Learning a Distance Metric by Empirical Loss Minimization

Wei Bian and Dacheng Tao
Centre for Quantum Computation and Intelligence Systems
University of Technology, Sydney, NSW 2007, Australia
wei.bian@student.uts.edu.au; dachent.tao@uts.edu.au

Abstract

In this paper, we study the problem of learning a metric and propose a loss function based metric learning framework, in which the metric is estimated by minimizing an empirical risk over a training set. With mild conditions on the instance distribution and the used loss function, we prove that the empirical risk converges to its expected counterpart at rate of root-n. In addition, with the assumption that the best metric that minimizes the expected risk is bounded, we prove that the learned metric is consistent. Two example algorithms are presented by using the proposed loss function based metric learning framework, each of which uses a log loss function and a smoothed hinge loss function, respectively. Experimental results suggest the effectiveness of the proposed algorithms.

1 Introduction

It is essential to choose a suitable metric for learning machines, e.g., k-means based clustering and nearest-neighbor based classification. A good metric helps to encode the geometry information of the instance distribution, and thus improves the performance of learning algorithms. In recent years, several metric learning algorithms have been proposed, e.g., neighbourhood component analysis [Goldberger *et al.*, 2004], metric learning via large margin nearest neighbor [Weinberger *et al.*, 2006], information-theoretic metric learning method [Davis *et al.*, 2007], robust metric learning [Zha *et al.*, 2009] and semisupervised metric learning [Baghshah and Shouraki, 2009]. They share a similar intrinsic idea that a good metric keeps the similar instances close and dissimilar ones apart, and show promising performances empirically. However,it is not easy to study their statistical properties, e.g., the consistency of the learned metric. Closely related to metric learning, several algorithms were developed for subspace learning [Bian and Tao, 2008; 2011], with various intuitions. Again, theoretical analyses are omitted.

In this paper, we propose a new metric learning framework, i.e., loss function based metric learning, in which the metric is estimated by minimizing an empirical risk over a training set. This framework enjoys one important advantage that the learned metric is consistent when the cardinality of the training set goes infinite. On the contrary, it is difficult to prove the consistency of existing metric learning algorithms and there is no results show that these algorithms are consistent. To prove the consistency of the loss function based metric learning, we first show that the empirical risk converges to its expected counterpart at rate $\mathcal{O}(1/\sqrt{n})$, and then we show the learned metric that minimizes the empirical risk converges (at least equivalently) to the best metric that minimizes the expected risk, given the assumption that the best metric is bounded.

We then develop two example algorithms by using the proposed loss function based metric learning framework. The first one uses the log loss for calculating the empirical risk, while the second one is based on the hinge loss but in a smoothed form (to achieve the computational efficiency). In both algorithms, we show that their objective functions are convex and each has a Lipschitz gradient. Therefore, we can optimize both objective functions iteratively based on a fast gradient method, which has the optimal convergence rate at $\mathcal{O}(1/k^2)$, wherein k is the number of iteration steps.

To evaluate the effectiveness of the proposed metric learning algorithms, we compare them against popular metric learning algorithms, e.g., NCA [Goldberger *et al.*, 2004], LMNN [Weinberger *et al.*, 2006], RCA [Shental *et al.*, 2002], and [Xing *et al.*, 2002]. On six data sets from the UCI machine learning repository [Asuncion and Newman, 2007], the nearest neighbor based classification shows that the proposed log loss and smoothed hinge loss based metric learning algorithms achieve competitive performance on all experiments.

2 Loss Function based Metric Learning

We observe a set of independent and identically-distributed (i.i.d.) instances $\mathcal{Z} = \{ (x_i, y_i) \mid i = 1, 2, ..., n\}$ that are drawn from an unknown joint distribution $P(x, y)$, which is defined on $\mathcal{X} \times \mathcal{Y}$, wherein $\mathcal{X}$ is a measurable space and $\mathcal{Y}$ is a finite discrete set. Define similar and dissimilar sets

$$\mathcal{S} : (x_i, x_{i'}) \in \mathcal{S} \text{ if } y_i = y_{i'},$$

$$\mathcal{D} : (x_i, x_{i'}) \in \mathcal{D} \text{ if } y_i \neq y_{i'}.$$

By defining a metric M, which is a positive semidefinite matrix, the distance between two instances x_i and $x_{i'}$ is given by

$$d(x_i, x_{i'}) = \sqrt{(x_i - x_{i'})^T M (x_i - x_{i'})}. \tag{1}$$

Metric learning finds a metric M that encodes the side information defined by instance pairs in $\mathcal{S}$ and $\mathcal{D}$. In particular, we expect that the learned metric M makes distances between samples of pairs in $\mathcal{S}$ small and distances between samples of pairs in $\mathcal{D}$ large.

