Machine Learning (2019) 108:267–295
https://doi.org/10.1007/s10994-018-5734-0

Fast generalization rates for distance metric learning

Improved theoretical analysis for smooth strongly convex distance metric learning

Han-Jia Ye[1] · **De-Chuan Zhan**[1] · **Yuan Jiang**[1]

Received: 30 June 2017 / Accepted: 16 June 2018 / Published online: 27 June 2018

Abstract
Distance metric learning (DML) aims to find a suitable measure to compute a distance between instances. Facilitated by side information, the learned metric can often improve the performance of similarity or distance based methods such as kNN. Theoretical analyses of DML focus on the learning effectiveness for squared Mahalanobis distance. Specifically, whether the Mahalanobis metric learned from the empirically sampled pairwise constraints is in accordance with the optimal metric optimized over the paired samples generated from the true distribution, and the sample complexity of this process. The excess risk could measure the quality of the generalization, i.e., the gap between the expected objective of empirical metric learned from a regularized objective with convex loss function and the one with the optimal metric. Given N training examples, existing analyses over this *non-i.i.d.* learning problem have proved the excess risk of DML converges to zero at a rate of $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. In this paper, we obtain a faster convergence rate of DML, $\mathcal{O}\left(\frac{1}{N}\right)$, when learning the distance metric with a smooth loss function and a strongly convex objective. In addition, when the problem is relatively easy, and the number of training samples is large enough, this rate can be further improved to $\mathcal{O}\left(\frac{1}{N^2}\right)$. Synthetic experiments validate that DML can achieve the specified faster generalization rate, and results under various settings help explore the theoretical properties of DML a lot.

Keywords Distance metric learning · Generalization analysis · Excess risk

1 Introduction

Similarity and distance measures often act as essential components in machine learning algorithms. For example, the Euclidean distance can be used to measure nearest neighbors in

Editor: Peter Flach..

✉ De-Chuan Zhan
zhandc@nju.edu.cn

[1] Nanjing University, Nanjing, China

both classification (kNN) and clustering tasks (KMeans). Since such a single form of distance cannot be universally applied, distance metric learning (DML) methods aim to find suitable similarity/distance metrics to compare instances better and lead to adaptable measurement for various real applications. With the help of side information, the learned metric makes similar instances close to each other while pushes dissimilar ones far away (Kulis 2012; Bellet et al. 2015). It has been validated that DML methods are able to find good distance and similarity measures effectively for classification (Weinberger et al. 2006; Davis et al. 2007), clustering (Xing et al. 2003; Law et al. 2016b; Park et al. 2015), ranking (McFee and Lanckriet 2010; Lim et al. 2013; Chechik et al. 2010), semantic discovering (Frome et al. 2007; Changpinyo et al. 2013; Ye et al. 2016b), recommendation (Hsieh et al. 2017), etc.

Given instance and label space $\mathcal{X}$ and $\mathcal{Y}$, example $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ is sampled from the latent joint distribution $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For a pair of instances $(\mathbf{x}_i, \mathbf{x}_j)$ with $\mathbf{x}_{i/j} \in \mathbb{R}^d$, their squared Mahalanobis distance is defined as:

$$\mathrm{Dis}^2_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top M(\mathbf{x}_i - \mathbf{x}_j) \ . \tag{1}$$

$M \in \mathcal{S}_d^+$ is a $d \times d$ positive semi-definite (PSD) matrix. The Mahalanobis distance metric not only measures the correlation between different features (Lim et al. 2013; Ye et al. 2016a), but also possesses good optimization properties (Qian et al. 2013, 2015).

Types of side information provide supervision during the metric training process, based on which distances between constructed pairs are optimized. Existing distance metric learning methods can be mainly categorized into two parts based on the way they use side information, i.e., focus on pairwise or higher order comparison constraints. Pairwise side information indicates whether two instances $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar or not directly (Shalev-Shwartz et al. 2004; Davis et al. 2007). Besides, higher order relationship is also popular for their owned rich direction information (Weinberger and Saul 2009; Hwang et al. 2013; Law et al. 2016a). For instance, relative comparison relationships are contained in a triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$, i.e., the target neighbor $\mathbf{x}_j$ is more similar to the center $\mathbf{x}_i$ than an imposter $\mathbf{x}_k$. We focus our analysis on the former pairwise side information in this paper.

Given N examples $\{\mathbf{z}_i = (\mathbf{x}_i, y_i)\}_{i=1}^N$ sampled *i.i.d.* from $\mathcal{Z}$, the Mahalanobis metric M can be learned based on their pairwise relationship:

$$\hat{M} = \mathop{\arg\min}_{M \in \mathcal{S}_d^+} \hat{F}(M) = \mathop{\arg\min}_{M \in \mathcal{S}_d^+} \underbrace{\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \ell(q_{ij}(\gamma - \mathrm{Dis}^2_M(\mathbf{x}_i, \mathbf{x}_j)))}_{\epsilon_N(M)} + \Omega(M) \ . \tag{2}$$

$q_{ij} = \mathbb{I}[y_i = y_j] \in \{-1, 1\}$ indicates whether two instances are affiliated to the same class. It equals 1 if $y_i = y_j$ and -1 otherwise. γ is a pre-defined threshold value. $\ell(\cdot)$ is a convex non-negative loss function, which is usually an upper bound of the 0-1 loss function. The objective in Eq. 2 utilizes all possible $N(N-1)$ pairwise relationship between two different instances, and requires the distance between them in accordance with their label supervision. This optimization problem finds a metric $\hat{M}$, measured with which the same class instances have distances smaller than γ, while instances in different classes are pushed away than the threshold. $\Omega(\cdot)$ is a non-negative convex regularizer on M, which controls the complexity and structure of the metric. $\epsilon_N(M)$ is the empirical risk of M, measured by loss function value with $N(N-1)$ pairs. These two components form the empirical objective $\hat{F}(M)$ of the distance metric learning problem. Benefited from the convex property of Eq. 2, we can get an optimal metric $\hat{M}$ with standard optimization techniques.

