Guaranteed Classification via Regularized Similarity Learning

Zheng-Chu Guo\(^1\) and Yiming Ying

College of Engineering, Mathematics and Physical Sciences
University of Exeter, EX4 4QF, UK.
{gzhengchu, mathying}@gmail.com

Abstract

Learning an appropriate (dis)similarity function from the available data is a central problem in machine learning, since the success of many machine learning algorithms critically depends on the choice of a similarity function to compare examples. Despite many approaches for similarity metric learning have been proposed, there is little theoretical study on the links between similarity metric learning and the classification performance of the result classifier. In this paper, we propose a regularized similarity learning formulation associated with general matrix-norms, and establish their generalization bounds. We show that the generalization error of the resulting linear classifier can be bounded by the derived generalization bound of similarity learning. This shows that a good generalization

\(^1\)The author’s current affiliation: Department of Mathematics, Zhejiang University, Hangzhou 310027, China
of the learnt similarity function guarantees a good classification of the resulting linear classifier. Our results extend and improve those obtained by Bellet at al. (Bellet et al., 2012). Due to the techniques dependent on the notion of uniform stability (Bousquet & Elisseeff, 2002), the bound obtained there holds true only for the Frobenius matrix-norm regularization. Our techniques using the Rademacher complexity (Bartlett & Mendelson, 2002) and its related Khinchin-type inequality enable us to establish bounds for regularized similarity learning formulations associated with general matrix-norms including sparse $L^1$-norm and mixed $(2, 1)$-norm.

1 Introduction

The success of many machine learning algorithms heavily depends on how to specify the similarity or distance metric between examples. For instance, the k-nearest neighbor (k-NN) classifier depends on a distance (dissimilarity) function to identify the nearest neighbors for classification. Most information retrieval methods rely on a similarity function to identify the data points that are most similar to a given query. Kernel methods rely on the kernel function to represent the similarity between examples. Hence, how to learn an appropriate (dis)similarity function from the available data is a central problem in machine learning, which we refer to as similarity metric learning throughout the paper.

Recently, a considerable amount of research efforts are devoted to similarity metric learning and many methods have been proposed. They can be broadly divided into two main categories. The first category of such methods is one-stage approach for
similarity metric learning, which means that the methods learn the similarity (kernel) function and classifier together. Multiple kernel learning (Lanckriet et al., 2004; Varma & Babu., 2009) is a notable one-stage approach, which aims to learn an optimal kernel combination from a prescribed set of positive semi-definite (PSD) kernels. Another exemplary one-stage approach is indefinite kernel learning, which is motivated by the fact that, in many applications, potential kernel matrices could be non-positive semi-definite. Such cases include hyperbolic tangent kernels (Smola et al., 2001), and the protein sequence similarity measures derived from Smith-Waterman and BLAST score (Saigo et al., 2004). Indefinite kernel learning (Chen et al., 2009; Ying et al., 2009) aims to learn a PSD kernel matrix from a prescribed indefinite kernel matrix, which are mostly restricted to the transductive settings. Recent methods (Wu & Zhou, 2005; Wu, 2013) analyzed regularization networks such as ridge regression and SVM given a prescribed indefinite kernel, instead of aiming to learn an indefinite kernel function from data. The generalization analysis for such one-stage methods was well studied (see, e.g., Chen et al., 2009; Cortes et al., 2010a; Ying & Campbell, 2009).

The second category of similarity metric learning is a two-stage method, which means that the processes of learning the similarity function and training the classifier are separate. One exemplar two-stage approach is referred to as metric learning (Bar-Hillel et al., 2005; Davis et al., 2007; Hoi et al., 2006; Jin et al., 2009; Blitzer et al., 2005; Xing et al., 2002; Ying et al., 2009), which often focuses on learning a Mahalanobis distance metric defined, for any \( x, x' \in \mathbb{R}^d \), by \( d_M(x, x') = \sqrt{(x - x')^T M (x - x')} \). Here, \( M \) is a positive semi-definite (PSD) matrix. Another example of such methods (Chechik et al., 2010; Maurer, 2008) is bilinear similarity learning, which focuses on learning
a similarity function defined, for any \( x, x' \in \mathbb{R}^d \), by \( s_M(x, x') = x^T M x' \) with \( M \) being a PSD matrix. The above methods are mainly motivated by the natural intuition that the similarity score between examples in the same class should be larger than that of examples from distinct classes. The k-NN classification using the similarity metric learnt from the above methods was empirically shown to achieve better accuracy than that using the standard Euclidean distance.

Although many two-stage approaches for similarity metric learning have been proposed, in contrast to the one-stage methods, there is relatively little theoretical study on the question whether similarity-based learning guarantees a good generalization of the resultant classification. For instance, generalization bounds were recently established for metric and similarity learning (Cao et al., 2012; Jin et al., 2009; Maurer, 2008) under different statistical assumptions on the data. However, there are no theoretical guarantees for such empirical success. In other words, it is not clear whether good generalization bounds for metric and similarity learning (Cao et al., 2012; Jin et al., 2009) can lead to a good classification performance of the resultant k-NN classifiers. Recently, Bellet et al. (Bellet et al., 2012) proposed a regularized similarity learning approach, which is mainly motivated by the \((\varepsilon, \gamma, \tau)\)-good similarity functions introduced in (Balcan & Blum, 2006; Balcan et al., 2008). In particular, they showed that the proposed similarity learning can theoretically guarantee a good generalization for classification. However, due to the techniques dependent on the notion of uniform stability (Bousquet & Elisseeff, 2002), the generalization bounds only hold true for strongly convex matrix-norm regularization (e.g. the Frobenius norm).

In this paper, we consider a new similarity learning formulation associated with general
matrix-norm regularization terms. Its generalization bounds are established for various matrix regularization including the Frobenius norm, sparse $L^1$-norm, and mixed $(2,1)$-norm (see definitions below). The learnt similarity matrix is used to design a sparse classification algorithm and we prove the generalization error of its resultant linear classifier can be bounded by the derived generalization bound for similarity learning. This implies that the proposed similarity learning with general matrix-norm regularization guarantees a good generalization for classification. Our techniques using the Rademacher complexity (Bartlett & Mendelson, 2002) and the important Khinchin-type inequality for the Rademacher variables enables us to derive bounds for general matrix-norm regularization including the sparse $L^1$-norm and mixed $(2,1)$-norm regularization.

The remainder of this paper is organized as follows. In Section 2, we propose the similarity learning formulations with general matrix-norm regularization terms and state the main theorems. In particular, the results will be illustrated using various examples. The related work is discussed in Section 3. The generalization bounds for similarity learning are established in Section 4. In Section 5, we develop a theoretical link between the generalization bounds of the proposed similarity learning method and the generalization error of the linear classifier built from the learnt similarity function. Section 6 estimates the Rademacher averages and gives the proof for examples stated in Section 2. Section 7 summarizes this paper and discuss some possible directions for future research.
2 Regularization Formulation and Main Results

In this section, we mainly introduce the regularized formulation of similarity learning and state our main results. Before we do that, let us introduce some notations and present some background material.

Denote, for any $n \in \mathbb{N}$, $\mathbb{N}_n = \{1, 2, \ldots, n\}$. Let $\mathbf{z} = \{z_i = (x_i, y_i) : i \in \mathbb{N}_m\}$ be a set of training samples, which is drawn identically and independently from a distribution $\rho$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Here, the input space $\mathcal{X}$ is a domain in $\mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$ is called the output space. Let $\mathbb{S}^d$ denote the set of $d \times d$ symmetric matrices. For any $x, x' \in \mathcal{X}$, we consider $K_A(x, x') = x^T Ax'$ as a bilinear similarity score parameterized by a symmetric matrix $A \in \mathbb{S}^d$. The symmetry of matrix $A$ guarantees the symmetry of the similarity score $K_A$, i.e. $K_A(x, x') = K_A(x', x)$.