2.1 Empirical Risk

Introducing random variable r such that $r = 1$ if $(x, x') \in \mathcal{S}$ and $r = -1$ if $(x, x') \in \mathcal{D}$, then we can use the following risk for metric learning

$$\mathcal{R}_\ell(f) = \int \ell(rf(x, x'))dP(x, x', r), \tag{2}$$

where the unknown distribution $P(x, x', r)$ is defined on the space $\mathcal{X}^2 \times \{-1, 1\}$ and $f(x, x', r)$ is a decision function.

Given the training data sampled from a joint unknown distribution $P(x, y)$, but not directly from $P(x, x', r)$. It is necessary to consider $P(x, x', r)$ being induced from $P(x, y)$: to obtain a triplet (x, x', r) sampled from $P(x, y)$, we first independently sample (x, y) and (x', y'), and then determine r by checking values of y and y'. Thus, based on the training set $\mathcal{Z}$, we can construct a corresponding set $\mathcal{Z}'$ for $P(x, x', r)$, i.e., $\mathcal{Z}' = \{(x_i, x_{i'}, r_{ii'})| \ 1 \leq i < i' \leq n\}$. It is worth emphasizing that $\mathcal{Z}'$ is not i.i.d. for $P(x, x', r)$. For example, even two triplets $(x_i, x_{i'}, r_{ii'})$ and $(x_j, x_{j'}, r_{jj'})$ can be dependent if any two indeces in a quadruple (i, i', j, j') are the same. This example shows that data directly sampled from $P(x, x', r)$ are not i.i.d. either.

Therefore, given the empirical risk defined by

$$\hat{\mathcal{R}}_\ell(f, \mathcal{Z}') = \frac{2}{(n-1)n} \sum_{i<i'} \ell(r_{ii'} f(x_i, x_{i'})), \tag{3}$$

it is theoretically necessary to answer the following two questions: 1) does this empirical risk converge to the expected risk (2)? and 2) if so, how can we use this empirical risk to obtain a consistent estimation of a metric?

2.2 Metric Estimation

To estimate the metric by minimizing the empirical risk, it is necessary to specify the decision function f and the loss function ℓ. According to [Maurer, 2005], one possible choice of f is

$$f(x, x'; c, M) = c - (x - x')^T M(x - x'), \tag{4}$$

where M is a metric and c is a positive variable denoting the decision threshold, i.e., we predict $r = 1$ if $f > 0$, or $r = -1$ otherwise. The choice of f implies an essential difference between metric learning and classification: the former predicts on an instance pair (x, x') while the later predicts on a single instance x. Regarding the loss function ℓ, there are many choices, e.g., the log loss and the hinge loss. We develop two example algorithms in Section 4, and consider a general loss function for subsequent theoretical analyses.

By combining the decision function (4) and a general loss function ℓ, the best metric $M^\star$ associated with the decision threshold $c^\star$ that minimizes the expected risk (2) are given by

$$(M^\star, c^\star) = \arg\min_{M,c} \mathcal{R}_\ell(f). \tag{5}$$

Property 2.1. *If the second order moment of the marginal distribution $P(x)$ is finite, i.e., $\mathbb{E}\|x\|^2 < \infty$, then for any convex loss function $\ell(u)$ that is upper bounded by $a_0|u| + b_0$, (both a_0 and b_0 are positive constants), the expected risk $\mathcal{R}_\ell(f)$ is convex and well defined on $\{M, c|M \succeq 0, c \geq 0\}$.*

Proof. For any fixed (M, c), we have

$$\begin{aligned} \ell(rf) &\leq a_0|rf| + b_0 \\ &= a_0 \left|c - (x - x')^T M(x - x')\right| + b_0 \\ &\leq 2a_0 x^T Mx + 2a_0 x'^T Mx' + a_0 c + b_0. \end{aligned} \tag{6}$$

Because x and x' are independent, and $\mathbb{E}\|x\|^2 < \infty$, we have $\mathcal{R}_\ell(f) = \mathbb{E}\ell(rf) < \infty$, i.e., it is well defined on $\{M, c|M \succeq 0, c \geq 0\}$. In additoin, f is linear in (M, c) and $\ell(u)$ is convex, so $\ell(rf)$ is convex jointly on (M, c). Finally, since the expectation operation is linear, $\mathcal{R}_\ell(f) = \mathbb{E}\ell(rf)$ is also convex. This completes the proof. □

The theoretical analyses in the rest of the paper are based on the following assumption.