The true risk of the above learning process can be defined as follows:

$$M^* = \arg\min_{M \in \mathcal{S}_d^+} F(M) = \arg\min_{M \in \mathcal{S}_d^+} \underbrace{\mathbb{E}_{\mathbf{z}_1, \mathbf{z}_2}[\ell(q_{12}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_1, \mathbf{x}_2)))]}_{\epsilon(M)} + \Omega(M) \ . \tag{3}$$

The expectation $\mathbb{E}(\cdot)$ in Eq. 3 involves the loss value w.r.t. two randomly sampled examples $\mathbf{z}_1$ and $\mathbf{z}_2$ from $\mathcal{Z}$. The expected objective $F(M)$ contains the expected risk $\epsilon(M)$ and the regularizer $\Omega(M)$, whose optimal solution is M^*. This expected risk reveals the generalization ability of a metric, which is measured over an unseen pair of examples. Based on this expected objective, we define the *excess risk*

$$F(\hat{M}) - F(M^*) = F(\hat{M}) - \min_M F(M) = \epsilon(\hat{M}) - \epsilon(M^*) + \Omega(\hat{M}) - \Omega(M^*) \ , \tag{4}$$

i.e., the difference between the expected objective of empirical optimal $\hat{M}$ and true optimal M^*. It justifies whether the distance metric learned by optimizing over empirical training example pairs in Eq. 2 is consistent with the one learned from pairs generated from true distribution in Eq. 3. Besides, it also reveals the difference of generalization ability between the empirical optimal metric $\hat{M}$ and the true optimal one M^* when testing distance measure on unseen randomly sampled pairs.

In this paper, we propose to bound the **excess risk** $F(\hat{M}) - F(M^*)$ given *smooth loss function* and *strongly convex objective*, and we focus on the rate the excess risk converging to zero w.r.t. the number of training examples. The main technologies for the proof are followed from Zhang et al. (2017). The analysis on distance metric learning scenario, however, is a *non-trivial* extension, since the measure over pairs of examples is *not i.i.d.* as in traditional classification task anymore. In our proof, we obtain a $\mathcal{O}\left(\frac{1}{N}\right)$ convergence rate of the excess risk for the empirical optimized metric $\hat{M}$. Besides, when the task is relative easy, i.e., $\epsilon(M^*)$ is small, the convergence rate can be further improved to $\mathcal{O}\left(\frac{1}{N^2}\right)$ given a large number of training examples. Our theory validates the fast learning rate of distance metric learning objective given a wide condition (as we will discuss in Sect. 4). Besides, it verifies that the distance metric learning can achieve the same fast convergence rates as in traditional classification tasks in spite of facing non-i.i.d. examples.

In the rest of this paper, we first discuss related work on distance metric learning from both algorithmic and theoretical perspectives. Then, the primary results and analyses are described in detail. After that, we extend our analysis on a pairwise ranking problem. After the theorem proof, we validate the theory on synthetic data, and various settings are investigated to show different properties of the learning problem. Last are the conclusion and discussion.

2 Related work

Types of side information are utilized to guide the properties of distance between pairs of examples in distance metric learning. Xing et al. (2003) propose to utilize pairwise constraints and *learn* the distance between instances, which can improve the performance of clustering. Shalev-Shwartz et al. (2004) deal with this pairwise learning problem in an online manner. Later, Weinberger and Saul (2009) propose to consider triplet type side information generated from Euclidean nearest neighbors from same and different classes. A large margin is also required to improve the generalization ability. Davis et al. (2007) tackle the distance metric learning problem with an information theoretical regularizer, and (McFee and Lanckriet 2010;

Lim et al. 2013) extend the learning of Mahalanobis metric to the application of ranking and retrieval problems. Sparse metric is considered in Huang et al. (2009) and Ying et al. (2009). Distance metric learning can also be used for the collaborative recommendation as in Hsieh et al. (2017). The relationship between metric learning and SVM objective is discussed in Do et al. (2012). Detailed overviews of distance metric learning can be found in Kulis (2012) and Bellet et al. (2015).

Statistical learning theory links the performance of the learned hypothesis on both training and unseen test data. Given enough training examples sampled i.i.d. from an unknown latent distribution, the divergence between the empirical and expected statistics vanishes. Sample complexity shows the order of training samples required to achieve a specified difference between this gap. In other words, the difference shrinks between the gap given the number of training examples. Here we denote the empirical optimal and true optimal solutions are optimized over the empirical training sets and the ground-truth distribution, respectively. Two kinds of bounds are commonly used to reveal the generalization ability of the learned hypothesis. Generalization bound focuses on the gap between empirical error and expected error concerning the same empirical optimal hypothesis (or uniformly over all hypotheses); while excess risk bound analyzes whether the empirical solution could perform as well as the expected optimal solution over the ground-truth distribution. The excess risk can be obtained based on the generalization bound result in some cases (Shalev-Shwartz and Ben-David 2014).

