{2}\rfloor}}. \quad (5)$$

Proof. For any $z = (x, y), z^{'} = (x^{'}, y^{'})$, let

$$\Phi_{M,b}(z, z^{'}) = [1 + r(y, y^{'})(d_M(x, x^{'}) - b)]_+.$$

For any $(M, b) \in \mathcal{F}$, it follows from inequality (4) that

$$\begin{aligned}\sup_{z,z'} \sup_{(M,b)\in\mathcal{F}} \Phi_{M,b}(z, z^{'}) &\leq 1 + \sup_{z,z'} \sup_{(M,b)\in\mathcal{F}} d_M(x, x^{'}) + b \\ &\leq 2(1 + X_*/\sqrt{\lambda}).\end{aligned}$$

Using this inequality and analyzing analogously to Cao et al. [4], the following inequality holds with probability $1-\delta$

$$\sup_{(M,b)\in\mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\boldsymbol{z}}(M, b)] \leq \mathbb{E}_{\boldsymbol{z}} \sup_{(M,b)\in\mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\boldsymbol{z}}(M, b)] + 4(1 + X_*/\sqrt{\lambda}) \left(\frac{2\log(1/\delta)}{n}\right)^{1/2}. \quad (6)$$

Cao et al. [4] essentially derived the following inequality (equation (17) in [4])

$$\mathbb{E}_{\boldsymbol{z}} \sup_{(M,b)\in\mathcal{F}} [\mathcal{E}(M, b) - \mathcal{E}_{\boldsymbol{z}}(M, b)] \leq 2\mathbb{E}_{\boldsymbol{z},\sigma} \frac{1}{\lfloor\frac{n}{2}\rfloor} \sup_{(M,b)\in\mathcal{F}} \sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \sigma_i \Phi_{M,b}(z_i, z_{\lfloor\frac{n}{2}\rfloor+i}),$$

which, coupled with the contraction property of Rademacher averages (Lemma 2 with $\psi_i(t) = [1 + r(y_i, y_{\lfloor\frac{n}{2}\rfloor+i})t]_+$), can be further upper bounded by

$$\begin{aligned}&2\mathbb{E}_{\boldsymbol{z},\sigma} \frac{1}{\lfloor\frac{n}{2}\rfloor} \sup_{(M,b)\in\mathcal{F}} \sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \sigma_i [d_M(x_i, x_{\lfloor\frac{n}{2}\rfloor+i}) - b] \\ &= 2R_n(\{M \in \mathbb{S}^{d\times(md)} : \|M\| \leq \lambda^{-1/2}\}) + \frac{2(1 + X_*/\sqrt{\lambda})}{\lfloor\frac{n}{2}\rfloor} \mathbb{E}_\sigma \Big|\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} \sigma_i\Big| \\ &\leq 2R_n(\{M \in \mathbb{S}^{d\times(md)} : \|M\| \leq \lambda^{-1/2}\}) + \frac{2(1 + X_*/\sqrt{\lambda})}{\sqrt{\lfloor\frac{n}{2}\rfloor}}.\end{aligned}$$

The proof is complete if we plug the above inequality into equation (6). □

As exhibited in Theorem 3, the estimation of Rademacher complexities is quite important to understand the behavior of $(M_{\boldsymbol{z}}, b_{\boldsymbol{z}})$. We now provide a general result on Rademacher complexity bounds by reformulating the distance metric (1) as the inner product between two matrices. Our discussion is based on the elegant Lemma 4 developed in [7], which provides a general technique for tackling Rademacher complexities of linear function classes using the concept of strong convexity.

Definition 2. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be β-strongly convex w.r.t. a norm $\|\cdot\|$ iff $\forall x, y \in \mathcal{X}$ and $\forall \alpha \in (0, 1)$, we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\beta}{2}\alpha(1-\alpha)\|x - y\|^2.$$

Lemma 4 ([7])**.** *Let f be a β-strongly convex function w.r.t. a norm $\|\cdot\|$ and assume that $f^*(0) = 0$. Suppose that $\mathcal{W} = \{w : f(w) \leq f_{max}\}$ and $\|x\|_* \leq X, \forall x \in \mathcal{X}$. Then, for any $\bar{x}_1, \ldots, \bar{x}_n \in \mathcal{X}$, we have*

$$\mathbb{E}_\sigma \left[\sup_{w\in\mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle w, \bar{x}_i\rangle\right] \leq X\sqrt{\frac{2f_{max}}{\beta n}},$$

where $\{\sigma_i\}_{i\in\mathbb{N}_n}$ is a sequence of i.i.d. Rademacher variables.

Theorem 5. *Let F be a β-strongly convex function w.r.t. a norm $\|\cdot\|$ on $\mathbb{R}^{d\times(md)}$ such that $F^*(0)=0$. Suppose that $F(M)\le f_{max}, \forall M\in\mathcal{M}$ and X_* is defined as equation (4), then we have that $R_n(\mathcal{M})\le X_*\sqrt{\frac{2f_{max}}{\beta\lfloor\frac{n}{2}\rfloor}}$.*