Assumption 2.2. *The best metric $M^\star$ and the decision threshold $c^\star$ are bounded, i.e., $0 \preceq M^\star \preceq \alpha I$ and $0 \leq c \leq \alpha$, wherein α is a sufficiently large positive constant.*

This assumption is natural for metric learning, because a well-posed problem should have a bounded metric with a bounded decision threshold. Let $\mathcal{Q} = \{ \ (M, c) \ |0 \preceq M^\star \preceq \alpha I, 0 \leq c \leq \alpha\}$, and then the empirical estimation for (5) can be redefined by

$$(\hat{M}, \hat{c}) = \arg \min_{(M,c) \in \mathcal{Q}} \hat{\mathcal{R}}_\ell(f, \mathcal{Z}'). \tag{7}$$

Similar to the proof for Property 2.1., we can show that $\hat{\mathcal{R}}_\ell(f, \mathcal{Z}')$ is convex, and thus (7) has the global optimal solution. In the rest of the paper, we refer to (7) as the loss function based metric learning framework. According to Property 2.1., any loss function ℓ that satisfies the condition

$$\ell(u) < a_0|u| + b_0 \tag{8}$$

can be used in this framework. Section 4 shows two examples based on the log loss and a smoothed hinge loss, respectively.

3 Consistency

In this section, we study the consistency of the proposed loss function based metric learning framework. We first prove that the empirical risk (3) converges to the expected risk (2) at rate $\mathcal{O}(1/\sqrt{n})$ under Assumption 2.2, wherein n is the cardinality of the training set. We then prove that the metric $\hat{M}$ learned by minimizing the empirical risk (3) is consistent.

3.1 Convergence of the Empirical Risk

According to discussions above, the empirical risk (3) is not an independent sum over the training set $\mathcal{Z}'$. Thus, the proof for its convergence to the expected risk is nontrivial. In this subsection, we prove that the empirical risk $\hat{\mathcal{R}}_\ell(f, \mathcal{Z}')$ converges to $\mathcal{R}_\ell(f)$ in probability at rate $\mathcal{O}(1/\sqrt{n})$. First, we have the following theorem.

Theorem 3.1. *Given the loss function ℓ that satisfies the condition (8), and the fourth order moment of the marginal distribution $P(x)$ is finite, i.e., $\mathbb{E}\|x\|^4 < \infty$, then*

$$\max_{(M,c)\in\mathcal{Q}} \mathbb{E}[\hat{\mathcal{R}}_\ell(f,\mathcal{Z}') - \mathcal{R}_\ell(f)]^2 = \mathcal{O}(n^{-1}),\ n \to \infty \quad (9)$$

Proof. In the rest of the proof, we denote $\ell(r_{ii'} f(x_i, x_i'))$ by $\ell_{ii'}$.

First, since the expectation operation is linear, we have

$$\mathbb{E}_{P(x,x',r)}\hat{\mathcal{R}}_\ell(f,\mathcal{Z}') = \mathcal{R}_\ell(f). \quad (10)$$

Then, the second moment of the empirical risk is given by

$$\begin{aligned} &\mathbb{E}_{P(x,x',r)}[\hat{\mathcal{R}}_\ell(f,\mathcal{Z}')]^2 \\ &= \mathbb{E}\left[\frac{2}{(n-1)n}\sum_{i<i'}\ell_{ii'}\right]^2 \\ &= \mathbb{E}\left[\frac{4}{(n-1)^2n^2}\sum_{i<i',\, j<j'}\ell_{ii'}\ell_{jj'}\right]. \end{aligned} \quad (11)$$

Define three index sets, $I_0 = \{(i,i',j,j')|1 \le i < i' \le n, 1 \le j < j' \le n,)\}$, $I_1 = \{(i,i',j,j')|(i,i',j,j') \in I_0$, and $i,i',j,j'$are different from each other$\}$, and $I_2 = I_0 - I_1$. It can be calculated that the cardinalities of I_1 and I_2 are $n(n-1)(n-2)(n-3)/4$ and $n(n-1)(4n-6)/4$, respectively.

Next, we split the summation in (11) into two terms T_1 and T_2, i.e.,

$$T_1 = \frac{4}{(n-1)^2n^2}\sum_{(i,i',j,j')\in I_1}\mathbb{E}\left[\ell_{ii'}\ell_{jj'}\right], \quad (12)$$

$$T_2 = \frac{4}{(n-1)^2n^2}\sum_{(i,i',j,j')\in I_2}\mathbb{E}\left[\ell_{ii'}\ell_{jj'}\right]. \quad (13)$$

For T_1, the independence between $\ell_{ii'}$ and $\ell_{jj'}$ leads to

$$\begin{aligned} T_1 &= \frac{4}{(n-1)^2n^2}\sum_{(i,i',j,j')\in I_1}[\mathbb{E}\ell]^2 \\ &= \frac{4}{(n-1)^2n^2}\frac{n(n-1)(n-2)(n-3)}{4}\left[\mathcal{R}_\ell(f)\right]^2 \\ &= \left[\mathcal{R}_\ell(f)\right]^2 + \mathcal{O}(\mathcal{K}_1 n^{-1}), \end{aligned} \quad (14)$$

where $\mathcal{K}_1 = \max_{(M,c)\in\mathcal{Q}}\mathcal{R}_\ell(f)$. The existence of $\mathcal{K}_1$ is guaranteed by the fact that $\mathcal{R}_\ell(f)$ is convex and non-negative function and $\mathcal{Q} = \{\ (M,c)\ |0 \preceq M \preceq \alpha I, 0 \le c \le \alpha\}$ is a closed set.