Different theoretical tools approach the generalization ability analysis. Stability (Bousquet and Elisseeff 2002) and Rademacher complexity (Bartlett and Mendelson 2002) are the usual tools, which can induce a generalization bound with sample complexity $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ in general cases. Bartlett et al. (2005) consider a Bennett type concentration inequality, together with localized Rademacher complexity measure for a self-bounded reweight hypothesis class, which can achieve a faster rate near $\mathcal{O}\left(\frac{1}{N}\right)$. Types of properties of objective functions are also utilized to improve the convergence rate. Sridharan et al. (2009) analyze the convergence rate for strongly convex objective, especially those with convex loss and strongly convex regularizer like support vector machine, can also achieve $\mathcal{O}\left(\frac{1}{N}\right)$ using the peeling and reweight technologies. Srebro et al. (2010) considers the objective with a smooth loss function, which can also improve the convergence rate. Both the strongly convex and smooth properties are analyzed in Zhang et al. (2017), where the complexity bound is built on the properties of the first order information of the objective. In the analysis of Zhang et al. (2017), the excess risk not only gets a fast rate as in previous methods with both smooth and strongly convex property parameters but can also be proved to achieve a faster rate as $\mathcal{O}\left(\frac{1}{N^2}\right)$ when the number of examples N is large enough, and the expected risk of the optimal hypothesis is small. In this paper, our analysis of distance metric learning is based on *non-i.i.d. pairs* extracted from the underlying distribution, and our results validate that learning a metric can also get a faster rate as in i.i.d. supervised learning.

The generalization ability of the learned distance metric could be analyzed based on different theoretical tools. Jin et al. (2010) use the algorithmic stability (Bousquet and Elisseeff 2002) to measure the consistency of the learned metric and propose an online optimization method. This analysis requires the learning objective be a convex one, and a Frobenius norm of metric is often used to obtain the stability. Bellet et al. (2015) provide a stability analysis for metric learning in a more general scenario. Similar methods can also be applied to the analyses in transfer learning case with biased regularizer (Perrot and Habrard 2015), and for multiple metrics extensions (Perrot et al. 2014). Robustness is also considered as a tool for sparse (regularized) metric learning (Bellet and Habrard 2015). Verma and Branson (2015)

impose the i.i.d. assumption on given pairs, and use Rademacher complexity (Bartlett and Mendelson 2002) to find the upper and lower sample complexity bound of the metric learning objective. Similarly, Mason et al. (2017) focus on the triplet side information case and also require the triplet i.i.d. assumption. The latent dimension and low-rank property of the metric are stressed in the theoretical result. Cao et al. (2016) use the same theoretical tool to do the analysis, but properties of U-statistics (Clémençon et al. 2008) are used to deal with the non-i.i.d. pairs. Sample complexities for hinge loss with several different types of regularizer are analyzed in Cao et al. (2016). There are also several analyses related to the relationship between the learned metric and the classification performance (Bellet et al. 2012; Guo and Ying 2014). Among all existing results for batch distance metric learning, there is a general $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ type of convergence rate from $\epsilon_N(\hat{M})$ to $\epsilon(\hat{M})$, or from $\epsilon(\hat{M})$ to $\epsilon(M^*)$. This result is tight since there are no more conditions on the components of distance metric learning problem. In this paper, we explore some practical properties of the loss functions and regularizers when learning a distance metric, so as to get a faster convergence rate for the **excess risk** of the regularized objective.

3 Distance metric learning analysis

In this section, we present our main theoretical result on distance metric learning in detail. We first describe some preliminaries, followed by the main theorem, and then discuss the result.

3.1 Preliminaries

For a differentiable convex function $F(M) : \mathcal{S}_d^+ \to \mathbb{R}$, it is λ-strongly convex if for any M and M':

$$F(M) \geq F(M') + \langle \nabla F(M'), M - M' \rangle + \frac{\lambda}{2}\|M - M'\|_F^2 .$$

$\nabla F(M')$ is the gradient of $F(\cdot)$ at M'. We use notation $\mathrm{Tr}(\cdot)$ for the trace of a matrix. The Frobenius norm $\|M\|_F^2 = \mathrm{Tr}(MM^\top)$. Inner product $\langle M, M' \rangle = \mathrm{Tr}(MM'^\top)$. A non-negative regularizer $\Omega(M) : \mathcal{S}_d^+ \to \mathbb{R}^+$ is L-Lipschitz w.r.t. Frobenius norm if for M and M':

$$|\Omega(M) - \Omega(M')| \leq L\|M - M'\|_F .$$

For a non-negative loss function $\ell(\cdot) : \mathbb{R} \to \mathbb{R}^+$, it is β-smooth if its gradient is β-Lipschitz, i.e., for $x, y \in \mathbb{R}$, $|\ell'(x) - \ell'(y)| \leq \beta|x - y|$. For matrix input, the β-smooth of a function $F(M)$ also has the following property:

$$F(M) \leq F(M') + \langle \nabla F(M'), M - M' \rangle + \frac{\beta}{2}\|M - M'\|_F^2 .$$

Given smooth property for a non-negative function, the norm of gradient can be bounded by the value of the function: $\|\nabla F(M)\|_F \leq 2\beta F(M)$.

