Robustness and Generalization for Metric Learning*

Aurélien Bellet Amaury Habrard
Laboratoire Hubert Curien UMR CNRS 5516,
Université Jean Monnet, 42000 Saint-Etienne, France
{aurelien.bellet,amaury.habrard}@univ-st-etienne.fr

Abstract

Metric learning has attracted a lot of interest over the last decade, but little work has been done about the generalization ability of such methods. In this paper, we address this issue by proposing an adaptation of the notion of algorithmic robustness, previously introduced by Xu and Mannor in classic supervised learning, to derive generalization bounds for metric learning. We also show that a weak notion of robustness is a necessary and sufficient condition to generalize, justifying that it is fundamental for metric learning. We provide some illustrative examples of our approach on a large class of existing algorithms. **Keywords:** Metric learning, Algorithmic robustness, Generalization bounds.

1 Introduction

The past ten years have seen a growing interest in supervised metric learning. Indeed, the relevance of a distance or a similarity, for a given task, is of crucial importance to the effectiveness of many classification or clustering methods. For this reason, a lot of research has been devoted to automatically learning distances or similarities from supervised data. Existing approaches rely on the fairly reasonable principle that, according to a good metric, pairs of examples with the same (resp. different) labels must be close to each other (resp. far away). Learning thus generally consists in finding the best parameters of the metric function given a set of labeled pairs. The most classic and commonly used approach in the literature focuses on Mahalanobis distance learning where the objective is to learn a positive semi-definite (PSD) matrix inducing a linear projection of the data where the Euclidean distance performs well. Other approaches have also considered arbitrary similarity functions with no PSD constraint. The learned distance or similarity is then typically used to improve the performance of nearest-neighbor methods.

From a theoretical standpoint, many papers have studied the convergence rate of the optimization problem used to learn the parameters of the metric. However and somewhat surprisingly, few studies have been done about the generalization ability of learned metrics on unseen data. This situation can be explained by the fact that one cannot assume that the learning pairs provided to a metric learning algorithm are independent and identically distributed (IID). Indeed, these pairs are generally given by an expert and/or extracted from a sample of individual instances. For example, common procedures for building such learning pairs are based either on the $k$ nearest or farthest neighbors of each example, some criterion of diversity, taking all the possible pairs or drawing pairs randomly from a learning sample. Online methods nevertheless offer guarantees, but only in the form of regret bounds assessing the deviation between the cumulative loss suffered by the online algorithm and the loss induced by the best hypothesis that can be chosen in hindsight. Apart from these results, as far as we know, very few papers have proposed a theoretical study on the generalization ability of supervised metric learning methods. The approach of Bian and Tao uses a statistical analysis to give generalization guarantees for loss minimization methods, but their results assume some hypotheses on the distribution of the examples and do not take into account any regularization on the metric. The most general contribution has been

*We would like to acknowledge support from the ANR LAMPADA 09-EMER-007-02 project and the PASCAL 2 Network of Excellence.

1 These pairs are sometimes replaced by triplets $(x, y, z)$ such that example $x$ must be closer to example $y$ than to example $z$, where $x$ and $y$ share the same label and $z$ does not.
proposed by Jin et al. [14] who adapted the framework of uniform stability [15] to regularized metric learning. However, their approach is based on a Frobenius norm regularizer and cannot be applied to any type of regularization, in particular sparsity-inducing norms [16].

In this paper, we propose to address this lack of theoretical framework by studying the generalization ability of metric learning algorithms according to a notion of *algorithmic robustness*. Algorithmic robustness, introduced by Xu et al. [17, 18], allows one to derive generalization bounds when given two “close” training and testing examples the variation between their associated loss is bounded. This notion of closeness examples relies on a partition of the input space into different regions such that two examples in the same region are said close. This framework has been successfully used in the classic supervised learning setting for deriving generalization bounds for SVM, Lasso and more. We propose here to adapt this notion of algorithmic robustness to metric learning that works both for similarity and distance learning. We show that, in this context, the problem of non-IIDness of the learning pairs can be worked around by simply assuming that the pairs are built from an IID sample of labeled examples. Moreover, following the work of Xu et al. [18], we provide a notion of weak robustness that is necessary and sufficient for metric learning algorithms to generalize well, highlighting that robustness is a fundamental property. We illustrate the applicability of our framework by deriving generalization bounds, using very few approach-specific arguments, for a larger class of problems than Jin et al. that can accommodate a vast choice of regularizers, without any assumption on the distribution of the examples.

The rest of the paper is organized as follows. We introduce some preliminaries and notations in Section 2. Our notion of algorithmic robustness for metric learning is presented in Section 3. The necessity and sufficiency of weak robustness is shown in Section 4. Section 5 is devoted to the illustration of our framework to actual metric learning algorithms. Finally, we conclude in Section 6.

## 2 Preliminaries

### 2.1 Notations

Let $X$ be the instance space, $Y$ a finite label set and let $Z = X \times Y$. In the following, $z = (x, y) \in Z$ means $x \in X$ and $y \in Y$. Let $\mu$ be an unknown probability distribution over $Z$. We assume that $X$ is a compact convex metric space w.r.t. a norm $\| \cdot \|$ such that $X \subset \mathbb{R}^d$, thus there exists a constant $R$ such that $\forall x \in X, \|x\| \leq R$. A similarity or distance function is a pairwise function $f : X \times X \to \mathbb{R}$. In the following, we use the generic term metric to refer to either a similarity or a distance function. We denote by $s$ a labeled training sample consisting of $n$ training instances $(s_1, \ldots, s_n)$ drawn IID from $\mu$. The sample of all possible pairs built from $s$ is denoted by $p_k$ such that $p_k = \{(s_1, s_1), \ldots, (s_1, s_n), \ldots, (s_n, s_n)\}$. A metric learning algorithm $\mathcal{A}$ takes as input a finite set of pairs from $(Z \times Z)^n$ and outputs a metric. We denote by $\mathcal{A}_{p_k}$ the metric learned by an algorithm $\mathcal{A}$ from a sample $p_k$ of pairs. For any pair of labeled examples $(z, z')$ and any metric $f$, we associate a loss function $l(f, z, z')$ which depends on the examples and their labels. This loss is assumed to be nonnegative and uniformly bounded by a constant $B$. We define the true generalization loss over $\mu$ by $\mathcal{L}(f) = \mathbb{E}_{z, z' \sim \mu}l(f, z, z')$. We denote the empirical loss over the sample $p_k$ by $l_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} l(f, s_i, s_j) = \frac{1}{n} \sum_{(s_i, s_j) \in p_k} l(f, s_i, s_j)$.