According to (6), on $\mathcal{Q} = \{\ (M,c)\ |0 \preceq M \preceq \alpha I, 0 \le c \le \alpha\}$, we have

$$\begin{aligned} \ell &\le 2a_0 x^T M x + 2a_0 x'^T M x' + a_0 c + b_0 \\ &\le 2a_0\alpha x^T x + 2a_0\alpha x'^T x + a_0\alpha + b_0. \end{aligned} \quad (15)$$

Furthermore, since $\mathbb{E}\|x\|^4 < \infty$, $\mathbb{E}(\ell^2)$ can then be upper bounded by a sufficiently large constant $\mathcal{K}_2$,

$$\mathbb{E}(\ell^2) \le \mathbb{E}(2a_0\alpha x^T x + 2a_0\alpha x'^T x + a_0\alpha + b_0)^2 < \mathcal{K}_2. \quad (16)$$

Thus, we have

$$\begin{aligned} |T_2| &= \frac{4}{(n-1)^2n^2}\left|\sum_{(i,i',j,j')\in I_2}\mathbb{E}\left[\ell_{ii'}\ell_{jj'}\right]\right| \\ &\le \frac{4}{(n-1)^2n^2}\sum_{(i,i',j,j')\in I_2}\mathbb{E}\left[\frac{\ell_{ii'}^2 + \ell_{jj'}^2}{2}\right] \\ &\le \frac{4}{(n-1)^2n^2}\frac{n(n-1)(4n-6)}{4}\mathbb{E}\ell^2 \\ &= \mathcal{O}(\mathcal{K}_2 n^{-1}). \end{aligned} \quad (17)$$

Combining (10), (14) and (17), we have

$$\mathbb{E}[\hat{\mathcal{R}}_\ell(f,\mathcal{Z}') - \mathcal{R}_\ell(f)]^2 = \mathcal{O}((\mathcal{K}_1 + \mathcal{K}_2)n^{-1}). \quad (18)$$

Since $\mathcal{K}_1$ and $\mathcal{K}_2$ are constants and do not depend on (M,c), we have (9). This completes the proof. □

By Chebyshev's inequality, (9) immediately gives the following Corollary.

Corollary 3.2. *The empirical risk converges to the expected risk at rate root-n, i.e.,*

$$\hat{\mathcal{R}}_\ell(f,\mathcal{Z}') - \mathcal{R}_\ell(f) = \mathcal{O}_p(1/\sqrt{n}),\ \forall (M,c) \in \mathcal{Q} \quad (19)$$

3.2 Consistency of the Learned Metric

We have proved that for any function f specified by (M,c), the empirical risk (3) converges to the expected risk (2). Next, the following proof for Theorem 3.3 shows that the learned distance metric $\hat{M}$ by minimizing the empirical risk converges to the best distance metric $M^\star$ that minimizes the expected risk.

Theorem 3.3. *Given the conditions in Theorem 3.1, the learned $(\hat{M},\hat{c})$ by minimizing the empirical risk (7) is consistent, i.e.,*

$$(\hat{M},\hat{c}) \xrightarrow{P} (M^\star, c^\star). \quad (20)$$

Proof. According to (7), $(\hat{M},\hat{c})$ minimizes the empirical risk $\hat{\mathcal{R}}_\ell(f,\mathcal{R})$, and thus we have

$$\hat{\mathcal{R}}_\ell(\hat{M},\hat{c}) \le \hat{\mathcal{R}}_\ell(M^\star,c^\star) \quad (21)$$

Similarly, $(M^\star,c^\star)$ minimizes the expected risk $\mathcal{R}_\ell(f)$, and thus it gives

$$\mathcal{R}_\ell(M^\star,c^\star) \le \mathcal{R}_\ell(\hat{M},\hat{c}) \quad (22)$$

Combining (21) and (22), we have

$$\begin{aligned} 0 \le &\mathcal{R}_\ell(\hat{M},\hat{c}) - \mathcal{R}_\ell(M^\star,c^\star) \\ = &\mathcal{R}_\ell(\hat{M},\hat{c}) - \hat{\mathcal{R}}_\ell(\hat{M},\hat{c}) + \hat{\mathcal{R}}_\ell(\hat{M},\hat{c}) - \mathcal{R}_\ell(M^\star,c^\star) \\ \le &\mathcal{R}_\ell(\hat{M},\hat{c}) - \hat{\mathcal{R}}_\ell(\hat{M},\hat{c}) + \hat{\mathcal{R}}_\ell(M^\star,c^\star) - \mathcal{R}_\ell(M^\star,c^\star) \end{aligned} \quad (23)$$