3.2 Excess risk bound for distance metric learning

For notation clarity, we define the outer product of the difference between two instances $\mathbf{x}_i$ and $\mathbf{x}_j$ as $A_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$. Furthermore, we assume it is bounded by A, i.e., $\|A_{ij}\|_F \leq A$. This bounded assumption can be implemented by an instance-wise normalization pre-processing. It is notable that in the following theoretical analysis, we only use the symmetric property of M to ensure the symmetric of distance between two instances as in Bellet et al. (2015) and Cao et al. (2016). The norm of the learned metric is also bounded, $\|M\|_F \leq R$. Then, we can obtain the following theorem for the distance metric learning problem:

Theorem 1 *Assume $F(M)$ is λ-strongly convex, $\hat{F}(M)$ is a convex function, the loss function $\ell(\cdot)$ is β-smooth, and $\Omega(M)$ is L-Lipschitz. Define:*

$$B = \sup_{\mathbf{z}_i, \mathbf{z}_j \sim \mathcal{Z}} \|\nabla \ell(q_{ij}(\gamma - \mathrm{Dis}^2_{M^*}(\mathbf{x}_i, \mathbf{x}_j)))\|_F \ ,$$

$$s = \lceil \log_2 2R + 2\log_2 N \rceil,$$

$$C_1 = 16\sqrt{2} + 8\sqrt{2\log s/\delta} \ , C_2 = 8\sqrt{2} + 8\sqrt{\log(s/\delta)} \ , C_3 = \frac{40B}{3}\log(s/\delta) \ .$$

For $0 < \delta < 1$, with probability $1 - 2\delta$:

$$F(\hat{M}) - F(M^*) \leq \max\left\{ \frac{B+L}{N^2} + \frac{\beta A^2}{2N^4}, \frac{4C_1^2 R^2 \beta^2 A^4}{\lambda N} + \frac{C_2^2}{\lambda N}\beta A^2 \epsilon(M^*) + \frac{2RC_3}{N} \right\}. \tag{5}$$

Furthermore, if

$$N \geq \frac{16\beta^2 A^2 C_1^2}{\lambda^2} \ , \tag{6}$$

we have with probability $1 - 2\delta$:

$$F(\hat{M}) - F(M^*) \leq \max\left\{ \frac{B+L}{N^2} + \frac{\beta A^2}{2N^4}, \frac{2C_2^2 \beta A^2 \epsilon(M^*)}{\lambda N} + \frac{2C_3^2}{\lambda N^2} \right\} . \tag{7}$$

Remark 1 Theorem 1 shows that the convergence rate of the regularized metric learning objective can achieve an $\mathcal{O}\left(\frac{1}{N}\right)$ rate with smooth loss function and strongly convex objective. Although there is a $\log_2 N$ term in s, it is eventually formed as a $\log\log N$ term and can be neglected in most cases. The r.h.s. of the bound in Eq. 5 is related to the parameter of the previous two properties. The larger the value of λ and the smaller the value of β, the r.h.s. of the above inequality will be tighter, thus can achieve a better convergence rate. This convergence rate of the objective function reflects the sample complexity to train a metric. To achieve the same value of error tolerance $\epsilon \ll 1$ between $F(\hat{M})$ and $F(M)$, the $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ convergence rate needs $\frac{1}{\epsilon^2}$ samples, while the $\mathcal{O}\left(\frac{1}{N}\right)$ convergence rate bound needs $\frac{1}{\epsilon}$ samples. This comparison means given fixed samples in a real task, the model possessing a faster convergence rate has a priority to achieve the required generalization error, thus could perform better on unseen instances.

Remark 2 It is notable that the faster convergence rate proved in Theorem 1 reveals the property of the regularized objective. Consider the structural risk minimization, the focus

on the convergence rate of the regularized objective is meaningful. If the expected loss function itself is strongly convex (as we will show in the experiments), then the improved rate can be applied to the loss counterpart. In a general case, when missing strongly convex and smooth property, the convergence of the loss term $\epsilon(M)$ has a lower bound with the order $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ (Guo and Ying 2014; Verma and Branson 2015), which is slower than the convergence rate of the regularized objective $F(M)$. This phenomenon has also been observed in i.i.d. learning case such as support vector machine (Sridharan et al. 2009).

Although the results in Theorem 1 focuses on the excess risk of the regularized objective, it is general enough to facilitate the analysis of the generalization ability w.r.t. the loss part in the case of Frobenius norm regularizer $\Omega(M) = \lambda' \|M\|_F^2$. In addition, we can approximate the strongly convex parameter of $F(M)$ as λ'. From Eq. 7, we have

$$\begin{aligned} F(\hat{M}) - F(M^*) &= \epsilon(\hat{M}) - \epsilon(M^*) + \lambda \|\hat{M}\|_F^2 - \lambda' \|M^*\|_F^2 \\ &\leq \max\left\{ \frac{B+L}{N^2} + \frac{\beta A^2}{2N^4}, \frac{2C_2^2 \beta A^2 \epsilon(M^*)}{\lambda' N} + \frac{2C_3^2}{\lambda' N^2} \right\} . \end{aligned}$$

Therefore, due to the non-negativity of the regularizer,

$$\epsilon(\hat{M}) - \epsilon(M^*) \leq \max\left\{ \frac{B+L}{N^2} + \frac{\beta A^2}{2N^4}, \frac{2C_2^2 \beta A^2 \epsilon(M^*)}{\lambda' N} + \frac{2C_3^2}{\lambda' N^2} \right\} + \lambda' \|M^*\|_F^2 .$$

By choosing the value of λ' as

$$\lambda' = \sqrt{\frac{2C_2^2 \beta A^2 \epsilon(M^*)}{\|M^*\|_F^2 N} + \frac{2C_3^2}{\|M^*\|_F^2 N^2}} ,$$

we can get

$$\begin{aligned} \epsilon(\hat{M}) - \epsilon(M^*) &\leq \max\left\{ \frac{B+L}{N^2} + \frac{\beta A^2}{2N^4} + \lambda' \|M^*\|_F^2, 2\|M^*\|_F \sqrt{\frac{2C_2^2 \beta A^2 \epsilon(M^*)}{N} + \frac{2C_3^2}{N^2}} \right\} \\ &\lessapprox 2\|M^*\|_F \left(\sqrt{\frac{2C_2^2 \beta A^2 \epsilon(M^*)}{N}} + \frac{2C_3}{N} \right) . \end{aligned} \tag{8}$$