The aim of similarity learning is to learn a matrix $A$ from a given set of training samples $\mathbf{z}$ such that the similarity score $K_A$ between examples from the same label is larger than that between examples from different labels. A natural approach to achieve the above aim is to minimize the following empirical error

$$
\mathcal{E}_\mathbf{z}(A) = \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left(1 - \frac{1}{mr} \sum_{j \in \mathbb{N}_m} y_i y_j K_A(x_i, x_j)\right)_+, \tag{1}
$$

where $r > 0$ is the margin. Note that $\sum_{j \in \mathbb{N}_m} y_i y_j K_A(x_i, x_j) = \sum_{\{j : y_j = y_i\}} K_A(x_i, x_j) - \sum_{\{j : y_j \neq y_i\}} K_A(x_i, x_j)$. Minimizing the above empirical error encourages, for any $i$, that, with margin $r$, the average similarity scores between examples with the same class as $y_i$ are relatively larger than those between examples with distinct classes from $y_i$. To avoid overfitting, we add a matrix-regularized term to the above empirical error and reach the
following regularization formulation

$$A_z = \arg \min_{A \in \mathbb{R}^d} \left[ \mathcal{E}_z(A) + \lambda \|A\| \right],$$  

(2)

where $\lambda > 0$ is the regularization parameter. Here, the notation $\|A\|$ denotes a general matrix norm. For instance, it can be the sparse $L^1$-norm $\|A\|_1 = \sum_{k \in \mathbb{N}_d} \sum_{\ell \in \mathbb{N}_d} |A_{k\ell}|$, the $(2, 1)$-mixed norm $\|A\|_{(2,1)} := \sum_{k \in \mathbb{N}_d} (\sum_{\ell \in \mathbb{N}_d} A_{k\ell}^2)^{\frac{1}{2}}$, the Frobenius norm $\|A\|_F = (\sum_{k, \ell \in \mathbb{N}_d} A_{k\ell}^2)^{\frac{1}{2}}$ or the trace norm $\|A\|_\text{tr} := \sum_{\ell \in \mathbb{N}_d} \sigma_\ell(A)$, where $\{\sigma_\ell(A) : \ell \in \mathbb{N}_d\}$ denote the singular values of matrix $A$. When $\|A\|$ is a Frobenius norm, formulation (2) has been proposed by Bellet at al. (Bellet et al., 2012).

The first contribution of this paper is to establish generalization bounds for regularized similarity learning (1) with general matrix-norms. Specifically, define

$$\mathcal{E}(A) = \int \left( 1 - \frac{1}{r} \int \int \right) yy' K_A(x, x') d\rho(x', y') + d\rho(x, y).$$  

(3)

The target of generalization analysis for similarity learning is to bound $\mathcal{E}(A_z) - \mathcal{E}_z(A_z)$.

Its special case with the Frobenius matrix norm was established in (Bellet et al., 2012). It used the uniform stability techniques (Bousquet & Elisseeff, 2002), which, however, can not deal with non-strongly convex matrix-norms such as the $L^1$-norm, $(2, 1)$-mixed norm and trace norm. Our new analysis techniques are able to deal with general matrix norms, which depend on the concept of Rademacher averages (Bartlett & Mendelson, 2002) defined as follows.

**Definition 1.** Let $\mathcal{F}$ be a class of uniformly bounded functions. For every integer $n$, we call

$$R_n(\mathcal{F}) := \mathbb{E}_z \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in \mathbb{N}_n} \sigma_i f(z_i) \right],$$
the Rademacher average over $\mathcal{F}$, where $\{z_i : i \in \mathbb{N}_n\}$ are independent random variables distributed according to some probability measure and $\{\sigma_i : i \in \mathbb{N}_n\}$ are independent Rademacher random variables, that is, $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$.

Before stating our generalization bounds for similarity learning, we first introduce some notations. For any $B, A \in \mathbb{R}^{n \times d}$, let $\langle B, A \rangle = \text{trace}(B^T A)$, where trace($\cdot$) denotes the trace of a matrix. For any matrix-norm $\| \cdot \|$, its dual norm $\| \cdot \|_*$ is defined, for any $B$, by $\|B\|_* = \sup_{\|A\| \leq 1} \text{trace}(B^T A)$. Denote $\|X\|_* = \sup_{x, x' \in \mathcal{X}} \|x'x^T\|_*$. Let the Rademacher average with respect to the dual matrix norm be defined by

$$R_m := \mathbb{E}_{z, \sigma} \left[ \sup_{\tilde{x} \in \mathcal{X}} \left\| \frac{1}{m} \sum_{i \in \mathbb{N}_m} \sigma_i y_i \tilde{x}_i \tilde{x}_i^T \right\|_* \right].$$

(4)

Now we can state the generalization bounds for similarity learning, which is closely related to the Rademacher averages with respect to the dual matrix-norm $\| \cdot \|_*$.

**Theorem 1.** Let $A_z$ be the solution to algorithm (2). Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds

$$\mathcal{E}_z(A_z) - \mathcal{E}(A_z) \leq \frac{6R_m}{r\lambda} + \frac{2X_*}{r\lambda} \sqrt{\frac{2 \ln \left( \frac{1}{\delta} \right)}{m}}.$$  

(5)

The proof for Theorem 1 will be given in Section 4. Recently, some studies (Cao et al., 2012; Kar & Jain, 2011; Kumar et al., 2012) gave generalization bounds for similarity (kernel) learning, where the involved empirical term is a U-statistics term. Hence, the natural idea in the above studies is to explore properties of U-statistics (Clémençon et al., 2008; De La Peña & Giné, 1999) for analyzing the related similarity learning formulations. The proof for Theorem 1 differs from earlier approaches (Cao et al., 2012; Kar & Jain, 2011) since the empirical term (1) in formulation (2) is not a U-statistics term; see more discussions in Section 3 below.
Following the exact argument as the proof for Theorem 1, similar result is also true if we switch the position of $\mathcal{E}_z(A_z)$ and $\mathcal{E}(A_z)$, i.e. for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\mathcal{E}(A_z) \leq \mathcal{E}_z(A_z) + \frac{6R_m}{r \lambda} + \frac{2X^*_s}{r \lambda} \sqrt{\frac{2\ln(1/\delta)}{m}}.$$ 

The second contribution of this paper is to investigate the theoretical relationship between similarity learning (2) and the generalization error of the linear classifier built from the learnt metric $A_z$. We show that the generalization bound for the similarity learning gives an upper bound for the generalization error of a linear classifier produced by the linear Support Vector Machine (SVM) (Vapnik, 1998) defined as follows:

$$f_z = \arg\min_{f \in \mathcal{F}_z} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} (1 - y_i f(x_i))_+ \right\}$$

s.t. $\Omega(f) =: \sum_{j \in \mathbb{N}_m} |\alpha_j| \leq 1/r$, (6)

where $\mathcal{F}_z = \{ f : f = \sum_{j \in \mathbb{N}_m} \alpha_j K_{A_z}(x_j, \cdot), \, \alpha_j \in \mathbb{R} \}$ is a sample-dependent hypothesis space. The empirical error of $f \in \mathcal{F}_z$ associated with $z$ is defined by

$$\mathcal{E}_z(f) = \frac{1}{m} \sum_{i \in \mathbb{N}_m} (1 - y_i f(x_i))_+.$$ 

The true generalization error is defined as

$$\mathcal{E}(f) = \int_Z (1 - y f(x))_+ \, d\rho(x, y).$$

Now we are in a position to state the relationship between the similarity learning and the generalization error of the linear classifier.

**Theorem 2.** Let $A_z$ and $f_z$ be defined by (2) and (6), respectively. Then, for any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds

$$\mathcal{E}(f_z) \leq \mathcal{E}_z(A_z) + \frac{4R_m}{\lambda r} + \frac{2X^*_s}{\lambda r} \sqrt{\frac{2\log \frac{1}{\delta}}{m}}.$$  

(7)
The proof for Theorem 2 will be established in Section 5.

Theorems 1 and 2 depend critically on two terms: the constant $X_*$ and the Rademacher average $R_m$. Below, we list the estimation of these two terms associated with different matrix norms. For any vector $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, denote $\|x\|_\infty = \max_{\ell \in \mathbb{N}_d} |x^\ell|$.

**Example 1.** Consider the matrix norm be the sparse $L^1$-norm defined, for any $A \in \mathbb{S}^d$, by $\|A\| = \sum_{k,\ell \in \mathbb{N}_d} |A_{k\ell}|$. Let $A_z$ and $f_z$ be defined respectively by (2) and (6). Then, we have the following results.

(a) $X_* \leq \sup_{x \in \mathcal{X}} \|x\|_\infty^2$ and $R_m \leq 2 \sup_{x \in \mathcal{X}} \|x\|_\infty^2 \sqrt{\frac{e \log(d+1)}{m}}$.