### 2.2 Robustness for classical supervised learning

The notion of algorithmic robustness, introduced by Xu and Mannor [17, 18] in the context of classic supervised learning, is based on the deviation between the losses associated to two training and testing instances that are close. An algorithm is said $(K, \epsilon(s))$-robust if there exists a partition of the space $Z = X \times Y$ into $K$ disjoint subsets such that for every learning and testing instances belonging to the same region of the partition, the deviation between their associated losses is bounded by a term $\epsilon(s)$. From this definition, the authors have proved a convergence bound for the difference between the empirical and true losses of the form $\epsilon(s) + B \sqrt{\frac{2K \ln 2 + 2 \ln 1/\delta}{n}}$ (with probability $1 - \delta$). This bound depends on $K$ and $\epsilon(s)$ which can be made as small as desired by refining this partition. When considering metric spaces, the partition of $Z$ can be obtained by the notion of covering number [19].


Thus, we present here our adaptation of robustness to metric learning. The idea is to use the partition of a metric learning definition proposed in this work, for any pair \((z, z')\) and a training pair \((z_1, z_2)\), if \(z, z_1\) belong to some region \(C_i\) and \(z', z_2\) to some region \(C_j\), then the deviation between the loss of these two pairs must be bounded.

**Definition 1** For a metric space \((X, \rho)\), and \(T \subseteq X\), we say that \(\hat{T} \subseteq T\) is a \(\gamma\)-cover of \(T\), if \(\forall t \in T\), \(\exists \hat{t} \in \hat{T}\) such that \(\rho(t, \hat{t}) \leq \gamma\). The \(\gamma\)-covering number of \(T\) is

\[
N(\gamma, T, \rho) = \min\{|\hat{T}| : \hat{T} \text{ is a } \gamma - \text{cover of } T\}.
\]

For example, when \(X\) is a compact convex space, for any \(\gamma > 0\), the quantity \(N(\gamma, X, \rho)\) is finite leading to a finite cover. If we consider the space \(Z\), we can note that the label set can be partitioned into \(|Y|\) subsets such that if two instances \(z_1 = (x_1, y_1), \: \: z_2 = (x_2, y_2)\) belong to the same subset, then \(y_1 = y_2\) and \(\rho(x_1, x_2) \leq \gamma\).

### 3 Robustness and Generalization for Metric Learning

We present here our adaptation of robustness to metric learning. The idea is to use the partition of \(Z\) as the corresponding instance of the other pair, as shown in Figure 1. A metric learning algorithm with this property is said robust. This notion is formalized as follows.

**Definition 2** An algorithm \(A\) is \((K, \epsilon(\cdot))\) robust for \(K \in \mathbb{N}\) and \(\epsilon(\cdot) : (Z \times Z)^K \to \mathbb{R}\) if \(Z\) can be partitioned into \(K\) disjoint sets, denoted by \(\{C_i\}_{i=1}^K\), such that for all sample \(s \in Z^n\) and the pair set \(p(s)\) associated to this sample, the following holds:

\[
\forall(s_1, s_2) \in p(s), \forall z_1, z_2 \in Z, \forall i, j = 1, \ldots, K : \text{if } s_1, z_1 \in C_i \text{ and } s_2, z_2 \in C_j \text{ then } \\
|l(A_{p_s}, s_1, s_2) - l(A_{p_s}, z_1, z_2)| \leq \epsilon(p_s).
\]

(1)

\(K\) and \(\epsilon(\cdot)\) quantify the robustness of the algorithm which depends on the learning sample. The property of robustness is required for every training pair of the sample; we will see later that this property can be relaxed.

Note that this definition of robustness can be easily extended to triplet based metric learning algorithms. Instead of considering all the pairs \(p_s\) from an IID sample \(s\), we take the admissible triplet set \(trip_s\) of \(s\) such that \((s_1, s_2, s_3) \in trip_s\) means \(s_1\) and \(s_2\) share the same label while \(s_1\) and \(s_3\) have different ones, with the interpretation that \(s_1\) must be more similar to \(s_2\) than to \(s_3\). The robustness property can then be expressed by: \(\forall(s_1, s_2, s_3) \in trip_s, \forall z_1, z_2, z_3 \in Z, \forall i, j, l = 1, \ldots, K : \text{if } s_1, z_1 \in C_i, \: s_2, z_2 \in C_j\) and \(s_3, z_3 \in C_l\) then

\[
|l(A_{trip_s}, s_1, s_2, s_3) - l(A_{trip_s}, z_1, z_2, z_3)| \leq \epsilon(trip_s).
\]

(2)
3.1 Generalization of robust algorithms

We now give our first result on the generalization of metric learning algorithms.

Proposition 1 \((\S\text{20})\) Let \(|N_1|, \ldots, |N_K|\) an IID multinomial random variable with parameters \(n\) and \((\mu(C_1), \ldots, \mu(C_K))\). By the Breteganolle-Huber-Carol inequality we have: 
\[
\Pr \left\{ |\sum_{i=1}^{K} \left| \frac{N_i}{n} - \mu(C_i) \right| \geq \lambda \right\} \leq 2^K \exp \left( \frac{-\lambda^2}{2} \right), \text{ hence with probability at least } 1 - \delta,
\]
\[
\sum_{i=1}^{K} \left| \frac{N_i}{n} - \mu(C_i) \right| \leq \sqrt{\frac{2K \ln 2 + 2 \ln(1/\delta)}{n}}. \tag{3}
\]

We now give our first result on the generalization of metric learning algorithms.