Therefore,

$$\begin{aligned} &\mathbb{E}[\mathcal{R}_\ell(\hat{M},\hat{c}) - \mathcal{R}_\ell(M^\star,c^\star)]^2 \\ &\le 2\mathbb{E}[\mathcal{R}_\ell(\hat{M},\hat{c}) - \hat{\mathcal{R}}_\ell(\hat{M},\hat{c})]^2 \\ &\quad + 2\mathbb{E}[\hat{\mathcal{R}}_\ell(\hat{M}^\star,\hat{c}^\star) - \mathcal{R}_\ell(M^\star,c^\star)]^2 \\ &\le 4\max_{(M,c)\in\mathcal{Q}}\mathbb{E}[\hat{\mathcal{R}}_\ell(M,c) - \mathcal{R}_\ell(M,c)]^2 \end{aligned} \quad (24)$$

By (9) and Chebyshev's inequality, we have

$$\mathcal{R}_\ell(\hat{M}, \hat{c}) \xrightarrow{P} \mathcal{R}_\ell(M^\star, c^\star) \tag{25}$$

When $\mathcal{R}_\ell(M, c)$ is strongly convex, (25) implies (20). When $\mathcal{R}_\ell(M, c)$ is weakly convex, $(M^\star, c^\star)$ is an equivalent solution set, and thus (25) implies that $(\hat{M}, \hat{c})$ converges to this equivalent solutions set. This completes the proof. □

4 Example Algorithms

We develop two example algorithms by using the proposed loss function based metric learning framework. They use the log loss and a smoothed hinge loss [Rennie, 2005], respectively. For each example, we show that its empirical risk has a Lipschitz gradient, and thus can be optimized by using a fast algorithm with the optimal convergence rate at $\mathcal{O}(1/k^2)$, wherein k is the number of iteration steps.

To simplify formulations in following subsections, we introduce a block-diagonal matrix $X = \text{diag}(M, c)$, and thus the decision function can be rewritten as

$$f(x, x'; c, M) = \langle A, X \rangle, \tag{26}$$

where $\langle A, B \rangle = \text{tr}(A^T B)$ and

$$A = \text{diag}\left(-(x - x')(x - x')^T, 1\right). \tag{27}$$

Furthermore, the training set $\mathcal{Z}' = \{(x_i, x_{i'}, r_{ii'})| \ 1 \leq i < i' \leq n\}$ contains $m = n(n-1)/2$ triplets. For convenience, we replace the paired index (i, i') with a single index j, and thus $\mathcal{Z}' = \{(x_j, x'_j, r_j)| \ 1 \leq j \leq m\}$. Each r_j is associated with a block-diagonal matrix $A_j = \text{diag}\left(-(x_j - x'_j)(x_j - x'_j)^T, 1\right)$.

4.1 The Log Loss based Example

In this example, we use the log loss function

$$\ell(u) = \ln\left(1 + \exp(-u)\right), \tag{28}$$

which has a gradient $|\nabla \ell(u)| \leq 1$ and satisfies the condition (8), i.e., $\ell(u) < |u| + 1$.

We replace the loss function in (7) with (28), and then the empirical risk can be rewritten as

$$\min_{X \in \mathcal{Q}} \hat{\mathcal{R}}(X) = \frac{1}{m} \sum_{j=1}^{m} \ln\left(1 + e^{-r_j \langle A_j, X \rangle}\right). \tag{29}$$

According to Property 4.1 shown below, we know that $\hat{\mathcal{R}}(X)$ in (29) has a Lipschitz gradient with the parameter $L = (1/4m)\sum_{j=1}^{m} \|A_j\|_F^2$.

Property 4.1. *For any direction $\Delta \in S^{(p+1)\times(p+1)}$, where p is the dimension of an instance x, we have*

$$\langle \nabla^2 \hat{\mathcal{R}}(X)\Delta, \Delta \rangle \leq L\|\Delta\|_F^2, \tag{30}$$

where $L = (1/4m)\sum_{j=1}^{m} \|A_j\|_F^2$.

Proof. First, the gradient of $\hat{\mathcal{R}}(X)$ is given by

$$\nabla \hat{\mathcal{R}}(X) = \frac{1}{m} \sum_{j=1}^{m} \frac{-rA_j}{1 + e^{r\langle A_j, X \rangle}}. \tag{31}$$