(b) For any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds

$$E_z(A_z) - E(A_z) \leq \frac{12 \sup_{x \in \mathcal{X}} \|x\|_\infty^2}{r \lambda} \sqrt{\frac{e \log(d+1)}{m}} + \frac{2 \sup_{x \in \mathcal{X}} \|x\|_\infty^2}{r \lambda} \sqrt{\frac{2 \ln \frac{1}{\delta}}{m}}.$$  

(8)

(c) For any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds

$$E(f_z) \leq E_z(A_z) + \frac{4 \sup_{x \in \mathcal{X}} \|x\|_\infty^2}{\lambda r} \sqrt{\frac{2 e \log(d+1)}{m}} + \frac{2 \sup_{x \in \mathcal{X}} \|x\|_\infty^2}{\lambda r} \sqrt{\frac{2 \log \frac{1}{\delta}}{m}}.$$  

(9)

For any vector $x \in \mathbb{R}^d$, let $\|x\|_F$ be the standard Euclidean norm. Considering the regularized similarity learning with the Frobenius matrix norm, we have the following result.

**Example 2.** Consider the Frobenius matrix norm defined, for any $A \in \mathbb{S}^d$, by $\|A\| = \sqrt{\sum_{k,\ell \in \mathbb{N}_d} |A_{k\ell}|^2}$. Let $A_z$ and $f_z$ be defined by (2) and (6), respectively. Then, we have the following estimation.

(a) $X_* \leq \sup_{x \in \mathcal{X}} \|x\|_F^2$ and $R_m \leq 2 \sup_{x \in \mathcal{X}} \|x\|_F^2 \sqrt{\frac{\lambda}{m}}$.  

10
(b) For any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds
\[
\mathcal{E}_z(A_z) - \mathcal{E}(A_z) \leq \frac{6 \sup_{x \in \mathcal{X}} \|x\|_F^2}{r \lambda \sqrt{m}} + \frac{2 \sup_{x \in \mathcal{X}} \|x\|_F^2}{\lambda \sqrt{2 \ln \left(\frac{1}{\delta}\right)}}.\quad (10)
\]

(c) For any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds
\[
\mathcal{E}(f_z) \leq \mathcal{E}_z(A_z) + \frac{4 \sup_{x \in \mathcal{X}} \|x\|_F^2}{\lambda r \sqrt{m}} + \frac{2 \sup_{x \in \mathcal{X}} \|x\|_F^2}{\lambda r} \sqrt{2 \log \frac{1}{\delta}}.\quad (11)
\]

We end this section with two remarks. Firstly, the above theorem and examples mean that a good similarity (i.e. a small $\mathcal{E}_z(A_z)$ for similarity learning) can guarantee a good classification (i.e. a small generalization error $\mathcal{E}(f_z)$). Secondly, the bounds in Example 2 is consistent with that in (Bellet et al., 2012).

3 Related Work

In this section, we discuss studies on similarity metric learning which are related to our work.

Many similarity metric learning methods have been motivated by the intuition that the similarity score between examples in the same class should be larger than that of examples from distinct classes (see, e.g., Bar-Hillel et al., 2005; Cao et al., 2012; Chechik et al., 2010; Hoi et al., 2006; Jin et al., 2009; Maurer, 2008; Blitzer et al., 2005; Xing et al., 2002). Jin et al. (Jin et al., 2009) established generalization bounds for regularized metric learning algorithms via the concept of uniform stability (Bartlett & Mendelson, 2002), which, however, only works for strongly convex matrix regularization terms. A very recent work (Cao et al., 2012) established generalization bounds for metric and similarity learning associated with general matrix norm regularization using techniques
of Rademacher averages and U-statistics. However, there was no theoretical links between the similarity metric learning and the generalization performance of classifiers based on the learnt similarity matrix. Here, we focus on the problem how to learn a good linear similarity function $K_A$ such that it can guarantee a good classifier derived from the learnt similarity function. In addition, our formulation (2) is quite distinct from similarity metric learning methods (Cao et al., 2012; Chechik et al., 2010), since they are based on pairwise or triplet-wise constraints and consider the following pairwise empirical objective function:

$$\frac{1}{m(m-1)} \sum_{i,j=1, i \neq j}^{m} (1 - y_i y_j (K_A(x_i, x_j) - r))^+.$$  

(12)

Our formulation (2) is less restrictive since the empirical objective function is defined over an average of similarity scores and it doesn’t require the positive semi-definiteness of the similarity function $K$.

Balcan et al. (Balcan et al., 2008) developed a theory of $(\epsilon, \gamma, \tau)$-good similarity function defined as follows. It attempts to investigate the theoretical relationship between the properties of a similarity function and its performance in linear classification.

**Definition 2.** (Balcan & Blum, 2006) A similarity function $K$ is a $(\epsilon, \gamma, \tau)$-good similarity function in hinge loss for a learning problem $P$ if there exists a random indicator function $R(x)$ defining a probabilistic set of “reasonable points” such that the following conditions hold:

1. $\mathbb{E}_{(x,y) \sim P}[1 - yg(x)/\gamma]^+ \leq \epsilon$, where $g(x) = \mathbb{E}_{(x',y') \sim P}[y'K(x, x')|R(x')]$,
2. $\Pr_{x'}[R(x')] \geq \tau$.

The first condition can be interpreted as “$1 - \epsilon$ proportion of points $x$ are on average $2\gamma$
more similar to random reasonable points of the same class than to random reasonable points of the distinct classes” and the second condition as “at least a $\tau$ proportion of the points should be reasonable.” The following theorem implies that if given an $(\epsilon, \gamma, \tau)$-good similarity function and enough landmarks, there exists a separator $\alpha$ with error arbitrarily close to $\epsilon$.

**Theorem 3.** (Balcan & Blum, 2006) Let $K$ be an $(\epsilon, \gamma, \tau)$-good similarity function in hinge loss for a learning problem $P$. For any $\epsilon_1 > 0$, and $0 < \delta \leq \gamma \epsilon_1 / 4$, let $S = \{x'_1, \ldots, x'_d_{\text{land}}\}$ be a potentially unlabeled sample of $d_{\text{land}} = \frac{2}{\tau} \left( \log (2/\delta) + 16 \frac{\log (2/d)}{\epsilon_1 \gamma^2} \right)$ landmarks drawn from $P$. Consider the mapping $\phi_i^S = K(x, x'_i)$, $i \in \{1, \ldots, d_{\text{land}}\}$.

Then, with probability at least $1 - \delta$ over the random sample $S$, the induced distribution $\phi^S(P)$ in $\mathbb{R}^{d_{\text{land}}}$ has a linear separator $\alpha$ of error at most $\epsilon + \epsilon_1$ at margin $\gamma$.

It was mentioned in (Balcan et al., 2008) that the linear separator can be estimated by solving the following linear programming if we have $d_u$ potentially unlabeled sample and $d_l$ labeled sample,

$$
\min_{\alpha} \left\{ \sum_{i=1}^{d_l} \left[ 1 - \sum_{j=1}^{d_u} \alpha_j y_i K(x_i, x'_j) \right] : \sum_{j=1}^{d_u} |\alpha_j| \leq 1/\gamma \right\}.
$$

(13)

The above algorithm (13) is quite similar to the linear SVM (6) used in our paper. Our work is distinct from Balcan et al. (Balcan et al., 2008) in the following two aspects. Firstly, the similarity function $K$ is predefined in algorithm (13), while we aim to learn a similarity function $K_{A_u}$ from a regularized similarity learning formulation (2). Secondly, although the separators are both trained from the linear SVM, the classification algorithm (13) in (Balcan et al., 2008) was designed using two different sets of examples, a set of labeled samples of size $d_l$ to train the classification algorithm and another
set of unlabeled samples with size $d_u$ to define the mapping $\phi^S$. In this paper, we used
the same set of training samples for both similarity learning (2) and the classification
algorithm (6).

Recent work by Bellet et al. (Bellet et al., 2012) is mostly close to ours. Specifically,
they considered similarity learning formulation (2) with the Frobenius norm regular-
ization. Generalization bounds for similarity learning were derived via uniform sta-
bility arguments (Bousquet & Elisseeff, 2002) which can not deal with, for instance,
the $L^1$-norm and $(2,1)$-norm regularization terms. In addition, the results about the
relationship between the similarity learning and the performance of the learnt matrix in
classification were quoted from (Balcan et al., 2008) and hence requires two separate
sets of samples to train the classifier.