Theorem 1 \(\text{If a learning algorithm } A \text{ is } (K, \epsilon(\cdot))-\text{robust and the training sample is made of the pairs } p_n \text{ obtained from a sample } s \text{ generated by n IID draws from } \mu, \text{ then for any } \delta > 0, \text{ with probability at least } 1 - \delta \text{ we have:}
\]
\[
|\mathcal{L}(A_{p_n}) - l_{\text{emp}}(A_{p_n})| \leq \epsilon(p_n) + 2B \sqrt{\frac{2K \ln 2 + 2 \ln 1/\delta}{n}}.
\]

Proof Let \(N_i\) be the set of index of points of \(s\) that fall into the \(C_i\). \(|N_1|, \ldots, |N_K|\) is a IID random variable with parameters \(n\) and \((\mu(C_1), \ldots, \mu(C_K))\). We have:
\[
|\mathcal{L}(A_{p_n}) - l_{\text{emp}}(A_{p_n})|
= \left| \sum_{i=1}^{K} \sum_{j=1}^{K} E_{z_1, z_2 \sim \mu}(l(A_{p_n}, z_1, z_2) | z_1 \in C_i, z_2 \in C_j) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} l(A_{p_n}, s_i, s_j) \right|
\leq \left| \sum_{i=1}^{K} \sum_{j=1}^{K} E_{z_1, z_2 \sim \mu}(l(A_{p_n}, z_1, z_2) | z_1 \in C_i, z_2 \in C_j) \mu(C_i) \frac{|N_j|}{n} \right| +
\left| \sum_{i=1}^{K} \sum_{j=1}^{K} E_{z_1, z_2 \sim \mu}(l(A_{p_n}, z_1, z_2) | z_1 \in C_i, z_2 \in C_j) \mu(C_i) \frac{|N_j|}{n} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} l(A_{p_n}, s_i, s_j) \right|
\leq (a) \left| \sum_{i=1}^{K} \sum_{j=1}^{K} E_{z_1, z_2 \sim \mu}(l(A_{p_n}, z_1, z_2) | z_1 \in C_i, z_2 \in C_j) \mu(C_i) \frac{|N_j|}{n} \right| +
\left| \sum_{i=1}^{K} \sum_{j=1}^{K} E_{z_1, z_2 \sim \mu}(l(A_{p_n}, z_1, z_2) | z_1 \in C_i, z_2 \in C_j) \mu(C_i) \frac{|N_j|}{n} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} l(A_{p_n}, s_i, s_j) \right|
\leq (b) \left| \sum_{i=1}^{K} \sum_{j=1}^{K} E_{z_1, z_2 \sim \mu}(l(A_{p_n}, z_1, z_2) | z_1 \in C_i, z_2 \in C_j) \mu(C_i) \frac{|N_j|}{n} \right| +
\left| \sum_{i=1}^{K} \sum_{j=1}^{K} E_{z_1, z_2 \sim \mu}(l(A_{p_n}, z_1, z_2) | z_1 \in C_i, z_2 \in C_j) \mu(C_i) \frac{|N_j|}{n} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} l(A_{p_n}, s_i, s_j) \right|
If a learning algorithm theorem illustrates the generalization guarantees associated to the pseudo-robustness property. The previous study requires the robustness property to be true for every learning pair. We show, with the following definition, that it is possible to relax the robustness to be true for only a subpart of the sample. Moreover, it also gives an insight into the behavior of metric learning approaches aiming at learning a distance to be plugged in a k-nearest neighbor classifier such as LMNN [3]. These methods do not optimize the distance according to all possible pairs, but only according to the nearest-neighbors of the same class.

\[
\sum_{i=1}^{K} \mu(C_i) - \frac{|N_j|}{n} + \frac{2}{n^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{s_i \in N_i, s_j \in N_j} \max \max_{z \in C_i, z' \in C_j} |l(A_{p_a}, z, z') - l(A_{p_a}, s_i, s_j)| + 2B \sqrt{\frac{2K \ln 2 + 2 \ln 1/\delta}{n}}.
\]

Inequalities (a) and (b) are due to the triangle inequality, (c) uses the fact that \( l \) is bounded by \( B \), that \( \sum_{i=1}^{K} \mu(C_i) = 1 \) by definition of a multinomial random variable and that \( \sum_{j=1}^{K} \frac{|N_j|}{n} = 1 \) by definition of the \( N_j \). Lastly, (d) is due to the hypothesis of robustness (Equation 1) and (e) to the application of Proposition [1].

The previous bound depends on \( K \) which is given by the cover chosen for \( Z \). If for any \( K \), the associated \( \epsilon(\cdot) \) is a constant (i.e. \( \epsilon_K(s) = \epsilon_K \)) for any \( s \), we can prove a bound holding uniformly for all \( K \):

\[
|\mathcal{L}(A_{p_a}) - \mathcal{L}_{emp}(A_{p_a})| \leq \inf_{K \geq 1} \left( \epsilon_K + 2B \sqrt{\frac{2K \ln 2 + 2 \ln 1/\delta}{n}} \right).
\]

This also gives an insight into the objective of any robust algorithm: according to a partition of the labeled input space, given two regions, minimize the maximum loss over pairs of examples belonging to each region.

For triplet based metric learning algorithms, by following the definition of robustness given by Equation [2] and adapting straightforward the losses to triplets such that they output zero for non-admissible ones, Theorem [1] can be easily extended to obtain the following generalization bound:

\[
|\mathcal{L}(A_{trip}) - \mathcal{L}_{emp}(A_{trip})| \leq \epsilon(trip_a) + 3B \sqrt{\frac{2K \ln 2 + 2 \ln 1/\delta}{n}}.
\]

### 3.2 Pseudo-robustness

The previous study requires the robustness property to be true for every learning pair. We show, with the following definition, that it is possible to relax the robustness to be true for only a subpart of the sample and yet be able to derive generalization guarantees.

**Definition 3** An algorithm \( A \) is \((K, \epsilon(\cdot), \nu(\cdot))\) pseudo robust for \( K \in \mathbb{N} \), \( \epsilon(\cdot) : (Z \times Z)^n \to \mathbb{R} \) and \( \nu(\cdot) : (Z \times Z)^n \to \{1, \ldots, n^2\} \), if \( Z \) can be partitioned into \( K \) disjoint sets, denoted by \( \{C_i\}_{i=1}^{K} \), such that for all \( s \in Z^n \) IID from \( \mu \), there exists a subset of training samples \( p_a \subseteq p_a \), with \( \nu(p_a) = \nu(p_a) \), such that the following holds:

\[
\forall(s_1, s_2) \in \hat{p}_a, \forall z_1, z_2 \in Z, \forall i, j = 1, \ldots, K: \text{if } s_1, z_1 \in C_i \text{ and } s_2, z_2 \in C_j \text{ then } |l(A_{p_a}, s_1, s_2) - l(A_{p_a}, z_1, z_2)| \leq \epsilon(p_a).
\]

We can easily observe that \((K, \epsilon(\cdot), \nu(\cdot))\)-robust is equivalent to \((K, \epsilon(\cdot), n^2)\) pseudo-robust. The following theorem illustrates the generalization guarantees associated to the pseudo-robustness property.