In Eq. 8, we neglect the fast rate term and focus on the r.h.s. part only with the possible slower rate. From the result, the convergence rate of the metric loss term has the order $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ as in previous analyses, which attaches the lower bound of similarity and distance metric learning for convex loss function. Since also with smooth loss assumption in our conditions, we can achieve a fast rate with order $\mathcal{O}\left(\frac{1}{N}\right)$ when $\epsilon(M^*) \to 0$. The improvement over loss function excess risk has been proved for the i.i.d. learning setting (Srebro et al. 2010). Both the faster and lower convergence rate for regularized objective and the loss term are validated in the experiments.

Remark 3 When the number of examples is large enough in Eq. 6, the convergence of the objective can achieve a $\mathcal{O}\left(\frac{1}{N^2}\right)$ given $\epsilon(M^*)$ is small enough. $\epsilon(M^*)$ measure the generalization performance of the optimal metric learned on the true risk, which reveals the difficulty of a particular task. For instance, as the realizability assumption in PAC learnability analysis (Shalev-Shwartz and Ben-David 2014), i.e., there exists an oracle hypothesis metric M^*

and all labels can be generated from it. This separable case leads to the condition $\epsilon(M^*) = 0$. On the other hand, for a problem that instances from different localities are easily separated, i.e., all required pairwise relationship generated from true data distribution can be easily satisfied by the best metric with large probability so that $\epsilon(M^*) \to 0$. Therefore, given enough training examples, the distance metric learning task can already perform well, and it is much better than the case with a small number of examples. The condition on the number of instances to get a faster rate indicates there may be an improvement over the generalization convergence rate with the increase of N. The precise number of N may be much smaller than the condition given in Eq. 6, which is also validated in the experiments.

Remark 4 The r.h.s. of the result is not related to the dimensionality of the data, which means it can be applied to even high dimensional cases. This dimension free property of distance metric learning generalization is also discussed in Verma and Branson (2015). Besides, higher dimension of metric induces larger hypothesis space, so that the optimal metric solution of $F(M)$ can perform better. Therefore, the value of $\epsilon(M^*)$ will be smaller in a high-dimensional case, which will prompt the condition for faster rates. This phenomenon has been validated in our experiment.

Remark 5 These two results are consistent with the improved rates in classification (Zhang et al. 2017) for i.i.d. examples, which shows that the learning process of distance metric learning can also get the same fast rates as in those classification tasks. In other words, learning based on the pairwise relationship can also obtain the same fast rate as the learning for just first order relationship.

Remark 6 The bounded condition of M, i.e., $\|M\|_F \leq R$, can be achieved by the Frobenius norm regularizer with bounded loss function some times. When $\Omega = \lambda\|M\|_F^2$, we have $\hat{F}(\hat{M}) + \lambda\|\hat{M}\|_F^2 \leq \hat{F}(0)$, which can be used to bound $\|\hat{M}\|_F$. Similarity, $\|M^*\|$ can also be bounded. Thus, we can restrict the discussion of M in a bounded domain.

Remark 7 In the theorem, the whole objective is required to be strongly convex *in expectation*, i.e., $F(M)$ is λ-strongly convex. This is a *weak* assumption w.r.t. requiring the empirical objective $\hat{F}(M)$ to be strongly convex for all input data (it is notable that in Theorem 1, we only require $\hat{F}(M)$ to be convex). In real applications, we can validate this assumption by checking the empirical objective. This weak condition extends our analysis to more real cases. An example with a strongly convex expected objective but a convex empirical objective can be found in the experiments.

4 Applications of the analysis of distance metric learning problems

It is notable that three essential properties of the expected objective $F(M)$ are required for the proof of the Theorem 1, namely the L-Lipschitz regularizer $\Omega(\cdot)$, the β-smooth loss function, and the λ-strongly convex objective. Several different types of loss functions and regularizers have been utilized in existing distance metric learning approach (Kulis 2012), and we will analyze the application of our theorem in existing methods in detail. By verifying function properties of regularizer (Lipschitz), loss (smooth), and objective (strongly convex), all the three conditions are easy to satisfy in a practical distance metric learning implementation.

For regularizer $\Omega(\cdot)$, it is required to be a Lipschitz function. This is a widely satisfied requirement among distance metric learning methods. For example, the commonly used (squared) Frobenius norm $\Omega_F(M) = \|M\|_F^2$, which is $2\|M\|_F$-Lipschitz in the domain. The

regularizer can also be the form $\Omega_T(M) = \mathrm{Tr}(MC)$, and C is a pre-defined matrix depended on real applications. When $C = I$, it requires the metric to be a low rank solution (Lim et al. 2013); while $C = \sum_{y_i=y_j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$, it is equivalent to minimize the distances between the same class instances, which is also used in Weinberger and Saul (2009). The Lipschitz parameter is related to the norm of C. The form $\Omega_L(M) = \mathrm{Tr}(M) - \log\det(M)$ is used in Davis et al. (2007), which intrinsically require the metric to be a PSD matrix. All these regularizers are Lipschitz in the PSD matrix set.