Most recently, there is a considerable interest on two-stage approaches for multiple
kernel learning (Cortes et al., 2010b; Kar, 2013) which perform competitively as the
one-stage approaches (Lanckriet et al., 2004; Varma & Babu., 2009). In particular, Kar
(Kar, 2013) studied generalization guarantees for the following regularization formula-
tion for learning similarity (kernel) function proposed in (Kumar et al., 2012):

$$\arg \min_{\mu \geq 0} \frac{2}{m(m-1)} \sum_{1 \leq i < j \leq n} \left( 1 - y_i y_j K_{\mu}(x_i, x_j) \right)_+ + \frac{\lambda}{2} \Omega(\mu),$$  \hspace{1cm} (14)

where $K_{\mu} = \sum_{\ell=1}^p \mu_\ell K_\ell$ is the positive linear combination of base kernels $\{K_\ell : \ell = 1, 2, \ldots, p\}$, $\lambda > 0$ and $\Omega(\cdot)$ is a regularization term which, for instance, can be the
Frobenius norm or the $L^1$ norm. Specifically, Kar (Kar, 2013) established elegant generalization bounds for the above two-stage multiple kernel learning using techniques
of Rademacher averages (Bartlett & Mendelson, 2002; Kakade et al., 2008, 2012) and
U-statistics (Cao et al., 2012; Clémenccon et al., 2008). The empirical error term (1) in
our formulation (2) is not a U-statistics term and the techniques in (Kar, 2013; Cao et al., 2012) can not directly be applied to our case.

Kar and Jain (Kar & Jain, 2011, 2012) introduced an extended framework of (Balcan & Blum, 2006; Balcan et al., 2008) in the general setting of supervised learning. The authors proposed a general goodness criterion for similarity functions, which can handle general supervised learning tasks and also subsumes the goodness of condition of (Balcan et al., 2008). There, efficient algorithms were constructed with provable generalization error bounds. The main distinction between these work and our work is that we aim to learn a similarity function while in their work a similarity function is defined in advance.

4 Generalization Bounds for Similarity Learning

In this section, we establish generalization bounds for the similarity learning formulation (2) with general matrix-norm regularization terms. Recall that the true error for similarity learning is defined by

$$\mathcal{E}(A) = \int_{\mathbb{Z}} \left(1 - \frac{1}{r} \int_{\mathbb{Z}} yy' K_A(x, x') d\rho(x', y') \right) + d\rho(x, y).$$

The target of generalization analysis for similarity learning is to bound the empirical error $\mathcal{E}_z(A_z)$ by the true error $\mathcal{E}(A_z)$.

By the definition (2) of $A_z$, we know that $\mathcal{E}_z(A_z) + \lambda \|A_z\| \leq \mathcal{E}_z(0) + \lambda \|0\| = 1$, which implies that $\|A_z\| \leq 1/\lambda$. Denote

$$\mathcal{A} = \left\{ A \in \mathbb{S}^d : \|A\| \leq 1/\lambda \right\}.$$
Hence, one can easily see that the solution $A_z$ to algorithm (2) belongs to $A$. Now we are ready to prove generalization bounds for similarity learning which was stated as Theorem 1 in Section 2.

**Proof of Theorem 1:** Our proof is divided into two steps.

**Step 1:** Let $\mathbb{E}_z$ denote the expectation with respect to samples $z$. Observe that 

$$
\mathcal{E}_z(A_z) - \mathcal{E}(A_z) \leq \sup_{A \in \mathcal{A}} \left[ \mathcal{E}_z(A) - \mathcal{E}(A) \right].
$$

Also, for any $z = (z_1, \ldots, z_k, \ldots, z_m)$ and $\tilde{z} = (z_1, \ldots, \tilde{z}_k, \ldots, z_m)$, $1 \leq k \leq m$, there holds

$$
\left| \sup_{A \in \mathcal{A}} \left[ \mathcal{E}_z(A) - \mathcal{E}(A) \right] - \sup_{A \in \mathcal{A}} \left[ \mathcal{E}_{\tilde{z}}(A) - \mathcal{E}(A) \right] \right| 
\leq \sup_{A \in \mathcal{A}} \left| \mathcal{E}_z(A) - \mathcal{E}_{\tilde{z}}(A) \right| 
\leq \frac{1}{m^2 r} \sup_{A \in \mathcal{A}} \left\{ \sum_{i=1, i \neq k}^{m} \left| y_i y_k K_A(x_k, x_i) - y_i \tilde{y}_k K_A(\tilde{x}_k, x_i) \right| 
+ \left| \sum_{j \in N^m} (y_k y_j K_A(x_k, x_j) - \tilde{y}_k y_j K_A(\tilde{x}_k, x_j)) \right| \right\}
\leq \frac{2}{m^2 r} \sup_{A \in \mathcal{A}} \sum_{i \in N^m} \left( |y_i y_k K_A(x_k, x_i)| + |y_i \tilde{y}_k K_A(\tilde{x}_k, x_i)| \right) 
\leq \frac{4X^*}{mr}.
$$

Applying the McDiarmid’s inequality (McDiarmid, 1989) (see Lemma 1 in the Appendix) to the term $\sup_{A \in \mathcal{A}} \left[ \mathcal{E}_z(A) - \mathcal{E}(A) \right]$, with probability at least $1 - \delta$, there holds

$$
\sup_{A \in \mathcal{A}} \left[ \mathcal{E}_z(A) - \mathcal{E}(A) \right] 
\leq \mathbb{E}_z \sup_{A \in \mathcal{A}} \left[ \mathcal{E}_z(A) - \mathcal{E}(A) \right] 
+ \frac{2X^*}{r \lambda} \sqrt{\frac{2 \ln \left( \frac{1}{\delta} \right)}{m}}.
$$

(15)

Now we are in a position to estimate the first term in the expectation form on the right-hand side of the above equation by standard symmetrization techniques.
Step 2: We divide the term \( \mathbb{E}_z \sup_{A \in A} \left[ \mathcal{E}_z(A) - \mathcal{E}(A) \right] \) into two parts as follows,

\[
\mathbb{E}_z \sup_{A \in A} \left[ \mathcal{E}_z(A) - \mathcal{E}(A) \right] = \mathbb{E}_z \sup_{A \in A} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{mr} \sum_{j \in \mathbb{N}_m} y_i y_j K_A(x_i, x_j) \right] + \mathcal{E}(A) \right\} \\
= \mathbb{E}_z \sup_{A \in A} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} y_i y_j K_A(x_i, x') \right] + \mathcal{E}(A) \\
- \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} y_i y_j K_A(x_i, x') \right] + \right\} \\
+ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{mr} \sum_{j \in \mathbb{N}_m} y_i y_j K_A(x_i, x_j) \right] + \right\} \\
\leq I_1 + I_2,
\]

where

\[
I_1 := \mathbb{E}_z \sup_{A \in A} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} y_i y_j K_A(x_i, x') \right] + \mathcal{E}(A) \right\},
\]

and

\[
I_2 := \mathbb{E}_z \sup_{A \in A} \left\{ - \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} y_i y_j K_A(x_i, x') \right] + \right\}.
\]

Now let \( z = \{ \bar{z}_1, \bar{z}_2, \ldots, \bar{z}_m \} \) be an i.i.d. sample which is independent of \( z \). We first estimate \( I_1 \) using the standard symmetrization techniques, to this end, we rewrite \( \mathcal{E}(A) \) as \( \mathbb{E}_z \left( \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} \bar{y}_i y_j K_A(\bar{x}_i, x') \right]_{+} \right) \). Then we have

\[
I_1 = \mathbb{E}_z \sup_{A \in A} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} y_i y_j K_A(x_i, x') \right]_{+} \right\} \\
- \mathbb{E}_z \left( \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} \bar{y}_i y_j K_A(\bar{x}_i, x') \right]_{+} \right) \\
\leq \mathbb{E}_{z,\bar{z}} \sup_{A \in A} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} \bar{y}_i y_j K_A(x_i, x') \right]_{+} \right\} \\
- \frac{1}{m} \sum_{i \in \mathbb{N}_m} \left[ 1 - \frac{1}{r} \mathbb{E}_{(x',y')} \bar{y}_i y_j K_A(\bar{x}_i, x') \right]_{+}
\]
By the standard Rademacher symmetrization technique and the contraction property of
the Rademacher average (see Lemma 2 in the Appendix), we further have

\[ I_1 \leq 2E_{z,\sigma} \sup_{A \in A} \left\{ \frac{1}{m} \sum_{i \in N_m} \sigma_i \left[ 1 - \frac{1}{r} E_{(x',y')} y_i y' K_A(x_i, x') \right]_+ \right\} \]

\[ \leq 4E_{z,\sigma} \sup_{A \in A} \left\langle \frac{1}{mr} \sum_{i \in N_m} \sigma_i y_i x_i \int y' x'^T d\rho(x', y'), A \right\rangle \]

\[ \leq \frac{4}{r} E_{z,\sigma} \left\| \frac{1}{m} \sum_{i \in N_m} \sigma_i y_i x_i \int y' x'^T d\rho(x', y') \right\|_* \]

\[ \leq \frac{4}{r} E_{z,\sigma} \sup_{\bar{x}} \left\| \frac{1}{m} \sum_{i \in N_m} \sigma_i y_i x_i \bar{x}^T \right\|_* , \]

where the last inequality follows from the fact that \( \langle A, B \rangle \leq ||A|| \cdot ||B||_* \leq \frac{1}{r} ||B||_* \)

for any \( A \in A \) and \( B \in \mathbb{R}^{d \times d} \).