**Theorem 2** If a learning algorithm \( A \) is \((K, \epsilon(\cdot), \nu(\cdot))\) pseudo-robust, the training pairs \( p_a \) come from a sample generated by \( n \) IID draws from \( \mu \), then for any \( \delta > 0 \), with probability at least \( 1 - \delta \) we have:

\[
|\mathcal{L}(A_{p_a}) - \mathcal{L}_{emp}(A_{p_a})| \leq \frac{\nu(p_a)}{n^2} \epsilon(p_a) + B \frac{n^2 - \nu(p_a)}{n^2} + 2 \sqrt{\frac{2K \ln 2 + 2 \ln 1/\delta}{n}}.
\]

The proof is similar to that of Theorem [1] and is given in Appendix [A.1].
and some pairs of different class. According to the previous theorem, this principle is founded provided that the robustness property is fulfilled for some of the pairs used to optimize the metric. Finally, note that this notion of pseudo-robustness can be also easily adapted to triplet based metric learning.

4 Necessity of Robustness

We prove here that a notion of weak robustness is actually necessary and sufficient to generalize in a metric learning setup. This result is based on an asymptotic analysis following the work of Xu and Mannor [18].

We consider pairs of instances coming from an increasing sample of training instances \( s = (s_1, s_2, \ldots) \) and from a sample of test instances \( t = (t_1, t_2, \ldots) \) such that both samples are assumed to be drawn IID from a distribution \( \mu \). We use \( s(n) \) and \( t(n) \) to denote the first \( n \) examples of the two samples respectively, while \( s^* \) denotes a fixed sequence of examples.

We use \( L(f, p_{s^*}) = \frac{1}{n} \sum_{(s_i, s_j) \in P_{s^*}} l(f, s_i, s_j) \) to refer to the average loss given a set of pairs for any learned metric \( f \), and \( \mathcal{L}(f) = \mathbb{E}_{z, z' \sim \mu} l(f, z, z') \) for the expected loss.

We first define a notion of generalizability for metric learning.

**Definition 4** 1. Given a training pair set \( p_{s^*} \) coming from a sequence of examples \( s^* \), a metric learning method \( \mathcal{A} \) generalizes w.r.t. \( p_{s^*} \) if \( \lim_n \mathcal{L}(\mathcal{A}_{p_{s^*}(n)}) - \mathcal{L}(\mathcal{A}_{p_{s^*}(n)}, p_{s^*}(n)) = 0 \).

2. A learning method \( \mathcal{A} \) generalizes with probability 1 if it generalizes with respect to the pairs \( p_s \) of almost all samples \( s \) IID from \( \mu \).

Note this notion of generalizability implies convergence in mean. We then introduce the notion of weak robustness for metric learning.

**Definition 5** 1. Given a set of training pairs \( p_{s^*} \) coming from a sequence of examples \( s^* \), a metric learning method \( \mathcal{A} \) is weakly robust with respect to \( p_{s^*} \) if there exists a sequence of \( \{D_n \subseteq \mathbb{Z}^n\} \) such that \( \mathbb{P}(t(n) \in D_n) \to 1 \) and

\[
\lim_n \max_{\tilde{s}(n) \in D_n} \left| L(\mathcal{A}_{p_{s^*}(n)}, \tilde{s}(n)) - L(\mathcal{A}_{p_{s^*}(n)}, p_{s^*}(n)) \right| = 0.
\]

2. A learning method \( \mathcal{A} \) is almost surely weakly robust if it is robust w.r.t. almost all \( s \).

The definition of robustness requires the labeled sample space to be partitioned into disjoints subsets such that some instances of pairs of train/test examples belong to the same partition, then they have similar loss. Weak robustness is a generalization of this notion where we consider the average loss of testing and training pairs: for a large (in the probabilistic sense) subset of data, the testing loss is close to the training loss, then the algorithm is weakly robust. From Proposition 1 we can see that if for any fixed \( \epsilon > 0 \) there exists \( K \) such that an algorithm \( \mathcal{A} \) is \( (K, \epsilon) \) robust, then \( \mathcal{A} \) is weakly robust. We now give the main result of this section about the necessity of robustness.

**Theorem 3** Given a fixed sequence of training examples \( s^* \), a metric learning method \( \mathcal{A} \) generalizes w.r.t. \( p_{s^*} \) if and only if it is weakly robust w.r.t. \( p_{s^*} \).

**Proof** Following [18], the sufficiency is obtained by the fact that the testing pairs are obtained from a sample \( t(n) \) constituted of \( n \) IID instances. We give the proof in Appendix A.2.

For the necessity, we need the following lemma which is a direct adaptation of a result introduced in [18] (Lemma 2). We provide the proof in Appendix A.3 for the sake of completeness.

**Lemma 1** Given \( s^* \), if a learning method is not weakly robust w.r.t. \( p_{s^*} \), there exists \( \epsilon^*, \delta^* > 0 \) such that the following holds for infinitely many \( n \):

\[
\mathbb{P}(|L(\mathcal{A}_{p_{s^*}(n)}, t(n)) - L(\mathcal{A}_{p_{s^*}(n)}, p_{s^*}(n))| \geq \epsilon^*) \geq \delta^*.
\]

Now, recall that \( l \) is positive and uniformly bounded by \( B \), thus by the McDiarmid inequality (recalled in Appendix A.3) we have that for any \( \epsilon, \delta > 0 \) there exists an index \( n^* \) such that for any \( n > n^* \), with
probability at least $1 - \delta$, we have $\frac{1}{n^2} \sum_{(t_i, t_j) \in p_{t(n)}} l(A_{p_{t(n)}}, t_i, t_j) - L(A_{p_{t(n)}}) \leq \epsilon$. This implies the convergence $L(A_{p_{t(n)}}, p_{t(n)}) - L(A_{p_{t(n)}}) \overset{p}{\rightarrow} 0$, and thus from a given index:

$$|L(A_{p_{t(n)}}, p_{t(n)}) - L(A_{p_{t(n)}})| \leq \frac{\epsilon^*}{2}. \quad (7)$$

Now, by contradiction, suppose algorithm $A$ is not weakly robust, Lemma 1 implies Equation 6 holds for infinitely many $n$. This combined with Equation 7 implies that for infinitely many $n$:

$$|L(A_{p_{t(n)}}, p_{t(n)}) - L(A_{p_{t(n)}}, p_{s(t(n)})| \geq \frac{\epsilon^*}{2}$$

which means $A$ does not generalize, thus the necessity of weak robustness is established.

The following corollary follows immediately from Theorem 3.

**Corollary 1** A metric learning method $A$ generalizes with probability 1 if and only if it is almost surely weakly robust.