Proof. Using the identity $\langle(X^1,\ldots,X^m),(Y^1,\ldots,Y^m)\rangle=\sum_{l=1}^m\langle X^l,Y^l\rangle$, the Rademacher complexity can be expressed as:

$$\hat{R}_n(\mathcal{M})=\frac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_\sigma\sup_{M\in\mathcal{M}}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i d_M(x_i,x_{i+\lfloor\frac{n}{2}\rfloor})$$

$$=\frac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_\sigma\sup_{M\in\mathcal{M}}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i\sum_{l=1}^{m}(x_i^l-x_{i+\lfloor\frac{n}{2}\rfloor}^l)^\top M^l(x_i^l-x_{i+\lfloor\frac{n}{2}\rfloor}^l)$$

$$=\frac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_\sigma\sup_{M\in\mathcal{M}}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i\sum_{l=1}^{m}\Big\langle(x_i^l-x_{i+\lfloor\frac{n}{2}\rfloor}^l)(x_i^l-x_{i+\lfloor\frac{n}{2}\rfloor}^l)^\top,M^l\Big\rangle$$

$$=\frac{1}{\lfloor\frac{n}{2}\rfloor}\mathbb{E}_\sigma\sup_{M\in\mathcal{M}}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\sigma_i\Big\langle\big((x_i^1-x_{i+\lfloor\frac{n}{2}\rfloor}^1)(x_i^1-x_{i+\lfloor\frac{n}{2}\rfloor}^1)^\top,\ldots,(x_i^m-x_{i+\lfloor\frac{n}{2}\rfloor}^m)(x_i^m-x_{i+\lfloor\frac{n}{2}\rfloor}^m)^\top\big),M\Big\rangle.$$

Applying Lemma 4 with $w=M$ and

$$\bar{x}_i=\big((x_i^1-x_{i+\lfloor\frac{n}{2}\rfloor}^1)(x_i^1-x_{i+\lfloor\frac{n}{2}\rfloor}^1)^\top,\ldots,(x_i^m-x_{i+\lfloor\frac{n}{2}\rfloor}^m)(x_i^m-x_{i+\lfloor\frac{n}{2}\rfloor}^m)^\top\big)$$

yields that $\hat{R}_n(\mathcal{M})\le X_*\sqrt{\frac{2f_{\max}}{\beta\lfloor\frac{n}{2}\rfloor}}$. The proof is complete if we take the expectation on both sides. □

4 Estimating Rademacher Complexities

This section is devoted to illustrating how the general Rademacher complexity bounds built in Theorem 5 can be applied to the derivation of generalization error bounds for different regularization schemes. As the Mahalanobis matrix for multi-modal metric learning has the structure $M=(M^1,\ldots,M^m)$, it is natural to use the group norm

$$\|M\|_{\Psi,\Phi}:=\big\|\big(\|M^1\|_\Psi,\|M^2\|_\Psi,\ldots,\|M^m\|_\Psi\big)\big\|_\Phi \tag{7}$$

in the regularization framework (2). Here Ψ is a norm on $\mathbb{R}^{d\times d}$ and Φ is a norm on $\mathbb{R}^m$. The relationship among different modalities is reflected by imposing the norm Φ on the vector $(\|M^l\|_\Psi)_{l\in\mathbb{N}_m}$. If Φ is absolutely symmetric, then the group norm $\|\cdot\|_{\Psi,\Phi}$ is indeed a norm on $\mathbb{R}^{d\times(md)}$ and its dual norm is

$$\|M\|_{\Psi_*,\Phi_*}=\big\|\big(\|M^1\|_{\Psi_*},\|M^2\|_{\Psi_*},\ldots,\|M^m\|_{\Psi_*}\big)\big\|_{\Phi_*},$$

where Ψ_* is the dual norm of Ψ and Φ_* is the dual norm of Φ [8]. This paper always assumes that Φ is an ℓ_p-norm $\|\cdot\|_p, p\ge1$, while Ψ can be either a mixed (q,s)-norm $\|\cdot\|_{q,s}, q,s\ge1$ or a Schatten norm $\|\cdot\|_{S(p)}, p\ge1$.

4.1 Strong convexity of group norms

The key point to apply Theorem 5 here is to construct an appropriate strongly convex function for different instantiations of the group norm $\|\cdot\|_{\Psi,\Phi}$. The following theorem provides us a powerful tool to achieve this aim, and it shows that the square of a group norm can be strongly smooth under some suitable conditions.

Definition 3. An everywhere differentiable function $f:\mathcal{X}\to\mathbb{R}$ is said to be β-strongly smooth w.r.t. $\|\cdot\|$ if $\forall x,y\in\mathcal{X}$ we have

$$f(x+y)\le f(x)+\langle\nabla f(x),y\rangle+\frac{\beta}{2}\|y\|^2.$$

Theorem 6 (Group Norms [9]). *Let Ψ, Φ be absolutely symmetric norms on $\mathbb{R}^{d\times d}$ and $\mathbb{R}^m$, respectively. Let $\Phi^2 \circ \surd : \mathbb{R}^m \to \mathbb{R}^m$ denote the following function*

$$(\Phi^2 \circ \surd)(x) := \Phi^2(\sqrt{|x_1|}, \dots, \sqrt{|x_m|}).$$

Suppose $\Phi^2 \circ \surd$ is a norm on $\mathbb{R}^m$. Furthermore, assume that both Ψ^2 and Φ^2 are β_1- and β_2-strongly smooth w.r.t. Ψ and Φ, respectively. Then, the function $\|\cdot\|^2_{\Psi,\Phi}$ is $(\beta_1+\beta_2)$-strongly smooth w.r.t. $\|\cdot\|_{\Psi,\Phi}$.

Remark 1. Kakade et al. [9] proved Theorem 6 when Ψ is a vector norm rather than a matrix norm. However, a closer look of their proof shows that this is also the case when Ψ is a norm on $\mathbb{R}^{d\times d}$.

Furthermore, the following lemma provides a dual property between convexity and smoothness: a function is strongly convex w.r.t. a norm if and only if its Fenchel conjugate is strongly smooth w.r.t. the dual norm.

Lemma 7 (Convexity/Smoothness Duality). *Assume that f is a closed and convex function. Then f is β-strongly convex w.r.t. a norm $\|\cdot\|$ if and only if f^* is $\frac{1}{\beta}$-strongly smooth w.r.t. the dual norm $\|\cdot\|_*$.*

Theorem 6, coupled with the convexity/smoothness duality, allows us to derive the following strongly convex functions for different instantiations of the group norm $\|\cdot\|_{\Psi,\Phi}$.