For the loss term $\ell(\cdot)$, hinge loss $\ell_h(x) = \max(1-x, 0)$ is often used to transform hard constraints for pairwise distances to soft ones (Shalev-Shwartz et al. 2004). In addition to requiring the input distance satisfy a similar or dissimilar distance constraint, it also preserve a margin 1 between the distance value and the threshold γ, which has been validated to improve the generalization ability in several cases (Weinberger and Saul 2009). The hinge loss is not a smooth function, since there are several possible sub-gradient directions at the input $x = 1$. Therefore, some smooth versions of it are also used, which can usually speed up the optimization (Beck and Teboulle 2009). For example, the squared hinge loss $\ell_s^1(x) = \max(1-x, 0)^2$, exponential smoothed hinge loss (Qian et al. 2013) $\ell_s^2(x) = \rho\log(1 + \exp(-\frac{1}{\rho}(x-1)))$, and smooth hinge loss (Qian et al. 2015) $\ell_s^3(x) = \max(1-x-\frac{\rho}{2}, 0)$ when $x \in (-\infty, 1-\rho)\cup(1, \infty)$, and $\ell_s^3(x) = \frac{1}{2\rho}(1-x)^2$ otherwise. The parameter $\rho \geq 0$ in $\ell_s^2(\cdot)$ and $\ell_s^3(\cdot)$ is to tune the smoothness of the loss function. The smaller the value of $\rho \to 0$, the larger the proportion the loss function closer to the hinge loss $\ell_h(\cdot)$, and the smaller the smoothness parameter β of it. Log loss $\ell_l(x) = \log(1+\exp(-x))$ also possesses the smooth property, which also acts as an important composition in distance metric training (Bian and Tao 2011). In addition, the square loss $\ell_2(x) = (x-1)^2$ satisfies the smooth assumption. In this case, the distance value of similar and dissimilar instances are required to be close to $\gamma - 1$ and $\gamma + 1$, respectively (we assume the value $\gamma - 1 > 0$).

The strongly convex requirement is also required on the *whole expected objective* $F(M)$. As in Remark 7, this condition can be checked by verifying strongly convexity of empirical objective $\hat{F}(M)$. It is noteworthy that in spite of the strongly convex property of the square loss $\ell_2(x)$, it cannot deduce the strongly convex property of the objective $\hat{F}(M)$ w.r.t. metric M directly. For example, when instances have low rank property, it is hard to obtain strongly convexity w.r.t. $\hat{F}(M)$.[1] Thus, a strong convex regularizer like the (squared) Frobenius norm $\Omega_F(M)$ is safe to ensure the strongly convex of the whole objective, just as that in the support vector machine. Besides, the strongly convex regularizer can be used together with other convex but not strongly convex regularizers to obtain strongly convex property and structure on metric at the same time. The larger the weight of $\Omega_F(M)$, the larger proportion the strongly convex property of the objective, and thus the faster convergence rate. Although the combined strategy helps to achieve the overall strongly convex property of the objective, it is not reasonable to set a large regularizer to achieve fast convergence rate, since it will also introduce a large *bias* to the empirical optimal solution of the loss function. The proportion of the bias can either be measured by the value of $|\Omega(\hat{M}) - \Omega(M^*)|$ or $\|\hat{M} - M^*\|_F$. Directly using $\Omega(M) = \|M\|_F^2$ implicitly assumes a zero metric as the prior, which may be far from the true optimal one. Therefore, an *iterative* procedure can be applied to sequentially estimate the optimal solution of the problem, i.e., learning the distance metric with multiple stages and assuming the stage-wise solutions provide and approach better optimal estimations. Set $M_{-1} = 0$ as the value of initial matrix, subsequent metrics can be learned based on

[1] In the experiment section, we give an example with strongly convex expected objective but only convex empirical objective, which is consistent with the requirement in Theorem 1.

$$F(M_g) = \min_{M \in \mathcal{S}_d^+} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \ell(q_{ij}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_i, \mathbf{x}_j))) + \lambda' \|M - M_{g-1}\|_F^2 .$$

Then the metric learned in the g-th stage reduce the bias towards optimal metric M^* by estimating it with the empirical optimal solution of the last stage. Similar stage-wise strategies have been proved to be effective in Weinberger and Saul (2009), Zhan et al. (2009) and Qian et al. (2015). Particularly in Qian et al. (2015), a smooth hinge loss $\ell_s^2(\cdot)$ is also used to satisfy the smooth condition.

5 Analysis on the ranking problem

In a ranking problem, there are N examples $\{\mathbf{z}_i = (\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$, but $y_i \in \mathbb{R}$ may be a real value reveals the preference order among different instances. The goal of ranking is to seek a scoring function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by minimizing the expected risk (Agarwal and Niyogi 2009; Rejchel 2012)

$$\Upsilon(f) = \mathbb{E}_{\mathbf{z}_1, \mathbf{z}_2}[\mathbb{I}[\mathrm{sign}(Y_1 - Y_2) f(\mathbf{x}_1, \mathbf{x}_2) < 0]] .$$

$\mathrm{sign}(\cdot) \in \{-1, 1\}$ shows sign of the input value. The expected risk involves a pair of examples, which is similar to the distance metric learning scenario. Its empirical counterpart can be formulated as

$$\hat{\Upsilon}(f) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j\neq i} \mathbb{I}[\mathrm{sign}(Y_i - Y_j) f(\mathbf{x}_i, \mathbf{x}_j) < 0] .$$