5 Examples of Robust Metric Learning Algorithms

We first restrict our attention to Mahalanobis distance learning algorithms of the following form:

$$\min_{M \succeq 0} c\|M\| + \frac{1}{n^2} \sum_{(s_i, s_j) \in p_s} g(y_{ij}|1 - f(M, x_i, x_j)|), \quad (8)$$

where $s_i = (x_i, y_i), s_j = (x_j, y_j), y_{ij} = 1$ if $y_i = y_j$ and $-1$ otherwise, $f(M, x_i, x_j) = (x_i - x_j)^TM(x_i - x_j)$ is the Mahalanobis distance parameterized by the $d \times d$ PSD matrix $M$, $\|\cdot\|$ some matrix norm and $c$ a regularization parameter. The loss function $l(f, s_i, s_j) = g(y_{ij}|1 - f(M, x_i, x_j)|)$ outputs a small value when its input is large positive and a large value when it is large negative. We assume $g$ to be nonnegative and Lipschitz continuous with Lipschitz constant $U$. Lastly, $g_0 = \sup_{s_i, s_j} g(y_{ij}|1 - f(0, x_i, x_j)|)$ is the largest loss when $M$ is 0.

To prove the robustness of (8), we will need the following theorem, which essentially says that if a metric learning algorithm achieves approximately the same testing loss for testing pairs that are close to each other, then it is robust.

**Theorem 4** Fix $\gamma > 0$ and a metric $\rho$ of $Z$. Suppose $A$ satisfies

$$|l(A_{p_s}, z_1, z_2) - l(A_{p_s}, z_1', z_2')| \leq \epsilon(p_s), \quad \forall z_1, z_2, z_1', z_2' : z_1, z_2 \in s, \rho(z_1, z_1') \leq \gamma, \rho(z_2, z_2') \leq \gamma$$

and $N(\gamma/2, Z, \rho) < \infty$. Then $A$ is $(N(\gamma/2, Z, \rho), \epsilon(p_s))$-robust.

**Proof** By definition of covering number, we can partition $X$ in $N(\gamma/2, X, \rho)$ subsets such that each subset has a diameter less or equal to $\gamma$. Furthermore, since $Y$ is a finite set, we can partition $Z$ into $|Y|N(\gamma/2, X, \rho)$ subsets $\{C_i\}$ such that $z_1, z_1' \in C_i \Rightarrow \rho(z_1, z_1') \leq \gamma$. Therefore,

$$|l(A_{p_s}, z_1, z_2) - l(A_{p_s}, z_1', z_2')| \leq \epsilon(p_s), \quad \forall z_1, z_2, z_1', z_2' : z_1, z_2 \in s, \rho(z_1, z_1') \leq \gamma, \rho(z_2, z_2') \leq \gamma$$

implies $z_1, z_2 \in s, z_1', z_1' \in C_i, z_2, z_2' \in C_j \Rightarrow |l(A_{p_s}, z_1, z_2) - l(A_{p_s}, z_1', z_2')| \leq \epsilon(p_s)$, which establishes the theorem.

We now prove the robustness of (8) when $\|M\|$ is the Frobenius norm.

**Example 1 (Frobenius norm) Algorithm [5] with $\|M\| = \|M\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^d m_{ij}^2}$ is $(|Y|N(\gamma/2, X, \|\cdot\|_F), \epsilon(p_s))$-robust.

**Proof** Let $M^*$ be the solution given training data $p_s$. Thus, due to optimality of $M^*$, we have

$$c\|M^*\|_F + \frac{1}{n^2} \sum_{(s_i, s_j) \in p_s} g(y_{ij}|1 - f(M, x_i, x_j)|) \leq c\|0\|_F + \frac{1}{n^2} \sum_{(s_i, s_j) \in p_s} g(y_{ij}|1 - f(0, x_i, x_j)|) = g_0$$
and thus $\|M^*\|_F \leq g_0/c$. We can partition $\mathcal{Z}$ as $\{Y|N(\gamma/2, X, \|\cdot\|_2)\}$ sets, such that if $z$ and $z'$ belong to the same set, then $y = y'$ and $\|x - x'\|_2 \leq \gamma$. Now, for $z_1, z_2, z'_1, z'_2 \in \mathcal{Z}$, if $y_1 = y'_1$, $\|x_1 - x'_1\|_2 \leq \gamma$, $y_2 = y'_2$ and $\|x_2 - x'_2\|_2 \leq \gamma$, then:

$$g(y_{1,2}[1 - f(M^*, x_1, x_2)]) - g(y'_{1,2}[1 - f(M^*, x'_1, x'_2)]) \leq U'(x_1 - x_2)^T M^*(x_1 - x_2) - (x'_1 - x'_2)^T M^*(x'_1 - x'_2)$$

$$= U'(x_1 - x_2)^T M^*(x_1 - x_2) - (x'_1 - x'_2)^T M^*(x'_1 - x'_2)$$

$$= U'(x_1 - x_2)^T M^*(x_1 - x'_2 - x'_1 + x_2) + (x_1 - x_2 - (x'_1 + x'_2))^{T} M^*(x'_1 + x'_2)$$

$$\leq U'(x_1 - x_2)^T M^*(x_1 - x'_1) + (x_1 - x_2)^T M^*(x'_2 - x_2)$$

$$+ (x_1 - x'_2)^T M^*(x'_1 + x'_2) + (x_1 - x_2)^T M^*(x'_1 + x'_2)$$

$$\leq U'(x_1 - x_2)^T M^*F \|x_1 - x'_1\|_2 + \|x_1 - x_2\|_2 \|M^*F\|_F \|x'_2 - x_2\|_2$$

$$+ \|x_1 - x'_2\|_2 \|M^*F\|_F \|x'_1 - x'_2\|_2 \leq \frac{8UR\gamma g_0}{c}$$

Hence, the example holds by Theorem 3.

Note that for the special case of Example 1 a generalization bound (with same order of convergence rate) based on uniform stability was derived in [10]. However, it is known that sparse algorithms are not stable [10], and thus stability-based analysis fails to assess the generalization ability of recent sparse metric learning approaches [3, 7, 8]. The key advantage of robustness over stability is that it can accommodate arbitrary p-norms (or even any regularizer which is bounded below by some p-norm), thanks to the equivalence of norms. (To illustrate this, we show the robustness when $\|M\|$ is either the $\ell_1$ norm (used in [3, 7]) which promotes sparsity at the component level, or the $\ell_{2,1}$ norm (used in [8]), which is particularly interesting in the context of Mahalanobis distance learning since it induces group sparsity at the column/row level.)