Corollary 1. *Let $1 < r_1, r_2, p \le 2$ be three positive numbers. Then the function $F(M) = \frac{1}{2}\|M\|^2_{(r_1,r_2),p}$ is $(s_1+s_2+q-3)^{-1}$-strongly convex w.r.t. $\|\cdot\|_{(r_1,r_2),p}$, where s_i is the dual exponent of $r_i, i = 1,2$ and q is the dual exponent of p.*

Proof. Introduce the norm $\Psi = \|\cdot\|_{s_1,s_2}$ on $\mathbb{R}^{d\times d}$ and the norm $\Phi = \|\cdot\|_q$ on $\mathbb{R}^m$. Kakade et al. [8] indicated that Ψ^2 is $2(s_1+s_2-2)$-strongly smooth w.r.t. Ψ and Φ^2 is $2(q-1)$-strongly smooth w.r.t. Φ. Furthermore, the fact $q \ge 2$ guarantees that $\Phi^2 \circ \surd = \|\cdot\|^2_q \circ \surd = \|\cdot\|_{q/2}$ is a norm on $\mathbb{R}^m$. Also, both Ψ and Φ are absolutely symmetric. Thus, an application of Theorem 6 shows that $\frac{1}{2}\|\cdot\|^2_{\Psi,\Phi}$ is (s_1+s_2+q-3)-strongly smooth w.r.t. $\|\cdot\|_{\Psi,\Phi}$. Lemma 17 in [8] implies that $\|\cdot\|_{\Psi,\Phi}$ is a norm on $\mathbb{R}^{d\times(md)}$ with dual norm $\|\cdot\|_{\Psi_*,\Phi_*} = \|\cdot\|_{(r_1,r_2),p}$.

For any norm $\|\cdot\|$, we know that the Fenchel conjugate of $\frac{1}{2}\|\cdot\|^2$ is $\frac{1}{2}\|\cdot\|^2_*$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ [3]. Consequently, the Fenchel conjugate of the function $\frac{1}{2}\|M\|^2_{\Psi,\Phi}$ is $F(M) = \frac{1}{2}\|M\|^2_{\Psi_*,\Phi_*}$. Putting the above discussions together and noticing the convexity/smoothness duality (Lemma 7), we immediately derive that $F(M)$ is $(s_1+s_1+q-3)^{-1}$-strongly convex w.r.t. the norm $\|\cdot\|_{(r_1,r_2),p}$. □

Corollary 2. *Let $1 < r, p \le 2$ be any two numbers. Then the function $F(M) = \frac{1}{2}\|M\|^2_{S(r),p}$ is $\min((q+1)^{-1}, (s+q-2)^{-1})$-strongly convex w.r.t. $\|\cdot\|_{S(r),p}$, where q and s are dual exponents of p and r, respectively.*

Proof. It is known [8] that $\|\cdot\|^2_{S(s)}$ is $2\max\{2, s-1\}$-strongly smooth w.r.t. the norm $\|\cdot\|_{S(s)}$. From the proof of Corollary 1, one can see that the conditions of Theorem 6 hold for the choice $\Psi = \|\cdot\|_{S(s)}$ on $\mathbb{R}^{d\times d}$ and $\Phi = \|\cdot\|_q$ on $\mathbb{R}^m$. Consequently, the function $\frac{1}{2}\|\cdot\|^2_{\Psi,\Phi}$ is $\max(q+1, s+q-2)$-strongly smooth w.r.t. $\|\cdot\|_{\Psi,\Phi}$. Now Lemma 7 shows that $F(M) = \frac{1}{2}\|M\|^2_{S(r),p} = \left(\frac{1}{2}\|M\|^2_{S(s),q}\right)^*$ is $\min((q+1)^{-1}, (s+q-2)^{-1})$-strongly convex w.r.t. $\|\cdot\|_{S(r),p}$. □

4.2 Specific examples

Here we estimate Rademacher complexities for different matrix regularization schemes. Our discussion covers the regularizer $\|\cdot\|_{S(1),1}$ used in [12, 15] and the regularizer $\|\cdot\|_{(2,2),2}$ exploited in [14]. For brevity, we first introduce the following notations for any $s, q \ge 1$:

$$X_{(s,s),q} := \sup_{x,x' \in \mathcal{X}} \|\big((x^1-(x')^1)(x^1-(x')^1)^\top, \dots, (x^m-(x')^m)(x^m-(x')^m)^\top\big)\|_{(s,s),q},$$

$$X_{S(s),q} := \sup_{x,x' \in \mathcal{X}} \|\big((x^1-(x')^1)(x^1-(x')^1)^\top, \dots, (x^m-(x')^m)(x^m-(x')^m)^\top\big)\|_{S(s),q},$$

$$X_{s,q} := \sup_{x,x' \in \mathcal{X}} \|(x^1-(x')^1, x^2-(x')^2, \dots, x^m-(x')^m)\|_{s,q}.$$

The following two lemmas show that the discussion of $X_{(s,s),q}$ and $X_{S(s),q}$ all reduce to the estimation of $X_{s,q}$.