Similarly, we can estimate \( I_2 \) as follows.

\[ I_2 = E_{z} \sup_{A \in A} \frac{1}{m} \sum_{i \in N_m} \left( \left[ 1 - \frac{1}{mr} \sum_{j \in N_m} y_i y_j K_A(x_i, x_j) \right] \right. \]

\[ - \left. \left[ 1 - \frac{1}{r} E_{(x',y')} y_i y' K_A(x_i, x') \right]_+ \right) \]

\[ \leq E_{z} \sup_{A \in A} \frac{1}{m} \sum_{i \in N_m} \left( \left| \frac{1}{r} E_{(x',y')} y_i y' K_A(x_i, x') - \frac{1}{mr} \sum_{j \in N_m} y_i y_j K_A(x_i, x_j) \right| \right) \]

\[ = E_{z} \sup_{A \in A} \frac{1}{mr} \sum_{i \in N_m} \left( \left| \frac{1}{r} E_{(x',y')} y_i y' x_i x_i^T - \frac{1}{m} \sum_{j \in N_m} y_j x_j x_i^T, A \right| \right) \]

\[ \leq \frac{1}{r} E_{z} \sup_{x \in \mathcal{X}} \left\| E_{(x',y')} y_i y' x_i^T - \frac{1}{m} \sum_{j \in N_m} y_j x_j x_i^T \right\|_* \]

\[ = \frac{1}{r} E_{z} \sup_{x \in \mathcal{X}} \left\| E_{(x',y')} \frac{1}{m} \sum_{j \in N_m} y_j x_j x_i^T - \frac{1}{m} \sum_{j \in N_m} y_j x_j x_i^T \right\|_* . \]

In the above estimation, the first inequality follows from the Lipschitz continuity of
the hinge loss function. Following the standard Rademacher symmetrization technique
(see, e.g., Bartlett & Mendelson, 2002), from the above estimation we can further esti-
mate $I_2$ as follows:

$$I_2 \leq \frac{1}{r \lambda} \mathbb{E}_{z, x^*} \sup_{x \in \mathcal{X}} \left\| \mathbb{E}_{z, x^*} \frac{1}{m} \sum_{j \in \mathbb{N}_m} y'_j x'_j x^T - \frac{1}{m} \sum_{j \in \mathbb{N}_m} y_j x_j x^T \right\|_*$$

$$\leq \frac{1}{r \lambda} \mathbb{E}_{z, x^*, o} \sup_{x \in \mathcal{X}} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j (y'_j x'_j x^T - y_j x_j x^T) \right\|_*$$

$$\leq \frac{2}{r \lambda} \mathbb{E}_{z, o} \sup_{x \in \mathcal{X}} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j x^T \right\|_*.$$

The desired result follows by combining (15) with the above estimation for $I_1$ and $I_2$.

This completes the proof for the theorem. □

5 Guaranteed Classification Via Good Similarity

In this section, we investigate the theoretical relationship between the generalization error of the similarity learning and that of the linear classifier built from the learnt similarity metric $K_{A_z}$. In particular, we will show that the generalization error of the similarity learning gives an upper bound for the generalization error of the linear classifier which was stated as Theorem 2 in Section 2.

Before giving the proof of Theorem 2, we first establish the generalization bounds for the linear SVM algorithm (6). Recalling that the linear SVM algorithm (6) was defined by

$$f_z = \arg \min_{f \in F_z} \left\{ \frac{1}{m} \sum_{i \in \mathbb{N}_m} (1 - y_i f(x_i))^+ \right\}$$

s.t. $\Omega(f) = \sum_{j \in \mathbb{N}_m} |\alpha_j| \leq 1/r$,

where

$$F_z = \left\{ f : f = \sum_{j \in \mathbb{N}_m} \alpha_j K_{A_z}(x, \cdot), \alpha_j \in \mathbb{R} \right\}.$$
The generalization analysis of the linear SVM algorithm (6) aims to estimate the term $\mathcal{E}(f_z) - \mathcal{E}_z(f_z)$. For any $z$, one can easily see that the solution to algorithm (6) belongs to the set $\mathcal{F}_{z,r}$, where

$$\mathcal{F}_{z,r} = \left\{ f = \sum_{j \in \mathbb{N}_m} \alpha_j K_{A_z}(x_j, \cdot) : \Omega(f) = \sum_{j \in \mathbb{N}_m} |\alpha_j| \leq 1/r, \alpha_j \in \mathbb{R} \right\}.$$ 

To perform the generalization analysis, we seek a sample-independent set which contains, for any $z$, the sample-dependent hypothesis space $\mathcal{F}_z$. Specifically, we define a sample independent hypothesis space by

$$\mathcal{F}_m = \left\{ f = \sum_{i \in \mathbb{N}_m} \alpha_i K_{A}(u_i, \cdot) : \|A\| \leq 1/\lambda, u_j \in X, \alpha_j \in \mathbb{R} \right\}.$$

Recalling that, for any $z$, $\|A_z\| \leq \lambda^{-1}$, one can easily see that $\mathcal{F}_z$ is a subset of $\mathcal{F}_m$. It follows that, for any $z$, the solution to the linear SVM algorithm (6) lies in the set $\mathcal{F}_{m,r}$, which is given by

$$\mathcal{F}_{m,r} = \left\{ f \in \mathcal{F}_m : \Omega(f) \leq 1/r \right\}.$$

The following theorem states the generalization bounds of the linear SVM for classification.

**Theorem 4.** Let $f_z$ be the solution to the algorithm (6). For any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\mathcal{E}(f_z) - \mathcal{E}_z(f_z) \leq 4R_m \frac{1}{\lambda r} + \frac{2X_s}{\lambda r} \sqrt{\frac{2 \log \frac{1}{\delta}}{m}}. \quad (16)$$

**Proof.** By McDiarmid’s inequality, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\mathcal{E}(f_z) - \mathcal{E}_z(f_z) \leq \sup_{f \in \mathcal{F}_{z,r}} \left( \mathcal{E}(f) - \mathcal{E}_z(f) \right) \leq \sup_{f \in \mathcal{F}_{m,r}} \left( \mathcal{E}(f) - \mathcal{E}_z(f) \right)$$

$$\leq \mathbb{E}_z \sup_{f \in \mathcal{F}_{m,r}} \left( \mathcal{E}(f) - \mathcal{E}_z(f) \right) + \frac{2X_s}{\lambda r} \sqrt{\frac{2 \log \frac{1}{\delta}}{m}}.$$
Next, all we need is to estimate the first part of the right hand-side of the above inequality. Let $\bar{z}$ be an independent sample (independent each other and $z$) and with the same distribution as $z$.

$$
\mathbb{E}_z \sup_{f \in \mathcal{F}_{m,r}} (\mathcal{E}(f) - \mathcal{E}_z(f)) = \mathbb{E}_z \sup_{f \in \mathcal{F}_{m,r}} (\mathbb{E}_z \mathcal{E}_z(f) - \mathcal{E}_z(f))
$$

$$
\leq \mathbb{E}_{z,\bar{z}} \sup_{f \in \mathcal{F}_{m,r}} (\mathcal{E}_{\bar{z}}(f) - \mathcal{E}_z(f))
$$

$$
\leq 2\mathbb{E}_{z,\sigma} \left[ \sup_{||A|| \leq 1/\lambda} \sup_{i \in N_m} |\alpha_i| \leq 1/r \left( \frac{1}{m} \sum_{i \in N_m} \sigma_i \left[ 1 - \sum_{j \in N_m} \alpha_j y_i K_A(x_i, u_j) \right] \right) \right]
$$

$$
\leq 4\mathbb{E}_{z,\sigma} \sup_{x \in \mathcal{X}} \left( \frac{1}{m} \sum_{i \in N_m} \sigma_i y_i \langle x_i x^T, A \rangle \right)
$$

$$
\leq \frac{4}{r} \mathbb{E}_{z,\sigma} \sup_{x \in \mathcal{X}} \left\| \frac{1}{m} \sum_{i \in N_m} \sigma_i y_i x_i x^T \right\|_*,
$$

Here we also use the standard Rademacher Symmetrization technique and the contractor property of the Rademacher average. Then the proof is completed. \qed

Now we are in a position to give the proof of Theorem 2.