Define function $\phi(\epsilon) = \langle \nabla \hat{\mathcal{R}}(X + \epsilon\Delta), \Delta \rangle$ with $\epsilon > 0$, and then we have

$$\begin{aligned} \phi(\epsilon) - \phi(0) &= \langle \nabla \hat{\mathcal{R}}(X + \epsilon\Delta) - \nabla \hat{\mathcal{R}}(X), \Delta \rangle \\ &= \frac{1}{m} \sum_{j=1}^{m} \left\langle \frac{-rA_j}{1 + e^{r\langle A_j, X + \epsilon\Delta \rangle}} - \frac{-rA_j}{1 + e^{r\langle A_j, X \rangle}}, \Delta \right\rangle \\ &= \frac{1}{m} \sum_{j=1}^{m} \left\langle \frac{-rA_j(1 - e^{\epsilon r\langle A_j, \Delta \rangle})}{\left(1 + e^{r\langle A_j, X + \epsilon\Delta \rangle}\right)\left(1 + e^{-r\langle A_j, X \rangle}\right)}, \Delta \right\rangle. \end{aligned} \tag{32}$$

Therefore,

$$\begin{aligned} \langle \nabla^2 \hat{\mathcal{R}}(X)\Delta, \Delta \rangle &= \phi'(0) = \lim_{\epsilon \to 0} \frac{\phi(\epsilon) - \phi(0)}{\epsilon} \\ &= \frac{1}{m} \sum_{j=1}^{m} \frac{r^2 \langle A_j, \Delta \rangle^2}{\left(1 + e^{r\langle A_j, X \rangle}\right)\left(1 + e^{-r\langle A_j, X \rangle}\right)} \\ &\leq \frac{1}{m} \sum_{j=1}^{m} \frac{r^2 \langle A_j, \Delta \rangle^2}{4} \leq \frac{1}{4m} \sum_{j=1}^{m} \|A_j\|_F^2 \|\Delta\|_F^2. \end{aligned} \tag{33}$$

This completes the proof. □

The log loss based empirical risk $\hat{\mathcal{R}}(X)$ is convex and has a Lipschitz gradient, so we can obtain the optimal solution of (29) by using the improved first-order method [Nesterov, 2004]. It has been proved that this method achieves the optimal convergence rate at $\mathcal{O}(1/k^2)$, wherein k is the number of iteration steps. Below, we show the optimization procedure for each iteration step in three stages.

Stage 1 We solve a standard first-order problem defined by

$$X_k^1 = \arg\min_{X \in \mathcal{Q}} \frac{L}{2}\|X - X_k\|_F^2 + \langle \nabla \hat{\mathcal{R}}(X_k), X - X_k \rangle, \tag{34}$$

which utilizes the gradient of the solution $X_k = \text{diag}(M_k, c_k)$ at the current iteration step. Note that $\nabla \hat{R}(X_k)$ can be calculated by using (31) and it is block-diagonal. The problem (34) is equivalent to

$$X_k^1 = \arg\min_{X \in \mathcal{Q}} \frac{L}{2}\|X - Y_k^1\|_F^2, \tag{35}$$

where Y_k^1 is block-diagonal and it is given by

$$Y_k^1 = \text{diag}\left(M_k - \frac{\nabla \hat{\mathcal{R}}(M_k)}{L}, c_k - \frac{\nabla \hat{\mathcal{R}}(c_k)}{L}\right). \tag{36}$$

According to (35), we have 1) the two diagonal blocks M and c in X are not coupled, so M_k^1 and c_k^1 in X_k^1 can be obtained independently; and 2) it is invariant to rotation, so it can be reduced to an optimization on a diagonal matrix by rotating it to the eigen-space of Y_k^1. This equivalent problem is easy to deal with. We show the solution directly and omit details. To obtain M_k^1, we conduct eigen-decomposition on the first block of Y_k^1 and then we have

$$\left(M_k - \frac{\nabla \hat{\mathcal{R}}(M_k)}{L}\right) = \sum_i \lambda_i \xi_i \xi_i^T. \tag{37}$$

According to the constraint $0 \preceq M \preceq \alpha I$ in $\mathcal{Q}$, we have

$$M_k^1 = \sum_{i \in I} \lambda_i \xi_i \xi_i^T, \tag{38}$$

where $I = \{\, i \,|0 \le \lambda_i \le \alpha\}$. For c_k^1, since $0 \le c \le \alpha$, we have

$$c_k^1 = \min\left(\max\left(0, c_k - L^{-1}\nabla\hat{\mathcal{R}}(c_k)\right), \alpha\right). \tag{39}$$

Stage 2 We solve another minimization that combines all the gradients in the previous iterations and makes use of a *prox-function* $p_p(X)$ for the primal feasible set $\mathcal{Q}$

$$X_k^2 = \arg\min_{X \in \mathcal{Q}} \frac{L}{\sigma_p} p_p(X) + \sum_{i=0}^{k} \alpha_i \langle \nabla\hat{\mathcal{R}}(X_i), X - X_i\rangle, \tag{40}$$

where the weighting parameter $\alpha_i = (i+1)/2$ according to [Nesterov, 2004] and $p_p(X)$ is required to be strongly convex with the parameter σ_p. We choose the *prox-function* as