Lemma 8. *Let $s, q \geq 1$ be two arbitrary numbers. For any $x^1, \ldots, x^m \in \mathbb{R}^d$, we have*

$$\|(x^1(x^1)^\top, x^2(x^2)^\top, \ldots, x^m(x^m)^\top)\|_{(s,s),q} = \|(x^1, x^2, \ldots, x^m)\|_{s,2q}^2.$$

Proof. Denote by $x^{l,i}$ the i-th component of the vector x^l. From the definition of group norm, we know that

$$\begin{aligned}\|(x^1(x^1)^\top, x^2(x^2)^\top, \ldots, x^m(x^m)^\top)\|_{(s,s),q} &= \left[\sum_{l=1}^m \Big(\sum_{i=1}^d \sum_{j=1}^d |x^{l,i}|^s |x^{l,j}|^s\Big)^{q/s}\right]^{1/q} \\ &= \left[\sum_{l=1}^m \Big(\sum_{i=1}^d |x^{l,i}|^s\Big)^{q/s} \Big(\sum_{j=1}^d |x^{l,j}|^s\Big)^{q/s}\right]^{1/q} \\ &= \left[\sum_{l=1}^m \|x^l\|_s^{2q}\right]^{1/q} \\ &= \|(x^1, x^2, \ldots, x^m)\|_{s,2q}^2.\end{aligned}$$

This completes the proof. □

Lemma 9. *Let $s, q \geq 1$ be two arbitrary numbers. For any $x^1, \ldots, x^m \in \mathbb{R}^d$, we have*

$$\|(x^1(x^1)^\top, x^2(x^2)^\top, \ldots, x^m(x^m)^\top)\|_{S(s),q} = \|(x^1, x^2, \ldots, x^m)\|_{2,2q}^2.$$

Proof. Denote by $x^{l,i}$ the i-th component of the vector x^l. For any $l \in \mathbb{N}_m$, note that the matrix $x^l(x^l)^\top$ is of rank one. Consequently, the matrix $x^l(x^l)^\top$ has only one non-zero eigenvalue, which is identical to its trace $\sum_{i=1}^d (x^{l,i})^2 = \|x^l\|_2^2$. That is,

$$\|x^l(x^l)^\top\|_{S(s)} = \|x^l\|_2^2, \qquad \forall s \geq 1.$$

Now, the group norm has the following equivalent form:

$$\begin{aligned}\|\big(x^1(x^1)^\top, x^2(x^2)^\top, \ldots, x^m(x^m)^\top\big)\|_{S(s),q} &= \left[\sum_{l=1}^m \|x^l(x^l)^\top\|_{S(s)}^q\right]^{1/q} \\ &= \left[\sum_{l=1}^m \|x^l\|_2^{2q}\right]^{1/q} \\ &= \|\big(x^1, x^2, \ldots, x^m\big)\|_{2,2q}^2.\end{aligned}$$

□

In the remainder of this section, we always assume that $(p, q), (r, s)$ are two dual pairs.

Example 1. *Consider the class*

$$\mathcal{M}_{S(1),1} = \{M \in \mathbb{R}^{d \times (md)} : \|M\|_{S(s),1} \leq M_{S(1),1}\}.$$

Then, the Rademacher complexity satisfies the inequality

$$R_n(\mathcal{M}_{S(1),1}) \leq e^2 X_{2,\infty}^2 M_{S(1),1} \sqrt{\frac{\log m + \log d}{\lfloor \frac{n}{2} \rfloor}}.$$

Proof. For any $r, p \in (1,2]$, Corollary 2 indicates that the function $F(M) := \frac{1}{2}\|M\|_{S(r),p}^2$ is $\min((q+1)^{-1}, (s+q-2)^{-1})$-strongly convex w.r.t. $\|\cdot\|_{S(r),p}$. Note that any $M \in \mathcal{M}_{S(1),1}$ meets the inequality $F(M) \leq \frac{1}{2}\|M\|_{S(1),1}^2 \leq \frac{1}{2}M_{S(1),1}^2$. Moreover, for any $W = (W_1, \ldots, W_m) \in \mathbb{R}^{d\times(md)}$, we have the following inequality connecting $\|W\|_{S(s),q}$ with $\|W\|_{S(\infty),\infty}$:

$$\begin{aligned}\|W\|_{S(s),q} &\leq m^{1/q} \sup_{1\leq l\leq m} \|W^l\|_{S(s)} \leq m^{1/q} d^{1/s} \sup_{1\leq l\leq m} \|W^l\|_{S(\infty)} \\ &= m^{1/q} d^{1/s} \|W\|_{S(\infty),\infty}.\end{aligned}$$

Thus, there holds that $X_{S(s),q} \leq m^{1/q} d^{1/s} X_{S(\infty),\infty}$. Applying Theorem 5 with $\|\cdot\| = \|\cdot\|_{S(r),p}$ and Lemma 9 here, we have

$$\begin{aligned}R_n(\mathcal{M}_{S(1),1}) &\leq X_{S(s),q}\sqrt{\frac{2\sup_{M\in\mathcal{M}_{S(1),1}} F(M)}{\min((q+1)^{-1}, (s+q-2)^{-1})\lfloor \frac{n}{2}\rfloor}} \\ &\leq m^{1/q} d^{1/s} X_{S(\infty),\infty} M_{S(1),1} \sqrt{\frac{\max(q+1, s+q-2)}{\lfloor \frac{n}{2}\rfloor}} \\ &= m^{1/q} d^{1/s} X_{2,\infty}^2 M_{S(1),1} \sqrt{\frac{\max(q+1, s+q-2)}{\lfloor \frac{n}{2}\rfloor}}.\end{aligned}$$

As the above inequality holds for any $q, s \geq 2$, plugging $q = \log m$ and $s = \log d$ into the above inequality completes the proof [1]. □

Example 2. *For the class*

$$\mathcal{M}_{S(r),1} = \{M \in \mathbb{R}^{d\times(md)} : \|M\|_{S(r),1} \leq M_{S(r),1}\}, \qquad r \in (1,2],$$

we have the following Rademacher complexity bound:

$$R_n(\mathcal{M}_{S(r),1}) \leq eX_{2,\infty}^2 M_{S(r),1} \sqrt{\frac{\log m + \max(1, s-2)}{\lfloor \frac{n}{2}\rfloor}}.$$