The proofs are reminiscent of that of Example 1 and can be found in Appendices A.5 and A.6.

**Example 2 ($\ell_1$ norm)** Algorithm [8] with $\|M\|_1 = \|M\|_1$ is $(\|Y|N(\gamma, X, \|\cdot\|_1), \frac{8UR\gamma g_0}{c})$-robust.

**Example 3 ($\ell_{2,1}$ norm)** Consider Algorithm [8] with $\|M\|_2,1 = \sum_i \|m_i\|_2$, where $m_i$ is the i-th column of $M$. This algorithm is $(\|Y|N(\gamma, X, \|\cdot\|_2, \frac{8UR\gamma g_0}{c})$-robust.

Some metric learning algorithms have kernelized versions, for instance [11, 12]. In the following example we show robustness for a kernelized formulation.

**Example 4 (Kernelization)** Consider the kernelized version of Algorithm [8]:

$$\min_{M \geq 0} \frac{1}{c} \|M\|_{H} + \frac{1}{c} \sum_{(s_i, s_j) \in C_{ps}} g(y_{ij}[1 - f(M, \phi(x_i, \phi(x_j)))],$$

(9)

where $\phi(\cdot)$ is a feature mapping to a kernel space $H$, $\|\cdot\|_{H}$ the norm function of $H$ and $k(\cdot, \cdot)$ the kernel function. Consider a cover of $X$ by $\|\cdot\|_2$ (X being compact) and let $f_2(\gamma) \triangleq \max_{a,b \in X, 0 - \|a - b\|_2 \leq \gamma} (k(a, a) + k(b, b) - 2k(a, b))$ and $B_{\gamma} = \max_{x \in X} \sqrt{k(x, x)}$. If the kernel function is continuous, $B_{\gamma}$ and $f_2(\gamma)$ are finite for any $\gamma > 0$ and thus Algorithm [8] is $(\|Y|N(\gamma, X, \|\cdot\|_2, \frac{8UB\sqrt{Tg_0}}{c})$-robust.

The proof is given in Appendix A.7.

**Remark 1** We can easily prove similar results for other forms of metrics using the same analysis.

**Remark 2** Using triplet-based robustness (Equation 4), we can for instance show the robustness of two popular triplet-based metric learning approaches [1, 2] for which no generalization guarantees were known (to the best of our knowledge). These algorithms have the following form:

$$\min_{M \geq 0} \frac{1}{|\text{trip}_p|} \sum_{(s_i, s_j, s_k) \in \text{trip}_p} [1 - (x_i - x_k)^T M(x_i - x_k) + (x_i - x_j)^T M(x_i - x_j)]_+,$$

In this case, the linear projection space of the data induced by the learned Mahalanobis distance is of lower dimension than the original space, allowing more efficient computations and smaller storage size.
where $\|M\| = \|M\|_F$ in $\mathbb{R}^2$ and $\|M\| = \|M\|_{1,2}$ in $\mathbb{R}^4$. These methods are $(\mathcal{N}(\gamma, Z, \|\cdot\|_2), \text{16LogDet}_\gamma)$-robust (by using the same proof technique as in Examples 1 and 3). The additional factor 2 comes from the use of triplets instead of pairs.

6 Conclusion

We proposed a new theoretical framework for evaluating the generalization ability of metric learning based on the notion of algorithm robustness originally introduced in [18]. We showed that a weak notion of robustness characterizes the generalizability of metric learning algorithms, justifying that robustness is fundamental for such algorithms. This framework allows us to derive generalization bounds for a large class of algorithms with different regularizations, such as sparsity inducing norms, making the approach more powerful than the (few) existing frameworks. Moreover, almost no algorithm-specific argument is needed to derive these bounds, which explains why they are often similar. Natural perspectives arise when considering different settings. For example, some algorithms use both pair and triplet based information as input such as [6]. Other future work could include studying even more general loss functions and regularizers (such as the LogDet divergence used in [4, 5]), unsupervised/semi-supervised methods or domain adaptation. Being able to characterize the generalization ability of metric learning directly with the kind of classifier using the metric-like k-NN is also an interesting and challenging direction.

A Appendix

A.1 Proof of Theorem 2 (pseudo-robustness)

Proof From the proof of Theorem 1, we can easily deduce that:

$$|\mathcal{L}(A_{p_{\delta}}) - l_{emp}(A_{p_{\delta}})| \leq 2B \sum_{i=1}^{K} \left( |N_i| \frac{n}{N} - \mu(C_i) \right) +$$

$$\sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{E}_{z_1, z_2 \sim \mu}(l(A_{p_{\delta}}, z_1, z_2) z_1 \in C_i, z_2 \in C_j) \left( |N_i| \frac{N_j}{N} n - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} l(A_{p_{\delta}}, s_i, s_j) \right).$$

Then, we have

$$\leq 2B \sum_{i=1}^{K} \left( |N_i| \frac{n}{N} - \mu(C_i) \right) +$$

$$\frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{s_o, s_i} \sum_{z_1 \in C_i, z_2 \in C_j} \max \max_{z_1, z_2} |l(A_{p_{\delta}}, z, z') - l(A_{p_{\delta}}, s_o, s_i)| +$$

$$\frac{1}{n^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{s_o, s_i} \sum_{z_1 \in C_i, z_2 \in C_j} \max \max_{z_1, z_2} |l(A_{p_{\delta}}, z, z') - l(A_{p_{\delta}}, s_o, s_i)| \leq \frac{p_1(p_{\delta})}{n^2} \epsilon(p_{\delta}) + B \frac{\mu^2 - p_1(p_{\delta})}{n^2} + 2 \sqrt{2K \ln 2 + 2 \ln \frac{1}{\delta}}.$$
all $n > N(\delta, \epsilon)$, $Pr(t(n) \in D_n) > 1 - \delta$ and
\[
\max_{\hat{s}(n) \in D_n} |L(A_{p_{\hat{s}}(n)}, p_{\hat{s}(n)}) - L(A_{p_{\epsilon}}(n), p_{\epsilon}(n))| < \epsilon. \tag{10}
\]
Therefore for any $n > N(\delta, \epsilon)$,
\[
|L(A_{p_{\epsilon}}(n)) - L(A_{p_{\hat{s}}(n)}, p_{\hat{s}(n)})| = |E_{t(n)}(L(A_{p_{\epsilon}}(n), p_{t(n)})) - L(A_{p_{\epsilon}}(n), p_{\hat{s}(n)})| = |Pr(t(n) \notin D_n)E(L(A_{p_{\epsilon}}(n), p_{t(n)}))| + \delta B + \max_{\hat{s}(n) \in D_n} |L(A_{p_{\hat{s}}(n)}, p_{\hat{s}(n)}) - L(A_{p_{\epsilon}}(n), p_{\hat{s}(n)})|)
\leq \delta B + \max_{\hat{s}(n) \in D_n} |L(A_{p_{\hat{s}}(n)}, p_{\hat{s}(n)}) - L(A_{p_{\epsilon}}(n), p_{\hat{s}(n)})|)
\leq \epsilon.
\]
The first inequality holds because the testing samples $t(n)$ consists of $n$ instances IID from $\mu$. The second equality is obtained by conditional expectation. The next inequality uses the positiveness and the upper bound $B$ of the loss function. Finally, we apply Equation (10). We thus conclude that $A$ generalizes for $p_{\epsilon}$, because $\epsilon$ and $\delta$ can be chosen arbitrary.