**Proof of Theorem 2:** If we take $\alpha^0 = (\frac{y_1}{m}, \ldots, \frac{y_m}{m})^T$, then $f^0_z = \frac{1}{m} \sum_{j \in N_m} y_j K_A(x_j, \cdot)$.

One can easily see that $\Omega(f^0_z) = \sum_{j \in N_m} |\alpha^0_j| = \frac{1}{r}$, that means $f^0_z \in \mathcal{F}_{z,r}$. From Theorem 4 and the definition of $f_z$, we get

$$
\mathcal{E}(f_z) \leq \mathcal{E}_z(f_z) + \frac{R_m}{\lambda r} + 2X_* \sqrt{\frac{\log \frac{1}{\delta}}{m}} \leq \mathcal{E}_z(f^0_z) + \frac{R_m}{\lambda r} + 2X_* \sqrt{\frac{\log \frac{1}{\delta}}{m}}
$$

$$
= \frac{1}{m} \sum_{i \in N_m} \left( 1 - \frac{1}{mr} \sum_{j \in N_m} y_i y_j K_A(x_i, x_j) \right) + \frac{R_m}{\lambda r} + 2X_* \sqrt{\frac{\log \frac{1}{\delta}}{m}}
$$

$$
= \mathcal{E}_z(A_z) + \frac{R_m}{\lambda r} + 2X_* \sqrt{\frac{\log \frac{1}{\delta}}{m}}.
$$

This completes the proof of the theorem. \qed
6 Estimating Rademacher Averages

The main theorems above critically depend on the estimation of the Rademacher average $R_m$ defined by equation (4). In this section, we establish a self-contained proof for this estimation and prove the examples listed in Section 2. For notational simplicity, denote by $x_i^\ell$ the $\ell$-th variable of the $i$-th sample $x_i \in \mathbb{R}^d$.

Proof of Example 1: The dual norm of $L^1$-norm is the $L^\infty$-norm. Hence,

$$X_* = \sup_{x,x' \in \mathcal{X}} \sup_{\ell,k \in \mathbb{N}_d} |x^\ell(x')^k| = \sup_{x \in \mathcal{X}} \|x\|_\infty^2.$$ (17)

Also, the Rademacher average can be rewritten as

$$R_m = \mathbb{E}_{\sigma} \sup_{x \in \mathcal{X}} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j^T \right\|_\infty \leq \sup_{x \in \mathcal{X}} \|x\|_\infty \mathbb{E}_{\sigma} \max_{\ell \in \mathbb{N}_d} \left| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j^\ell \right|. \tag{18}$$

Now let $U_\ell(\sigma) = \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j^\ell$, for any $\ell \in \mathbb{N}_d$. By Jensen’s inequality, for any $\eta > 0$, we have

$$e^{\eta^2 (\mathbb{E}_{\sigma} \max_{\ell \in \mathbb{N}_d} |U_\ell(\sigma)|)^2} - 1 \leq \mathbb{E}_{\sigma} [e^{\eta^2 (\max_{\ell \in \mathbb{N}_d} |U_\ell(\sigma)|)^2} - 1]$$

$$= \mathbb{E}_{\sigma} [\max_{\ell \in \mathbb{N}_d} e^{\eta^2 |U_\ell(\sigma)|^2} - 1] \leq \sum_{\ell \in \mathbb{N}_d} \mathbb{E}_{\sigma} [e^{\eta^2 |U_\ell(\sigma)|^2} - 1]. \tag{19}$$

Furthermore, for any $\ell \in \mathbb{N}_d$, there holds

$$\mathbb{E}_{\sigma} [e^{\eta^2 |U_\ell(\sigma)|^2} - 1] = \sum_{k \geq 1} \frac{1}{k!} \eta^{2k} \mathbb{E}_{\sigma} |U_\ell|^{2k} \leq \sum_{k \geq 1} \frac{1}{k!} \eta^{2k} (2k - 1)^k (\mathbb{E}_{\sigma} |U_\ell|^2)^k \leq \sum_{k \geq 1} (2e\eta^2 \mathbb{E}_{\sigma} |U_\ell|^2)^k,$$

where the first inequality follows from the Khinchin-type inequality (see Lemma 3 in the Appendix), and the second inequality holds due to the Stirling’s inequality: $e^{-k} k^k \leq k!$. Now set $\eta = \sqrt{e \max_{\ell \in \mathbb{N}_d} (\mathbb{E}_{\sigma} |U_\ell|^2)^{1/2}} - 1$. The above inequality can be upper bounded by

$$\mathbb{E} [e^{\eta^2 |U_\ell(\sigma)|^2} - 1] \leq \sum_{k \geq 1} 2^{-k} = 1, \quad \forall \ell \in \mathbb{N}_d.$$
Putting the above estimation back into (19) implies that
\[ e^{\eta^2 (\mathbb{E}_{|U_\ell(\sigma)|})^2} - 1 \leq d. \]
That means
\[
\mathbb{E}_\sigma \max_{\ell \in \mathbb{N}_d} |U_\ell(\sigma)| = \mathbb{E}_\sigma \max_{\ell \in \mathbb{N}_d} \left| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_\ell^T \right| \leq \sqrt{\log(d + 1) \eta^{-2}}
\]
\[ = 2 \sqrt{e \log(d + 1)} \max_{\ell \in \mathbb{N}_d} \left( \mathbb{E}_\sigma \left| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_\ell^T \right|^2 \right)^{\frac{1}{2}}
\]
\[ = 2 \sqrt{e \log(d + 1)} \max_{\ell \in \mathbb{N}_d} \left( \mathbb{E}_\sigma \left| \frac{1}{m^2} \sum_{j, k \in \mathbb{N}_m} \sigma_j \sigma_k y_j y_k x_\ell^T x_\ell^T \right|^2 \right)^{\frac{1}{2}}
\]
\[ = 2 \sqrt{e \log(d + 1)} \max_{\ell \in \mathbb{N}_d} \left( \frac{1}{m^2} \sum_{j \in \mathbb{N}_m} (x_\ell^T)^2 \right)^{\frac{1}{2}} \leq 2 \sup_{x \in \mathbb{X}} \|x\|_\infty \sqrt{\frac{e \log(d + 1)}{m}}. \]
Putting the above estimation back into (18) implies that
\[ \mathcal{R}_m \leq 2 \sup_{x \in \mathbb{X}} \|x\|_2^2 \sqrt{\frac{e \log(d + 1)}{m}}. \]
The other desired results in the example follow directly from combining the above estimation with Theorems 1 and 2.

We turn our attention to similarity learning formulation (2) with the Frobenius norm regularization.

**Proof of Example 2:** The dual norm of the Frobenius norm is itself. Consequently,
\[ X_* = \sup_{x, x' \in \mathbb{X}} \|x^T x'\|_F = \sup_{x \in \mathbb{X}} \|x\|_F^2. \]
The Rademacher average can be rewritten as
\[ \mathcal{R}_m = \mathbb{E}_{z, \sigma} \sup_{x \in \mathbb{X}} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j x_j^T \right\|_F. \]
By Cauchy’s Inequality, there holds

\[
R_m = \mathbb{E}_{z, \sigma} \sup_{x \in \mathcal{X}} \|x\|_F \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j \right\|_F \\
\leq \sup_{x \in \mathcal{X}} \|x\|_F \mathbb{E}_z \left( \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j \right)_F^{1/2} \\
= \sup_{x \in \mathcal{X}} \|x\|_F \mathbb{E}_z \left( \sum_{j \in \mathbb{N}_m} \|x_j\|_F^2 \right)^{1/2} / m \leq \sup_{x \in \mathcal{X}} \|x\|_F \frac{1}{\sqrt{m}}. \tag{21}
\]

Then, the desired results can be derived by combining the above estimation with Theorems 1 and 2. □

The above generalization bound for similarity learning formulation (2) with the Frobenius norm regularization is consistent with that given in (Bellet et al., 2012), where the result holds true under the assumption that \( \sup_{x \in \mathcal{X}} \|x\|_F \leq 1 \). Below, we provide the estimation of \( R_m \) respectively for the mixed \((2,1)\)-norm and the trace norm.