Proof. Introduce the function $F(M) := \frac{1}{2}\|M\|_{S(r),p}^2, 1 < p \leq 2$. Corollary 2 implies that $F(M)$ is $\min((q+1)^{-1}, (s+q-2)^{-1})$-strongly convex w.r.t. $\|\cdot\|_{S(r),p}$, while its magnitude over $\mathcal{M}_{S(r),1}$ can be controlled by

$$F(M) = \frac{1}{2}\|M\|_{S(r),p}^2 \leq \frac{1}{2}\|M\|_{S(r),1}^2 \leq \frac{1}{2}M_{S(r),1}^2, \qquad \forall M \in \mathcal{M}_{S(r),1}.$$

Moreover, we have $X_{S(s),q} \leq m^{1/q} X_{S(s),\infty}$. Putting the above discussion together and applying here Theorem 5 with $\|\cdot\| = \|\cdot\|_{S(r),p}$, we derive the inequality

$$\begin{aligned}R_n(\mathcal{M}_{S(r),1}) &\leq X_{S(s),q}\sqrt{\frac{2\sup_{M\in\mathcal{M}_{S(r),1}} F(M)}{\min((q+1)^{-1}, (s+q-2)^{-1})\lfloor \frac{n}{2}\rfloor}} \\ &\leq X_{S(s),\infty} M_{S(r),1} m^{1/q} \sqrt{\frac{\max(q+1, s+q-2)}{\lfloor \frac{n}{2}\rfloor}}.\end{aligned}$$

Taking the choice $q = \log m$ and using Lemma 9 to control $X_{S(s),\infty}$, we obtain the promised inequality. □

Example 3. *The Rademacher complexity of*

$$\mathcal{M}_{(1,1),1} = \{M \in \mathbb{R}^{d\times(md)} : \|M\|_{(1,1),1} \leq M_{(1,1),1}\}$$

satisfies the inequality

$$R_n(\mathcal{M}_{(1,1),1}) \leq eX_{\infty,\infty}^2 M_{(1,1),1} \sqrt{\frac{\log(md^2) - 1}{\lfloor \frac{n}{2}\rfloor}}.$$

[1] If either $m < e^2$ or $d < e^2$, one can simply take $q = 2$ or $s = 2$ to get similar Rademacher complexity bounds. For simplicity, here and in the following examples we omit the discussions for these situations.

Proof. Consider the function $F(M) := \frac{1}{2}\|M\|^2_{(p,p),p}, p \in (1,2]$. It is known [9] that $F(M)$ is $(p-1)$-strongly convex w.r.t. $\|\cdot\|_{(p,p),p}$. Furthermore, we have

$$X_{(q,q),q} \le (md^2)^{1/q} X_{(\infty,\infty),\infty} = (md^2)^{1/q} X^2_{\infty,\infty}.$$

Consequently, there holds that

$$R_n(\mathcal{M}_{(1,1),1}) \le (md^2)^{1/q} X^2_{\infty,\infty} M_{(1,1),1} \sqrt{\frac{q-1}{\lfloor \frac{n}{2} \rfloor}}.$$

Plugging $q = \log(md^2)$ into the above inequality yields the desired bound. □

Example 4. *Suppose that the class $\mathcal{M}_{(r,r),1}$ is defined by*

$$\mathcal{M}_{(r,r),1} = \{M \in \mathbb{R}^{d\times(md)} : \|M\|_{(r,r),1} \le M_{(r,r),1}\}, \qquad r \in (1,2].$$

Then the Rademacher complexity satisfies the inequality

$$R_n(\mathcal{M}_{(r,r),1}) \le eX^2_{s,\infty} M_{(r,r),1} \sqrt{\frac{s-2+\log m}{\lfloor \frac{n}{2} \rfloor}}.$$

Proof. Corollary 1 indicates that for any $p \in (1,2]$, the function $F(M) := \frac{1}{2}\|M\|^2_{(r,r),p}$ is $(s+q-2)^{-1}$-strongly convex w.r.t. $\|\cdot\|_{(r,r),p}{}^2$. For any $M \in \mathcal{M}_{(r,r),1}$, there holds that

$$F(M) = \frac{1}{2}\|M\|^2_{(r,r),p} \le \frac{1}{2}\|M\|^2_{(r,r),1} \le \frac{1}{2}M^2_{(r,r),1}.$$

Moreover, it can be directly verified that $X_{(s,s),q} \le m^{1/q} X_{(s,s),\infty}$. Consequently, applying Theorem 5 with $\|\cdot\| = \|\cdot\|_{(r,r),p}$ implies that

$$R_n(\mathcal{M}_{(r,r),1}) \le X_{(s,s),q} \sqrt{\frac{2\sup_{M\in\mathcal{M}_{(r,r),1}} F(M)}{(s+q-2)^{-1}\lfloor \frac{n}{2} \rfloor}} \le m^{1/q} X_{(s,s),\infty} M_{(r,r),1} \sqrt{\frac{s+q-2}{\lfloor \frac{n}{2} \rfloor}}.$$

Taking the assignment $q = \log m$ in the above inequality, we have

$$R_n(\mathcal{M}_{(r,r),1}) \le eX_{(s,s),\infty} M_{(r,r),1} \sqrt{\frac{s-2+\log m}{\lfloor \frac{n}{2} \rfloor}}.$$

The proof is complete if we apply Lemma 8 here to bound $X_{(s,s),\infty}$. □

For simplicity, we now provide Rademacher complexity bounds for $\mathcal{M}_{S(r),p}$ and $\mathcal{M}_{(2,2),2}$ without presenting the proof.