The indicator function $\mathbb{I}[\cdot]$ is often replaced as a convex surrogate $\ell(\cdot)$, when we implement f with linear predictor $\mathbf{w} \in \mathbb{R}^d$, we can redefine the expected risk and empirical risk of ranking problem as following:

$$\Upsilon(\mathbf{w}) = \mathbb{E}_{\mathbf{z}_1, \mathbf{z}_2}[\ell(\mathrm{sign}(Y_1 - Y_2)\mathbf{w}^\top(\mathbf{x}_1 - \mathbf{x}_2))] + \Omega(\mathbf{w}) . \tag{9}$$

$$\hat{\Upsilon}(\mathbf{w}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j\neq i} \ell(\mathrm{sign}(Y_i - Y_j)\mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j)) + \Omega(\mathbf{w}) . \tag{10}$$

$\Omega(\mathbf{w})$ is a convex regularizer on $\mathbf{w}$. The optimal solution of Eqs. 9 and 10 are $\mathbf{w}^*$ and $\hat{\mathbf{w}}$, respectively. Similar to the proof of Theorem 1, we can get the following theorem, which provides the relationship between $\Upsilon(\mathbf{w})$ and $\hat{\Upsilon}(\mathbf{w})$.

Theorem 2 *Assume $\Upsilon(\mathbf{w})$ is λ-strongly convex, $\hat{\Upsilon}(\mathbf{w})$ is convex function, loss function $\ell(\cdot)$ is β-smooth, and $\Omega(\mathbf{w})$ is L-Lipschitz. Define:*

$$\|\mathbf{w}\|_F \leq R, \ \|\mathbf{x}\| \leq A ,$$

$$B = \sup_{\mathbf{z}_i, \mathbf{z}_j \sim \mathcal{Z}} \|\nabla \ell(\mathrm{sign}(y_i - y_j)(\mathbf{w}^{*\top}(\mathbf{x}_i - \mathbf{x}_j)))\|_F ,$$

$$s = \lceil \log_2 2R + 2\log_2 N \rceil,$$

$$C_1 = 128\sqrt{2} + 32\sqrt{2\log s/\delta} , C_2 = 16\sqrt{2} + 16\sqrt{\log(s/\delta)} , C_3 = \frac{40B}{3}\log(s/\delta) .$$

For $0 < \delta < 1$, *with probability* $1 - 2\delta$:

$$\Upsilon(\hat{\mathbf{w}}) - \Upsilon(\mathbf{w}^*) \leq \max\left\{\frac{B+L}{N^2} + \frac{\beta A^2}{2N^4}, \frac{4C_1^2R^2\beta^2A^4}{\lambda N} + \frac{C_2^2}{\lambda N}\beta A^2\epsilon(\mathbf{w}^*) + \frac{2RC_3}{N}\right\}.$$

Furthermore, if

$$N \geq \frac{16\beta^2A^2C_1^2}{\lambda^2},$$

we have with probability $1 - 2\delta$:

$$\Upsilon(\hat{\mathbf{w}}) - \Upsilon(\mathbf{w}^*) \leq \max\left\{\frac{B+L}{N^2} + \frac{\beta A^2}{2N^4}, \frac{2C_2^2\beta A^2\epsilon(\mathbf{w}^*)}{\lambda N} + \frac{2C_3^2}{\lambda N^2}\right\}.$$

Remark 8 The analysis of the ranking problem has been researched from various perspectives (Clémençon et al. 2008; Agarwal and Niyogi 2009; Rejchel 2012). In the above Theorem 2, we utilize the property of loss function and regularizer to improve the generalization rate of a ranking problem. The convergence rate is related to the property of the objective function. In summary, the pairwise linear ranking problem and distance metric learning task have similar convergence results.

6 Proof of the Theorem

In this section, we present the proof details of Theorem 1 for distance metric learning problem. The Proof of Theorem 2 for ranking can be constructed similarly.

The main differences between the proof for distance metric learning objective and the classification task are two-fold. First, in the i.i.d. classification tasks, symmetrization technique (Bartlett and Mendelson 2002) is often used to introduce Rademacher random variables as a measure of hypothesis complexity. For the non-i.i.d. pairwise objective in distance metric learning, this symmetrization does not apply. U-statistics (Clémençon et al. 2008) is used to transform the original non-i.i.d. pairs to the sum of i.i.d. blocks. Second, a Bernstein-type inequality for vectors is used in Zhang et al. (2017) to bound the first order difference of excess risk on the optimal solution, which requires the input is the sum of i.i.d. random vectors. In our proof, a peeling strategy plus a Bernstein-type inequality for U-statistics is applied to the non-i.i.d. random variables, which helps achieve the same rate.

At first, we prove the *smoothness* property for the gradient of the loss function. The norm notation presents the Frobenius norm by default. For M and M', we have

$$\begin{aligned}
&\|\nabla\ell(q_{ij}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_i, \mathbf{x}_j))) - \nabla\ell(q_{ij}(\gamma - \mathrm{Dis}_{M'}^2(\mathbf{x}_i, \mathbf{x}_j)))\| \\
&= \|\ell'(q_{ij}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_i, \mathbf{x}_j)))(-q_{ij}A_{ij}) - \ell'(q_{ij}(\gamma - \mathrm{Dis}_{M'}^2(\mathbf{x}_i, \mathbf{x}_j)))(-q_{ij}A_{ij})\| \\
&= \|\ell'(q_{ij}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_i, \mathbf{x}_j))) - \ell'(q_{ij}(\gamma - \mathrm{Dis}_{M'}^2(\mathbf{x}_i, \mathbf{x}_j)))\|\|A_{ij}\| \\
&\leq \beta|q_{ij}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_i, \mathbf{x}_j)) - q_{ij}(\gamma - \mathrm{Dis}_{M'}^2(\mathbf{x}_i, \mathbf{x}_j))|A \\
&= \beta A|\mathrm{Dis}_M^2(\mathbf{x}_i, \mathbf{x}_j) - \mathrm{Dis}_{M'}^2(\mathbf{x}_i, \mathbf{x}_j)| \\
&= \beta A|\langle A_{ij}, M - M'\rangle| \\
&\leq \beta A^2\|M - M'\| \qquad (11)
\end{aligned}$$

Inequality Eq. 11 comes from the smooth property of loss function $\ell(\cdot)$, which means its differential is β-Lipschitz. From the above proof, the gradient of loss function $\nabla\ell(\cdot)$ is still a smooth function over M, and its smooth parameter is βA^2.