**Example 3.** Consider similarity learning formulation (2) with the mixed \((2,1)\)-norm regularization \( \|A\|_{(2,1)} = \sum_{k \in \mathbb{N}_d} \left( \sum_{\ell \in \mathbb{N}_d} |A_{k\ell}|^2 \right)^{1/2} \). Then, we have the following estimation.

(a) \( X_* \leq \left[ \sup_{x \in \mathcal{X}} \|x\|_F \right] \left[ \sup_{x \in \mathcal{X}} \|x\|_\infty \right] \) and

\[
R_m \leq 2 \left[ \sup_{x \in \mathcal{X}} \|x\|_F \right] \left[ \sup_{x \in \mathcal{X}} \|x\|_\infty \right] \sqrt{\frac{e \log(d + 1)}{m}}.
\]

(b) For any \( 0 < \delta < 1 \), with confidence at least \( 1 - \delta \), there holds

\[
\mathcal{E}_z(A_*) - \mathcal{E}(A_*), \leq 12 \left[ \sup_{x \in \mathcal{X}} \|x\|_F \right] \left[ \sup_{x \in \mathcal{X}} \|x\|_\infty \right] \sqrt{\frac{e \log(d + 1)}{m}} + 2 \left[ \sup_{x \in \mathcal{X}} \|x\|_F \right] \left[ \sup_{x \in \mathcal{X}} \|x\|_\infty \right] \sqrt{\frac{2 \ln(\frac{1}{\delta})}{m}}. \tag{22}
\]

(c) For any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \), there holds

\[
\mathcal{E}(f_*) \leq \mathcal{E}_z(A_*) + \frac{4 \left[ \sup_{x \in \mathcal{X}} \|x\|_F \right] \left[ \sup_{x \in \mathcal{X}} \|x\|_\infty \right]}{\lambda} \sqrt{\frac{2e \log(d + 1)}{m}} + \frac{2 \left[ \sup_{x \in \mathcal{X}} \|x\|_F \right] \left[ \sup_{x \in \mathcal{X}} \|x\|_\infty \right]}{\lambda} \sqrt{\frac{2 \log\frac{1}{\delta}}{m}}.
\]
Proof. The dual norm of the \((2,1)\)-norm is the \((2,\infty)\)-norm, which implies that \(X^* = \sup_{x,x' \in \mathcal{X}} \|x'x^T\|_{(2,\infty)} = \sup_{x \in \mathcal{X}} \|x\|_F \sup_{x' \in \mathcal{X}} \|x'\|_\infty\) and

\[
\mathbb{E}_{z,\sigma} \sup_{x \in \mathcal{X}} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_jx_jx^T \right\|_* \leq \sup_{x \in \mathcal{X}} \|x\|_F \mathbb{E}_{z,\sigma} \max_{\ell \in \mathbb{N}_d} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_jx^T \right\|_\ell \leq 2 \sup_{x \in \mathcal{X}} \|x\|_F \sup_{x' \in \mathcal{X}} \|x'\|_\infty \sqrt{\frac{e \log(d+1)}{m}},
\]

where the last inequality follows from estimation (20). We complete the proof by combining the above estimation with Theorems 1 and 2.

We briefly discuss the case of the trace norm regularization, i.e., \(\|A\| = \|A\|_{tr}\). In this case, the dual norm of trace norm is the spectral norm defined, for any \(B \in \mathbb{S}^d\), by \(\|B\|_* = \max_{\ell \in \mathbb{N}_d} \sigma_\ell(B)\) where \(\{\sigma_\ell : \ell \in \mathbb{N}_d\}\) are the singular values of matrix \(B\).

Observe, for any \(u,v \in \mathbb{R}^d\), that \(\|uv^T\|_* = \|u\|_F \|v\|_F\). Hence, the constant \(X^* = \sup_{x,x' \in \mathcal{X}} \|x'x^T\|_* = \sup_{x \in \mathcal{X}} \|x\|_F^2\). In addition,

\[
\mathcal{R}_m = \mathbb{E}_{z,\sigma} \sup_{x \in \mathcal{X}} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_jx_jx^T \right\|_* = \mathbb{E}_{z,\sigma} \sup_{x \in \mathcal{X}} \left\| x \right\|_F \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_jx_j \right\|_F = \sup_{x \in \mathcal{X}} \|x\|_F \mathbb{E}_{z,\sigma} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_jx_j \right\|_F \leq \sup_{x \in \mathcal{X}} \|x\|_F \mathbb{E}_{z,\sigma} \left( \sum_{j \in \mathbb{N}_m} \|x_j\|_F^2 \right)^{1/2} / m.
\]

Indeed, the above estimation for \(\mathcal{R}_m\) is optimal. To see this, we observe from (De La Peña & Giné, 1999, Theorem 1.3.2) that

\[
\left( \mathbb{E}_{\sigma} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_jx_j \right\|_F^2 \right)^{1/2} \leq \sqrt{2} \mathbb{E}_{\sigma} \left\| \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_jx_j \right\|_F.
\]

25
Combining the above fact with (23), we can obtain

\[ R_m = \sup_{x \in \mathcal{X}} \| x \|_F E_{z, \sigma} \left( \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j \right)_F \]

\[ \geq \frac{1}{\sqrt{2}} \sup_{x \in \mathcal{X}} \| x \|_F E_{z, \sigma} \left( \frac{1}{m} \sum_{j \in \mathbb{N}_m} \sigma_j y_j x_j \right)_F^2 \]

\[ = \frac{1}{m \sqrt{2}} \sup_{x \in \mathcal{X}} \| x \|_F E \left( \sum_{j \in \mathbb{N}_m} \| x_j \|_F^2 \right)^{\frac{1}{2}}. \]

Hence, the estimation (23) for \( R_m \) is optimal up to the constant \( \frac{1}{\sqrt{2}} \). Furthermore, ignoring further estimation \( \frac{1}{m} E \left( \sum_{j \in \mathbb{N}_m} \| x_j \|_F^2 \right)^{\frac{1}{2}} \leq \frac{1}{\sqrt{m}} \sup_{x \in \mathcal{X}} \| x \|_F \), the above estimations mean that the estimation for \( R_m \) in the case of trace-norm regularization are the same as the estimation (21) for the Frobenius norm regularization. Consequently, the generalization bounds for similarity learning and the relationship between similarity learning and the linear SVM are the same as those stated in Example 2. It is a bit disappointing that there is no improvement when using the trace norm. The possible reason is that the spectral norm of \( B \) and the Frobenius norm of \( B \) are the same when \( B \) takes the form \( B = xy^T \) for any \( x, y \in \mathbb{R}^d \).

We end this section with a comment on an alternate way to estimate the Rademacher average \( R_m \). Kakade (Kakade et al., 2008, 2012) developed elegant techniques for estimating Rademacher averages for linear predictors. In particular, the following theorem was established:

**Theorem 5.** (Kakade et al., 2008, 2012) Let \( \mathcal{W} \) be a closed convex set and let \( f : \mathcal{W} \to \mathbb{R} \) be a \( \beta \)-strongly convex with respect to \( \| \cdot \| \) and assume that \( f^*(0) = 0 \). Assume \( \mathcal{W} \subseteq \{ w : f(w) \leq f_{\text{max}} \} \). Furthermore, let \( \mathcal{X} = \{ x : \| x \|_* \leq X \} \) and \( \mathcal{F} = \{ w \to \langle w, x \rangle : w \in \mathcal{W}, x \in \mathcal{X} \} \). Then, we have

\[ R_n(\mathcal{F}) \leq X \sqrt{\frac{2f_{\text{max}}}{\beta n}}. \]
To apply Theorem 5, we rewrite the Rademacher average $\mathcal{R}_m$ as

$$
\mathcal{R}_m = \mathbb{E}_{z,\sigma} \left[ \sup_{\tilde{x} \in \mathcal{X}} \left\| \frac{1}{m} \sum_{i \in N_m} \sigma_i y_i x_i \tilde{x}^T \right\|_2 \right]
= \mathbb{E}_{z,\sigma} \left[ \sup_{\tilde{x} \in \mathcal{X}} \sup_{\|A\| \leq 1, A \in S^d} \left\langle \frac{1}{m} \sum_{i \in N_m} \sigma_i y_i x_i \tilde{x}^T, A \right\rangle \right] 
= \mathbb{E}_{z,\sigma} \left[ \sup_{\tilde{x} \in \mathcal{X}} \sup_{\|A\| \leq 1, A \in S^d} \left\langle \frac{1}{m} \sum_{i \in N_m} \sigma_i y_i x_i, A \tilde{x} \right\rangle \right].
$$