Example 5. *The Rademacher complexity of*

$$\mathcal{M}_{S(r),p} = \{M \in \mathbb{R}^{d\times(md)} : \|M\|_{S(r),p} \le M_{S(r),p}\}$$

is upper bounded by

$$R_n(\mathcal{M}_{S(r),p}) \le X^2_{2,2q} M_{S(r),p} \sqrt{\frac{\max(q+1, s+q-2)}{\lfloor \frac{n}{2} \rfloor}}.$$

Example 6. *The Rademacher complexity of*

$$\mathcal{M}_{(2,2),2} = \{M \in \mathbb{R}^{d\times(md)} : \|M\|_{(2,2),2} \le M_{(2,2),2}\}$$

satisfies the inequality

$$R_n(\mathcal{M}_{(2,2),2}) \le X^2_{2,4} M_{(2,2),2} \sqrt{\frac{1}{n}}.$$

class	$\mathcal{M}_{(1,1),1}$	$\mathcal{M}_{(2,2),2}$	$\mathcal{M}_{(r,r),1}, r \in (1,2]$
bound	$X^2_{\infty,\infty} M_{(1,1),1} \sqrt{\frac{\log(md^2)}{n}}$	$X^2_{2,4} M_{(2,2),2} \sqrt{\frac{1}{n}}$	$X^2_{s,\infty} M_{(r,r),1} \sqrt{\frac{s+\log m}{n}}$

Table 1: Rademacher complexity bounds for mixed (q, s)-norm based regularizers.

class	$\mathcal{M}_{S(r),1}, r \in (1,2]$	$\mathcal{M}_{S(1),1}$	$\mathcal{M}_{S(r),p}$
bound	$X^2_{2,\infty} M_{S(r),1} \sqrt{\frac{\log m+s}{n}}$	$X^2_{2,\infty} M_{S(1),1} \sqrt{\frac{\log m+\log d}{n}}$	$X^2_{2,2q} M_{S(r),p} \sqrt{\frac{s+q}{n}}$

Table 2: Rademacher complexity bounds for Schatten-norm based regularizers.

Combining the above Rademacher complexity bounds and Theorem 3, it is straightforward to derive the generalization bounds for problem (2) under different regularizers. For brevity, we omit these discussions here. We now summarize in Table 1 and 2 the Rademacher complexity bounds in Examples 1-6, ignoring the exact constant factors here.

Remark 2. Our discussion extends the generalization analysis in [4] from single-modal context to multi-modal case. Furthermore, our result is more general in that we provide a universal technique to approach Rademacher complexities, while the complexity bounds in [4] are somewhat disperse and the deduction process largely relies on the specific classes considered.

Remark 3. In the context of single-modal learning, Cao et al. [4] derived the Rademacher complexity bounds of the form $\sup_{x,x'} \|x-x'\|_2^2 \cdot n^{-1/2}$ for the regularizer $\|\cdot\|_{2,2}$, and of the form $\sup_{x,x'} \|x-x'\|_\infty^2 \sqrt{\frac{\log d}{n}}$ for the regularizer $\|\cdot\|_{1,1}$. It can be clearly seen that our results recover these bounds when $m = 1$. Furthermore, we also provide Rademacher complexity bounds for Schatten-norm based regularizers, which were not addressed in [4].

5 Discussions

In this section, we discuss the implication of our generalization analysis for multi-modal metric learning and its extension to similarity learning.

Multi-modal similarity learning. Although we only focus our attention to metric learning here, it should be mentioned that these discussions can be transformed in a straightforward manner to the setting of multi-modal similarity learning. Indeed, one can define the following empirical error for similarity learning:

$$\widetilde{\mathcal{E}_{\boldsymbol{z}}}(M,b) := \frac{1}{n(n-1)} \sum_{i,j \in \mathbb{N}_n, i \neq j} [1 - r(y_i, y_j)(s_M(x_i, x_j) - b)]_+,$$

where the similarity function $s_M(x_i, x_j)$ is

$$s_M(x_i, x_j) := \sum_{l=1}^m (x_i^l)^\top M^l x_j^l.$$

One can also define the generalization error $\widetilde{\mathcal{E}}(M,b)$ in a similar manner. Suppose that the model $(\widetilde{M_{\boldsymbol{z}}}, \widetilde{b}_{\boldsymbol{z}})$ is obtained by minimizing the regularization functional (2) except that the involved distance function $d_M(x_i, x_j)$ is replaced by the similarity function $s_M(x_i, x_j)$. For any matrix class $\mathcal{M}$, one can define the following Rademacher complexity for multi-modal similarity learning:

$$\widetilde{R}_n(\mathcal{M}) := \frac{1}{\lfloor \frac{n}{2} \rfloor} \mathbb{E} \sup_{M \in \mathcal{M}} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i s_M(x_i, x_{\lfloor \frac{n}{2} \rfloor + i}).$$

[2] In the case $r_1 = r_2 = r$, the constant $(s_1 + s_2 + q - 3)^{-1}$ in Corollary 1 can be slightly improved to $(s + q - 2)^{-1}$, which is due to the fact that $\frac{1}{2}\|\cdot\|^2_{(s,s)}$ is $(s-1)$-strongly smooth w.r.t. $\|\cdot\|_{(s,s)}$.

Analogous to Theorem 3, with probability $1-\delta$, the generalization error of $(\widetilde{M}_{\boldsymbol{z}}, \widetilde{b}_{\boldsymbol{z}})$ satisfies the inequality

$$\widetilde{\mathcal{E}}(\widetilde{M}_{\boldsymbol{z}}, \widetilde{b}_{\boldsymbol{z}}) - \widetilde{\mathcal{E}}_{\boldsymbol{z}}(\widetilde{M}_{\boldsymbol{z}}, \widetilde{b}_{\boldsymbol{z}}) \leq 2\widetilde{R}_n(\{M \in \mathbb{S}^{d\times(md)} : \|M\| \leq \lambda^{-1/2}\}) + 2(1+\widetilde{X}_*/\sqrt{\lambda})\frac{1+2\sqrt{\log\frac{1}{\delta}}}{\lfloor\frac{n}{2}\rfloor},$$

where

$$\widetilde{X}_* := \sup_{x,x'\in\mathcal{X}} \big\|\big((x')^1(x^1)^\top, (x')^2(x^2)^\top, \ldots, (x')^m(x^m)^\top\big)\big\|_*.$$