$$p_p(X) = \|X\|_F^2, \tag{41}$$

which vanishes at $X_0 = O$ and has the convexity parameter $\sigma_p = 2$. Similar to the method used to solve (34), we can get $X_k^2 = \mathrm{diag}(M_k^2, c_k^2)$ by

$$M_k^2 = \sum_{i \in I} \sigma_i \zeta_i \zeta_i^T, \tag{42}$$

where $I = \{\, i \,|0 \le \sigma_i \le \alpha\}$ and $\{\sigma_i, \zeta_i\}$ are the corresponding eigenvalues and eigenvectors of

$$-L^{-1} \sum_{i=0}^{k} \alpha_i \nabla\hat{\mathcal{R}}(X_i) = \sum_i \sigma_i \zeta_i \zeta_i^T, \tag{43}$$

and

$$c_k^2 = \min\left(\max\left(0, -L^{-1}\sum_{i=0}^{k} \alpha_i \nabla\hat{\mathcal{R}}(c_i)\right), \alpha\right). \tag{44}$$

Stage 3 $X_{k+1} = \mathrm{diag}(M_{k+1}, c_{k+1})$ for the next iteration step is given by the weighted combination

$$X_{k+1} = \tau_k X_k^1 + (1 - \tau_k) X_k^2, \tag{45}$$

where it is suggested that $\tau_k = (k+1)/(k+3)$ [Nesterov, 2004]. Algorithm 1 summarizes the above procedure.

4.2 The Smoothed Hinge Loss Case

In the second example, we utilize the smoothed hinge loss (Rennie 2005), which is defined by

$$\ell_{sh}(u) = \begin{cases} 0.5 - u, & u < 0 \\ 0.5(1-u)^2, & 0 \le u \le 1 \\ 0, & u > 1. \end{cases} \tag{46}$$

Its gradient is given by

$$\nabla\ell_{sh}(u) = \begin{cases} -1, & u < 0 \\ u - 1, & 0 \le u \le 1 \\ 0, & u > 1. \end{cases} \tag{47}$$

Algorithm 1 Loss function based metric learning

```
Input: The training set $\mathcal{Z}' = \{(x_j, x_j', r_j) | 1 \le j \le m\}$
Output: The metric $M$ and decision threshold $c$
For $k = 0, 1, 2, ...$
    - Compute $\hat{\mathcal{R}}(X_k)$ by (28) for log loss, or by (48) for the smoothed hinge loss
    - Compute $\nabla\hat{\mathcal{R}}(X_k)$ by (31) for the log loss, or by (49) for the smoothed hinge loss
    - Update $X_k^1 = \mathrm{diag}(M_k^1, c_k^1)$ by (38) and (39)
    - Update $X_k^2 = \mathrm{diag}(M_k^2, c_k^2)$ by (42) and (44)
    - Update $X_{k+1} = ((k+1)/(k+3))X_k^1 + (2/(k+3))X_k^2$
Until $|\hat{\mathcal{R}}(X_{k+1}) - \hat{\mathcal{R}}(X_k)| < \varepsilon$
```

The smoothed hinge loss satisfies the condition (8) by $|\ell(u)| \le |u| + 0.5$. With the smoothed hinge loss, the empirical risk is given by

$$\hat{\mathcal{R}}(X) = \frac{1}{m}\sum_{j=1}^{m} \ell_{sh}\left(r_j \langle A_j, X\rangle\right), \tag{48}$$

and its gradient is

$$\nabla\hat{\mathcal{R}}(X) = \frac{1}{m}\sum_{j=1}^{m} \nabla\ell_{sh}\left(r_j \langle A_j, X\rangle\right) r_j A_j. \tag{49}$$

The following property shows that the empirical risk (48) has a Lipschitz gradient (the proof is similar to the one for Property 4.1).

Property 4.2. *For any direction $\Delta \in S^{(p+1)\times(p+1)}$, where p is the dimension of the instance x, we have*

$$\langle \nabla^2\hat{\mathcal{R}}(X)\Delta, \Delta\rangle \le L\|\Delta\|_F^2, \tag{50}$$

where $L = (1/m)\sum_{j=1}^{m} \|A_j\|_F^2$.

Since (48) has a Lipschitz gradient with $L = (1/m)\sum_{j=1}^{m} \|A_j\|_F^2$, we can solve it by using the same iterative procedure and Algorithm 1 as for the log loss case.