(24)

Now let $\mathcal{W} := \{ w \mapsto \langle w, x \rangle, w = A\tilde{x}, \|A\| \leq 1, A \in S^d \}$. Let us consider the sparse $L^1$-norm defined, for any $A \in S^d$, by

$$
\|A\|_1 = \sum_{k,\ell \in \mathbb{N}_d} |A_{k\ell}|.
$$

In this case, we observe that

$$
\|w\|_1 \leq \|A\|_1 \|\tilde{x}\|_\infty \leq \|A\|_1 \sup_{x \in \mathcal{X}} \|x\|_\infty \leq \sup_{x \in \mathcal{X}} \|x\|_\infty.
$$

Let $f(w) = \|w\|_q^2$ with $q = \frac{\log d}{\log d - 1}$ which is $(\frac{1}{\log d})$-strongly convex with respect to the norm $\|\cdot\|_1$. Then, for any $w \in \mathcal{W}$, we have that $\|w\|_q^2 \leq \|w\|_1^2 \leq \sup_{x \in \mathcal{X}} \|x\|_\infty^2$. Combining these observations with (24) allows us to obtain the estimation $\mathcal{R}_m \leq \sup_{x \in \mathcal{X}} \|x\|_\infty^2 \sqrt{\frac{2\log d}{m}}$.

Similarly, for the $(2, 1)$-mixed norm $\|A\|_{(2,1)} = \sum_{k \in \mathbb{N}_d} (\sum_{\ell \in \mathbb{N}_d} |A_{k\ell}|^2)^{1/2}$, observe that

$$
\|w\|_1 \leq \|A\|_{(2,1)} \|\tilde{x}\|_F \leq \sup_{x \in \mathcal{X}} \|x\|_F.
$$

Applying Theorem 5 with $f(w) = \|w\|_q^2 (q = \frac{\log d}{\log d - 1})$ again, we will have the estimation $\mathcal{R}_m \leq \sup_{x \in \mathcal{X}} \|x\|_F \sup_{x \in \mathcal{X}} \|x\|_\infty \sqrt{\frac{2\log d}{m}}$.

Hence, the estimations for the above two cases are similar to our estimations in the above examples. Our estimation is more straightforward by directly using the Khinchin-type inequality in contrast to the advanced convex-analysis techniques used in (Kakade et al., 2008, 2012).

However, for the case of trace-norm regularization (i.e., $\|A\| = \|A\|_{tr}$), one would expect, using the techniques in (Kakade et al., 2008, 2012), that the estimation for $\mathcal{R}_m$ is the same as that in the case for the sparse $L^1$-norm. The main hurdle for such result is the estimation of $\|w\|_1 = \|A\tilde{x}\|_1$ by the trace-norm of $A$. Indeed, by the discussion following our estimation (23) directly using Khinchin-type inequality, we know that
our estimation (23) is optimal. Hence, one can not expect the estimation for $\mathcal{R}_m$ for the case for trace-norm regularization is the same as that in the case for sparse $L^1$-norm regularization in our particular case of similarity learning formulation (2).

We end this section with an open question. It is not clear to us on how to establish a generic result for estimating the interesting Rademacher average $\mathcal{R}_m$ given by equation (24). Such a generic result is expected to be very similar to the result stated above as Theorem 5, which was established by Kakade et al. (Kakade et al., 2008, 2012). The main advantage of establishing such a generic result would enable a unifying estimation of $\mathcal{R}_m$ for different matrix-norms, which can then be instantiated into Examples 1, 2 and 3.

7 Conclusion

In this paper, we considered a regularized similarity learning formulation (2). Its generalization bounds were established for various matrix-norm regularization terms such as the Frobenius norm, sparse $L^1$-norm, and mixed $(2,1)$-norm. We proved the generalization error of the linear classifier based on the learnt similarity function can be bounded by the derived generalization bound of similarity learning. This guarantees the goodness of the generalization of similarity learning (2) with general matrix-norm regularization and thus the classification generalization of the resulting linear classifier.

Our techniques using the Rademacher complexity (Bartlett & Mendelson, 2002) and the important Khinchin-type inequality for the Rademacher variables allow us to obtain new bounds for similarity learning with general matrix-norm regularization terms.
There are several possible directions for future work. Firstly, we may consider similarity algorithms with general loss functions. It is expected that under some convexity conditions on the loss functions, better results could be obtained. Secondly, we usually focus on the excess misclassification error when considering classification problems. Hence, in the future, we would like to consider the theoretical link between the generalization bounds of the similarity learning and the excess misclassification error of the classifier built from the learnt similarity function.

Acknowledgments

We are grateful to the referees for their invaluable comments and suggestions on this paper. This work was supported by the EPSRC under grant EP/J001384/1. The corresponding author is Yiming Ying.

References


Appendix

In this appendix, the following facts are used for establishing generalization bounds in section 4 and section 5.

**Definition 3.** We say the function $f: \prod_{k=1}^{m} \Omega_k \rightarrow \mathbb{R}$ with bounded differences $\{c_k\}_{k=1}^{m}$ if, for all $1 \leq k \leq m$,

$$\max_{z_1, \ldots, z_{k-1}, z_k, z_{k+1}, \ldots, z_m} \left| f(z_1, \ldots, z_{k-1}, z_k, z_{k+1}, \ldots, z_m) - f(z_1, \ldots, z_{k-1}, z'_k, z_{k+1}, \ldots, z_m) \right| \leq c_k$$

**Lemma 1.** (McDiarmid’s inequality (McDiarmid, 1989)) Suppose $f: \prod_{k=1}^{m} \Omega_k \rightarrow \mathbb{R}$ with bounded differences $\{c_k\}_{k=1}^{m}$ then, for all $\epsilon > 0$, there holds

$$\Pr_{\mathbf{z}} \left\{ f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \geq \epsilon \right\} \leq e^{-\frac{2\epsilon^2}{\sum_{k=1}^{m} c_k^2}}.$$  

We need the following contraction property of the Rademacher averages which is essentially implied by Theorem 4.12 in Ledoux and Talagrand (Ledoux & Talagrand, 1991), see also (Bartlett & Mendelson, 2002; Koltchinskii & Panchenko, 2002).

**Lemma 2.** Let $F$ be a class of uniformly bounded real-valued functions on $(\Omega, \mu)$ and $m \in \mathbb{N}$. If for each $i \in \{1, \ldots, m\}$, $\phi_i: \mathbb{R} \rightarrow \mathbb{R}$ is a function having a Lipschitz constant $c_i$, then for any $\{x_i\}_{i \in \mathbb{N}_m}$,

$$\mathbb{E}_\epsilon \left( \sup_{f \in F} \sum_{i \in \mathbb{N}_m} \epsilon_i \phi_i(f(x_i)) \right) \leq 2 \mathbb{E}_\epsilon \left( \sup_{f \in F} \sum_{i \in \mathbb{N}_m} c_i \epsilon_i f(x_i) \right).$$  

(25)

Another important property of the Rademacher average which is used in the proof of the generalization bounds of the similarity learning is the following Khinchin-type inequality (see, e.g., De La Peña & Giné, 1999, Theorem 3.2.2).
Lemma 3. For $n \in \mathbb{N}$, let $\{f_i \in \mathbb{R} : i \in \mathbb{N}_n\}$, and $\{\sigma_i : i \in \mathbb{N}_n\}$ be a family of i.i.d. Rademacher random variables. Then, for any $1 < p < q < \infty$ we have

$$\left(\mathbb{E}_{\sigma} \left| \sum_{i \in \mathbb{N}_n} \sigma_i f_i \right|^q \right)^{\frac{1}{q}} \leq \left( \frac{q - 1}{p - 1} \right)^{\frac{1}{2}} \left( \mathbb{E}_{\sigma} \left| \sum_{i \in \mathbb{N}_n} \sigma_i f_i \right|^p \right)^{\frac{1}{p}}.$$