Furthermore, for any class $\mathcal{M}$ meeting the conditions of Theorem 5, one can derive the Rademacher complexity bound: $\widetilde{R}_n(\mathcal{M}) \leq \widetilde{X}_*\sqrt{\frac{2f_{\max}}{\beta\lfloor\frac{n}{2}\rfloor}}$. There also holds similar Rademacher complexity bounds for the specific classes considered in Section 4.2. For example, for the class $\mathcal{M}_{S(1),1}$ we have

$$\widetilde{R}_n(\mathcal{M}_{S(1),1}) \leq e^2\widetilde{X}^2_{2,\infty}M_{S(1),1}\sqrt{\frac{\log m + \log d}{\lfloor\frac{n}{2}\rfloor}},$$

where $\widetilde{X}_{2,\infty} := \sup_{x\in\mathcal{X}} \|(x^1, x^2, \ldots, x^m)\|_{2,\infty}$.

Learning under weighted distance metric. In some multi-modal learning problems, it may happen that some modalities are more important than the others in doing the prediction. In such cases, it would be a better strategy to consider a weighted summation when incorporating the sub-distances associated to each modality together. That is, the distance metric is of the form:

$$d_{M,\mu}(x_i, x_j) := \sum_{l=1}^m \mu_l d_{M^l}(x_i^l, x_j^l) = \sum_{l=1}^m \mu_l (x_i^l - x_j^l)^\top M^l (x_i^l - x_j^l), \tag{8}$$

where $\mu_l > 0$ represents our belief on the importance of the l-th modality. Such strategy was considered by Xia et al. [15]. Although this distance function largely enriches the flexibility of multi-modal metric learning, our previous discussion for the unweighted case can be easily extended to tackle problem (8). Specifically, for each $x \in \mathcal{X}$, we introduce a related vector

$$\bar{x} = \big(\sqrt{\mu_1}(x^1)^\top, \sqrt{\mu_2}(x^2)^\top, \ldots, \sqrt{\mu_m}(x^m)^\top\big)^\top.$$

Now it is obvious that

$$d_{M,\mu}(x, x') = d_{M,\mathbf{1}}(\bar{x}, \bar{x}'), \qquad \forall x, x' \in \mathcal{X},$$

where $\mathbf{1}$ is the vector with all components equal to 1. Consequently, to study the generalization performance of $(M_{\boldsymbol{z}}, b_{\boldsymbol{z}})$ under the distance function $d_{M,\mu}$, it suffices to consider the unweighted distance metric $d_{M,\mathbf{1}}$ with the transformed input data $\bar{x}$.

Comparison of bounds with different regularizers. We now compare the generalization bounds for multi-modal metric learning under different regularization terms. To this end, assume the best model is $(M^*, b^*) := \arg\min_{M,b} \mathcal{E}(M, b)$ and the classes $\mathcal{M}_{\Psi,\Phi}$ in Examples 1-6 are of the form $\mathcal{M}_{\Psi,\Phi} = \{M \in \mathbb{R}^{d\times(md)} : \|M\|_{\Psi,\Phi} \leq \|M^*\|_{\Psi,\Phi}\}$ with different instantiations of Ψ, Φ.

Our first remark is that the Frobenius-norm based regularizers (regularizers with Frobenius-norm on each modality) are preferable to the Schatten-norm based regularizers (regularizers with Schatten-norm on each modality). Indeed, for any $r \in (1, 2]$, the Rademacher complexity bounds for $\mathcal{M}_{S(r),1}$ and $\mathcal{M}_{(2,2),1}$ satisfy the inequality

$$\begin{aligned} X^2_{2,\infty}\|M^*\|_{S(r),1}\sqrt{\frac{\log m + s}{n}} &\geq X^2_{2,\infty}\|M^*\|_{S(2),1}\sqrt{\frac{\log m + 2}{n}} \qquad \text{(since } r \leq 2\text{)} \\ &= X^2_{2,\infty}\|M^*\|_{(2,2),1}\sqrt{\frac{\log m + 2}{n}}, \end{aligned}$$

where we have used the fact that $\|\cdot\|_{S(2)} = \|\cdot\|_{2,2}$. That is, our Rademacher complexity bounds for $\mathcal{M}_{(2,2),1}$ are always smaller than that for $\mathcal{M}_{S(r),1}, r \in (1, 2]$. The same argument also holds for the classes $\mathcal{M}_{S(r),p}$

and $\mathcal{M}_{(2,2),p}, p \geq 1, r \in (1,2]$. The underlying reason for this phenomenon is that, for any rank-one matrix, its Schatten norm is reduced to the Frobenius norm.

We now consider the choice among $\mathcal{M}_{(1,1),1}, \mathcal{M}_{(2,2),2}$ and $\mathcal{M}_{(2,2),1}$. Both $\|M^*\|_{(1,1),1}$ and $\|M^*\|_{(2,2),2}$ ignore the group structure of M^* and boil down to the ℓ_1- and ℓ_2-norm on M^*, respectively. While $\|M^*\|_{(2,2),1}$ behaves like an ℓ_1-norm on the vector $(\|M^l\|_{(2,2)})_{l\in\mathbb{N}_m}$ and thus assumes a structural sparsity. This has the effect of "variable section" in the sense that some unrelated modalities are removed. Unlike the case for Schatten-norm based regularizers, as we will see, none of these regularizes dominates the rest and which one should be used depends on a priori information of the concrete problem. Specifically, on the one hand the norm on M^* satisfies the inequality