5 Experiments

In this section, we evaluate the effectiveness of the proposed example algorithms, i.e., the log loss based metric learning algorism (LLML) and the smoothed hinge loss based metric learning algorithm (sHLML), by comparing them against three popular metric learning algorithms, i.e., the distance metric learning method (DML) proposed in [Xing *et al.*, 2002], neighborhood component analysis [Goldberger *et al.*, 2004] (NCA), and large margin nearest neighbor method [Weinberger *et al.*, 2006] (LMNN). We only consider the fully supervised learning case, in which all labels of the instances in training set are known, and thus we have the similarity or dissimilarity information of all instance pairs from the training set. However, it is worth emphasizing that the proposed example algorithms can readily deal with the case that only side information [Xing *et al.*, 2002] on the training set is known,

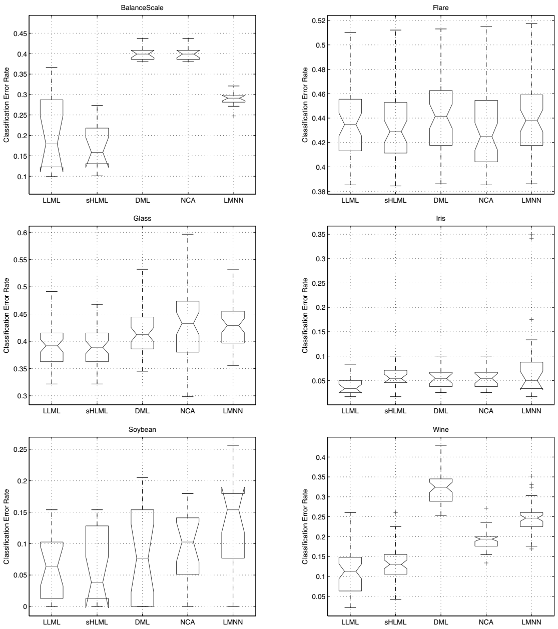

Figure 1: Performance evaluation on data sets from the UCI machine learning repository. Each method is represented by a box with whiskers. The box has lines at the lower quartile, median, and upper quartile values of the classification error rates on twenty independent experiments, and the error rate outside 1.5 times of the interquartile range from the ends of the box is regarded as whisker.

i.e., the labels of the instances in the training set are unknown but only the similarity and the dissimilarity information between (not necessarily all) instance pairs is known. Six data sets from the UCI Machine Learning Repository [Asuncion and Newman, 2007] are used in our experiments, including ”BalanceScale”,“Flare”,“Glass”,“Iris”,“Soybean” and “Wine”. On each data set, we randomly select 20 percents data as the training set and use the rest as the test set. A metric is learned by using different algorithms from the training data, and then the nearest neighbour classification with the learned metric is conducted on the test data. Twenty independent experiments on each data set are conducted, and the classification error rate of different metric learning algorithms are shown in Figure 1 by using the boxplot. From the results, one can see that the proposed loss function based algorithms have competitive performances on all six data set. In particular, significant improvements are achieved on the “BalanceScale”, “Soybean”, and “Wine” data sets.

6 Conclusion

In this paper, we have proposed the loss function based metric learning framework, in which the metric is estimated by minimizing an empirical risk. We proved that under natural assumptions on the instance distribution and the used loss function, the learned metric is consistent. Based on this framework, we have developed two example algorithms, which are based on the log loss and a smoothed hinge loss, respectively. Since the empirical risk of each algorithm has a Lipschitz gradient, they can be optimized with the optimal convergence rate. Sufficient experimental results on the data sets from UCI machine learning repository confirmed their effectiveness compared against popular metric learning algorithms.

Acknowledgments

This work was supported in part by Australian Research Council Future Fellowship (Grant No. FT100100971), Australian Research Council (ARC) Discovery Project DP1093762 and Discovery Project DP0988016.

References

[Asuncion and Newman, 2007] A. Asuncion and D.J. Newman. UCI machine learnin‘g repository, 2007.

[Baghshah and Shouraki, 2009] Mahdieh Soleymani Baghshah and Saeed Bagheri Shouraki. Semi-supervised metric learning using pairwise constraints. In *In IJCAI*, 2009.

[Bian and Tao, 2008] Wei Bian and Dacheng Tao. Harmonic mean for subspace selection. In *ICPR*, 2008.

[Bian and Tao, 2011] Wei Bian and Dacheng Tao. Max-min distance analysis by using sequential sdp relaxation for dimension reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5), 2011.

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

[Goldberger *et al.*, 2004] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *In NIPS*, 2004.

[Maurer, 2005] Andreas Maurer. Generalization bounds for subspace selection and hyperbolic pca. In *SLSFS*, pages 185–197, 2005.

[Nesterov, 2004] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Boston:Kluwer, 2004.

[Rennie, 2005] Jason D. M. Rennie. Smooth hinge classification 1 smooth hinge loss, 2005.

[Shental *et al.*, 2002] Noam Shental, Tomer Hertz, Daphna Weinshall, and Misha Pavel. Adjustment learning and relevant component analysis. In *IN ECCV*, 2002.

[Weinberger *et al.*, 2006] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *In NIPS*, 2006.

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

[Zha *et al.*, 2009] Zheng-Jun Zha, Tao Mei, Meng Wang, Zengfu Wang, and Xian-Sheng Hua. Robust distance metric learning with auxiliary knowledge. In *In IJCAI*, 2009.