A.3 Proof of Lemma 1

Proof This proof follows exactly the same principle as the proof of Lemma 2 from [13]. By contradiction, assume $\epsilon^*$ and $\delta^*$ do not exist. Let $\epsilon_v = \delta_v = 1/v$ for $v = 1, 2, ..., $ then there exists a non decreasing sequence $\{N(v)\}_{v=1}^\infty$ such that for all $v$, if $n \geq N(v)$ then $Pr([L(A_{p_{\epsilon^*}}(n), p_{t(n)}) - L(A_{p_{\epsilon^*}}(n), p_{\hat{s}(n)})]) \geq \epsilon_v < \delta_v$. For each $n$ we define
\[
D_v(n) = \{s(n)|L(A_{p_{\epsilon^*}}(n), p_{s(n)}) - L(A_{p_{\epsilon^*}}(n), p_{\hat{s}(n)})| < \epsilon_v\}.
\]
For each $n \geq N(v)$ we have
\[
Pr(t(n) \in D_v(n)) = 1 - Pr([L(A_{p_{\epsilon^*}}(n), p_{t(n)}) - L(A_{p_{\epsilon^*}}(n), p_{\hat{s}(n)})]) \geq \epsilon_v > 1 - \delta_v.
\]
For $n \geq N(1)$, define $D_n = D_{v(n)}$, where $v(n) = \max(v|N(v) \leq n; v \leq n)$. Thus for all, $n \geq N(1)$ we have $Pr(t(n) \in D_n) > 1 - \delta_v(n)$ and
\[
\sup_{\hat{s}(n) \in D_n} |L(A_{p_{\epsilon^*}}(n), p_{\hat{s}(n)}) - L(A_{p_{\epsilon^*}}(n), p_{\hat{s}(n)})| < \epsilon_v(n).
\]
Note that $v(n)$ tends to infinity, it follows that $\delta_v(n) \to 0$ and $\epsilon_v(n) \to 0$. Therefore, $Pr(t(n) \in D_n) \to 1$ and
\[
\lim_{n \to \infty} \{\sup_{\hat{s}(n) \in D_n} |L(A_{p_{\epsilon^*}}(n), p_{\hat{s}(n)}) - L(A_{p_{\epsilon^*}}(n), p_{\hat{s}(n)})|) = 0.
\]
That is $A$ is weakly robust. w.r.t. $p_{\epsilon}$ which is a desired contradiction.

A.4 Mc Diarmid inequality

Let $X_1, \ldots, X_n$ be $n$ independent random variables taking values in $X$ and let $Z = f(X_1, \ldots, X_n)$. If for each $1 \leq i \leq n$, there exists a constant $c_i$ such that
\[
\sup_{x_1, \ldots, x_n, x'_i \in X} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x'_i, \ldots, x_n)| \leq c_i, \forall 1 \leq i \leq n,
\]
then for any $\epsilon > 0$, $Pr[|Z - E[Z]| \geq \epsilon] \leq 2 \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$.
A.5 Proof of Example 2 ($\ell_1$ norm)

**Proof** Let $M^*$ be the solution given training data $p_n$. Due to optimality of $M^*$, we have $\|M^*\|_1 \leq g_0/c$. We can partition $Z$ as $\{Y|N(\gamma/2, X, \|\cdot\|_1)\}$ sets, such that if $z$ and $z'$ belong to the same set, then $y = y'$ and $\|x - x'\|_1 \leq \gamma$. Now, for $z_1, z_2, x_1, x_2 \in Z$, if $y_1 = y_2$, $\|x_1 - x_1'\|_1 \leq \gamma$, $y_2 = y_2'$ and $\|x_2 - x_2'\|_1 \leq \gamma$, then:

$$
\begin{eqnarray*}
|g(y_{12}[1 - f(M^*, x_1, x_2)]) - g(y_{12}[1 - f(M^*, x_1', x_2')])| & \leq & U((\|x_1 - x_2\|_2|M^*\|_1\|x_1 - x_1'\|_1 + \|x_1 - x_2\|_2|M^*\|_1\|x_2 - x_2'\|_1 + \|x_1 - x_1'\|_1|M^*\|_1\|x_2' - x_2\|_1 + \|x_2 - x_2'\|_1|M^*\|_1\|x_2 - x_2'\|_1) \\
& & + |g(y_{12}[1 - f(M^*, x_1, x_2)]) - g(y_{12}[1 - f(M^*, x_1', x_2')])| \\
& \leq & \frac{8U\gamma g_0}{c}.
\end{eqnarray*}
$$

\hfill \Box

A.6 Proof of Example 3 ($\ell_2,1$ norm)

**Proof** Let $M^*$ be the solution given training data $p_n$. Due to optimality of $M^*$, we have $\|M^*\|_{2,1} \leq g_0/c$. We can partition $Z$ in the same way as in the proof of Example 1 and use the inequality $\|M^*\|_F \leq \|M^*\|_{2,1}$ (from Theorem 3 of Feng [21]) to derive the same bound:

$$
\begin{eqnarray*}
|g(y_{12}[1 - f(M^*, x_1, x_2)]) - g(y_{12}[1 - f(M^*, x_1', x_2')])| & \leq & U((\|x_1 - x_2\|_2|M^*\|_F\|x_1 - x_1'\|_1 + \|x_1 - x_2\|_2|M^*\|_F\|x_2 - x_2'\|_1 + \|x_1 - x_1'\|_1|M^*\|_F\|x_2' - x_2\|_1 + \|x_2 - x_2'\|_1|M^*\|_F\|x_2 - x_2'\|_1) \\
& & + |g(y_{12}[1 - f(M^*, x_1, x_2)]) - g(y_{12}[1 - f(M^*, x_1', x_2')])| \\
& \leq & \frac{8U\gamma g_0}{c}.
\end{eqnarray*}
$$