$$\|M^*\|_{(2,2),2} \leq \|M^*\|_{(2,2),1} \leq \|M^*\|_{(1,1),1}.$$

On the other hand, we have the following relationship for X:

$$X_{2,4} \geq X_{2,\infty} \geq X_{\infty,\infty}.$$

Thus, considering either $\|M^*\|_{\Psi,\Phi}$ or X_{Ψ_*,Φ_*} while ignoring the other would lead to a reverse preference on the choice of regularizers. Only if we know how $\|M^*\|_{\Psi,\Phi}$ grows versus X_{Ψ_*,Φ_*} can we make a right selection. For example, if we know that M^* is sparse and the input vector x is dense, then the regularizer $\|\cdot\|_{(1,1),1}$ would be the best choice. On the other hand, if M^* admits a structural sparsity (i.e., the vector $(\|(M^*)^1\|_{2,2}, \ldots, \|(M^*)^m\|_{2,2})$ is sparse) and the input vector x admits a structural density (i.e., the vector $(\|x^1\|, \ldots, \|x^m\|)$ is dense), then both $\|M^*\|_{(2,2),1}$ and $X_{2,\infty}$ would be small and thus $\|\cdot\|_{(2,2),1}$ may be an appropriate regularizer.

6 Conclusions

Motivated by the growing interest on multi-modal metric learning recently, we initiate the theoretical work on studying their generalization performance in this paper. We establish its generalization bounds using the concept of Rademacher complexity. By restating the distance metric (1) as the inner product of two matrices, we also provide a general result on estimating Rademacher complexities. We demonstrate the potential of this general Rademacher complexity bound by deriving novel generalization bounds for different regularizers (e.g., $\|\cdot\|_{S(r),p}, \|\cdot\|_{(r,r),p}$) under the unified framework. Below we mention some interesting problems for future research.

Firstly, in many applications the metric built on the metric learning stage serves as the foundation on which other machine learning algorithms (e.g., classification, regression) are subsequently implemented. Thus, it would be very interesting to establish the theoretical link between the generalization bounds of metric learning and the generalization performance of the resulting classifiers/regressors. Guo and Ying [5] developed such a connection by showing that regularized similarity learning can guarantee the goodness of the resulting classifier in the single-modal setting. In future, we would like to investigate how their discussion can be extended to the multi-modal case.

Secondly, we only consider linear metric learning in this paper. It remains a challenging question to study the generalization bounds for nonlinear multi-modal metric learning, where each modality is encoded by a kernel function and the Mahalanobis metric is searched in the feature space [12, 13]. A probable starting point would be the techniques developed for the kernel learning problems [11, 18, 19].

Acknowledgements

The work was done when Yunwen was visiting the University of Exeter, supported by the PhD short-time mobility scholarship from Wuhan University, 2013. The corresponding author is Yiming Ying.

References

[1] P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Stat.*, 33(4): 1497–1537, 2005.

[2] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[3] S. P. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge Univ. Press, New York, 2004.

[4] Q. Cao, Z.-C. Guo, and Y. Ying. Generalization bounds for metric and similarity learning. *arXiv preprint:1207.5437*, 2012.

[5] Z.-C. Guo and Y. Ying. Guaranteed classification via regularized similarity learning. *Neural. Comput.*, pages 497–522, 2014.

[6] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: theory and algorithm. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in 22nd Neural Information Processing Systems*, NIPS '09, pages 862–870, Vancouver, British Columbia, Canada, 2009.

[7] S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in the 21st Neural Information Processing Systems*, NIPS '08, pages 793–800, Vancouver, British Columbia, Canada, 2008.

[8] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, available at http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2009.

[9] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Regularization techniques for learning with matrices. *J. Mach. Learn. Res.*, 13:1865–1890, 2012.

[10] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, Berlin, 1991.

[11] Y. Lei and L. Ding. Refined Rademacher chaos complexity bounds with applications to the multikernel learning problem. *Neural. Comput.*, 26(4):1–22, 2014.

[12] B. McFee and G. Lanckriet. Learning multi-modal similarity. *J. Mach. Learn. Res.*, 12:491–523, 2011.

[13] J. Wang, H. Do, A. Woznica, and A. Kalousis. Metric learning with multiple kernels. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in the 24th Neural Information Processing Systems*, NIPS '11, pages 1170–1178, Granada, Spain, 2011.

[14] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao. Online multimodal deep similarity learning with application to image retrieval. In A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D. A. Shamma, M. Worring, and R. Zimmermann, editors, *Proceedings of the 21st ACM Multimedia Conference*, MM '13, pages 153–162, Barcelona, Spain, 2013. ACM.

[15] H. Xia, P. Wu, and S. C. Hoi. Online multi-modal distance learning for scalable multimedia retrieval. In S. Leonardi, A. Panconesi, P. Ferragina, and A. Gionis, editors, *Proceedings of the 6th ACM international conference on Web search and data mining*, WSDM '13, pages 455–464, Rome, Italy, 2013. ACM.

[16] P. Xie and E. P. Xing. Multi-modal distance metric learning. In F. Rossi, editor, *Proceedings of the 23rd international joint conference on Artificial Intelligence*, IJCAI '13, pages 1806–1812, Beijing, China, 2013. AAAI Press.

[17] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2, 2006.

[18] Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In S. Dasgupta and A. Klivans, editors, *Proceedings of the 22nd Annual Conference on Learning Theory*, COLT '09, Montreal, Quebec, Canada, 2009.

[19] Y. Ying and C. Campbell. Rademacher chaos complexities for learning the kernel problem. *Neural. Comput.*, 22(11):2858–2886, 2010.

[20] Y. Ying, K. Huang, and C. Campbell. Sparse metric learning via smooth optimization. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in 22nd Neural Information Processing Systems*, NIPS '09, pages 2214–2222, Vancouver, British Columbia, Canada, 2009.