\hfill \Box

A.7 Proof of Example 4 (Kernelization)

We assume $\mathbb{H}$ to be an Hilbert space with an inner product operator $\langle \cdot, \cdot \rangle$. The mapping $\phi(\cdot)$ is continuous from $X$ to $H$. The norm $\| \cdot \|_H : \mathbb{H} \to \mathbb{R}$ is defined as $\|w\|_H = \sqrt{\langle w, w \rangle}$ for all $w \in \mathbb{H}$, for matrices $\|M\|_H$ we take the entry wise norm by considering a matrix as a vector, corresponding to the Frobenius norm. The kernel function is defined as $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$.

$B_\gamma$ and $f_\Sigma(\gamma)$ are finite by the compactness of $X$ and continuity of $k(\cdot, \cdot)$. Let $M^*$ be the solution given training data $p_n$ by the optimality of $M^*$ and using the same trick as the other examples we have: $\|M^*\|_H \leq g_0/c$. Then, by considering a partition of $Z$ into $\{Y|N(\gamma/2, X, \|\cdot\|_2)\}$ disjoint subsets such that if $(x_1, y_1)$ and $(x_2, y_2)$ belong to the same set then $y_1 = y_2$ and $\|x_1 - x_2\|_2 \leq \gamma$.

We have:

$$
\begin{eqnarray*}
|g(y_{ij}[1 - f(M^*, \phi(x_1), \phi(x_2))]) - g(y_{ij}[1 - f(M^*, \phi(x_1'), \phi(x_2'))])| & \leq & U((\|\phi(x_1) - \phi(x_2)\|T M^*(\phi(x_1) - \phi(x_1')) + \|\phi(x_1) - \phi(x_2)\|T M^*(\phi(x_1') - \phi(x_2))) \\
& & + \|\phi(x_1) - \phi(x_1')\|T M^*(\phi(x_1') + \phi(x_2')) + \|\phi(x_1') - \phi(x_2)\|T M^*(\phi(x_1') + \phi(x_2'))) \\
& \leq & U((\|\phi(x_1)\|T M^*(\phi(x_1') - \phi(x_1')) + \|\phi(x_2)\|T M^*(\phi(x_1') - \phi(x_1')) + \|\phi(x_1)\|T M^*(\phi(x_2') - \phi(x_2))) \\
& & + \|\phi(x_1) - \phi(x_1')\|T M^*(\phi(x_2') + \phi(x_2')) + \|\phi(x_1') - \phi(x_2)\|T M^*(\phi(x_2') + \phi(x_2')) + \|\phi(x_2') - \phi(x_2)\|T M^*(\phi(x_2') + \phi(x_2'))).
\end{eqnarray*}
$$
Then, note that
\[
|\phi(x_1)^T M^* (\phi(x_1) - \phi(x'_1))| \leq \sqrt{\langle \phi(x_1), \phi(x_1) \rangle} \|M^*\|_H \sqrt{\langle \phi(x'_1) - \phi(x'_2), \phi(x'_1) - \phi(x'_2) \rangle}
\]
\[
\leq B \gamma g \sqrt{f_H(\gamma)}.
\]
Thus, by applying the same principle to all the terms in the right part of inequality (11), we obtain:
\[
|g(y_{ij}[1 - f(M^*, \phi(x_1), \phi(x_2))]) - g(y_{ij}[1 - f(M^*, \phi(x'_1), \phi(x'_2))])| \leq \frac{8UB \gamma g \sqrt{f_H(\gamma)} g_0}{c}.
\]

References

[1] M. Schultz and T. Joachims. Learning a Distance Metric from Relative Comparisons. In *NIPS*, 2003.
[2] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. In *ICML*, 2004.
[3] R. Rosales and G. Fung. Learning Sparse Metrics via Linear Programming. In *KDD*, pages 367–373, 2006.
[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.
[5] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman. Online Metric Learning and Fast Similarity Search. In *NIPS*, pages 761–768, 2008.
[6] K. Q. Weinberger and L. K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, 10:207–244, 2009.
[7] G.-J. Qi, J. Tang, Z.-J. Zha, T.-S. Chua, and H.-J. Zhang. An Efficient Sparse Metric Learning in High-Dimensional Space via l1-Penalized Log-Determinant Regularization. In *ICML*, 2009.
[8] Y. Ying, K. Huang, and C. Campbell. Sparse Metric Learning via Smooth Optimization. In *NIPS*, pages 2214–2222, 2009.
[9] G. Chechik, U. Shalit, V. Sharma, and S. Bengio. An Online Algorithm for Large Scale Image Similarity Learning. In *NIPS*, pages 306–314, 2009.
[10] A. M. Qamar. *Generalized Cosine and Similarity Metrics: A supervised learning approach based on nearest-neighbors*. PhD thesis, University of Grenoble, 2010.
[11] U. Shalit, D. Weinshall, and G. Chechik. Online learning in the manifold of low-rank matrices. In *NIPS*, pages 2128–2136, 2010.
[12] P. Kar and P. Jain. Similarity-based Learning via Data Driven Embeddings. In *NIPS*, 2011.
[13] W. Bian and D. Tao. Learning a Distance Metric by Empirical Loss Minimization. In *IJCAI*, pages 1186–1191, 2011.
[14] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In *NIPS*, pages 862–870, 2009.
[15] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
[16] H. Xu, C. Caramanis, and S. Mannor. Sparse Algorithms Are Not Stable: A No-Free-Lunch Theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):187–193, 2012.
[17] H. Xu and S. Mannor. Robustness and generalization. In *COLT*, pages 503–515, 2010.
[18] H. Xu and S. Mannor. Robustness and generalization. *Machine Learning Journal*, 86(3):391–423, 2012.
[19] A.N. Kolmogorov and V.M. Tihomirov. $\varepsilon$-entropy and $\varepsilon$-capacity of sets in functional spaces. *American Mathematical Society Translations*, 17(series 2):277–364, 1961.

[20] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer, 2000.

[21] B.Q. Feng. Equivalence constants for certain matrix norms. *Linear Algebra and Its Applications*, 374:247–253, 2003.