Proof of Theorem 1 Since all possible metrics are restricted in the domain $\|M-M^*\|_F \leq 2R$, we consider to split this range into two parts based on the proportion of the solution's closeness to the optimal one. Besides, we use $\Delta M = M - M^*$ and $\Delta\hat{M} = \hat{M} - M^*$ to simplify notations in some of the following derivations. In the relative easy case $0 \leq \|\Delta M\|_F = \|M-M^*\|_F \leq \frac{1}{N^2}$, we have the following bound for the excess risk

$$\begin{aligned} F(\hat{M}) - F(M^*) &= \epsilon(\hat{M}) - \epsilon(M^*) + \Omega(\hat{M}) - \Omega(M^*) \\ &\leq \langle \hat{M} - M^*, \nabla\epsilon(M^*)\rangle + \frac{A^2\beta}{2}\|\hat{M} - M^*\|^2 + L\|\hat{M} - M^*\| \qquad (12) \\ &\leq \|\hat{M} - M^*\|\|\nabla\epsilon(M^*)\| + \frac{A^2\beta}{2}\frac{1}{N^4} + L\frac{1}{N^2} \\ &\leq \frac{B}{N^2} + \frac{A^2\beta}{2}\frac{1}{N^4} = \frac{B+L}{N^2} + \frac{A^2\beta}{2N^4} \qquad (13) \end{aligned}$$

In Eq. 12, we use the smooth property of the loss function $\epsilon(\cdot)$ and the L-Lipschitz property of the regularizer $\Omega(\cdot)$ w.r.t. metric M. While in Eq. 13, we can bound the norm of $\|\nabla\epsilon(M^*)\|$ as

$$\begin{aligned} \|\nabla\epsilon(M^*)\| &= \|\mathbb{E}_{\mathbf{z}_i,\mathbf{z}_j\sim\mathcal{Z}}\left[\nabla\ell(q_{ij}(\gamma - \mathrm{Dis}^2_{M^*}(\mathbf{x}_i,\mathbf{x}_j)))\right]\| \\ &\leq \mathbb{E}_{\mathbf{z}_i,\mathbf{z}_j\sim\mathcal{Z}}\left[\|\nabla\ell(q_{ij}(\gamma - \mathrm{Dis}^2_{M^*}(\mathbf{x}_i,\mathbf{x}_j)))\|\right] \leq B\ . \end{aligned}$$

When $\frac{1}{N^2} \leq \|\Delta M\|_F = \|M - M^*\|_F \leq 2R$, we follow Zhang et al. (2017) to get the upper bound of the excess risk:

$$\begin{aligned} & F(\hat{M}) - F(M^*) + \frac{\lambda}{2}\|\Delta\hat{M}\|^2 \leq \langle\nabla F(\hat{M}), \Delta\hat{M}\rangle \\ &= \langle\nabla F(\hat{M}) - \nabla F(M^*), \Delta\hat{M}\rangle + \langle\nabla F(\hat{M}), \Delta\hat{M}\rangle \\ &= \langle\nabla F(\hat{M}) - \nabla F(M^*) - [\nabla\hat{F}(\hat{M}) - \nabla\hat{F}(M^*)], \Delta\hat{M}\rangle \\ &\quad + \langle[\nabla\hat{F}(\hat{M}) - \nabla\hat{F}(M^*)] + \nabla F(\hat{M}), \Delta\hat{M}\rangle \\ &\leq \langle\nabla F(\hat{M}) - \nabla F(M^*) - [\nabla\hat{F}(\hat{M}) - \nabla\hat{F}(M^*)], \Delta\hat{M}\rangle + \langle\nabla F(M^*) - \nabla\hat{F}(M^*), \Delta\hat{M}\rangle \qquad (14) \\ &= \langle\nabla\epsilon(\hat{M}) - \nabla\epsilon(M^*) - [\nabla\epsilon_N(\hat{M}) - \nabla\epsilon_N(M^*)], \Delta\hat{M}\rangle + \langle\nabla\epsilon(M^*) - \nabla\epsilon_N(M^*), \Delta\hat{M}\rangle \\ &\leq \sup_{\|\Delta M\|\leq\|\Delta\hat{M}\|} \langle\nabla\epsilon(M) - \nabla\epsilon(M^*) - [\nabla\epsilon_N(M) - \nabla\epsilon_N(M^*)], \Delta M\rangle \qquad (15) \\ &\quad + \sup_{\|\Delta M\|\leq\|\Delta\hat{M}\|} \langle\nabla\epsilon(M^*) - \nabla\epsilon_N(M^*), \Delta M\rangle \qquad (16) \end{aligned}$$

Equation 14 comes from the variational inequality (Boyd and Vandenberghe 2004) of the convex optimal solution $\forall M, \langle\nabla\hat{F}(\hat{M}), M - \hat{M}\rangle \geq 0$. To bound the above inequalities in Eqs. 15 and 16, we partition the range of $\|\Delta M\| = \|M - M^*\|$ into $s = \lceil\log_2 2R + 2\log_2 N\rceil$ segments such that

$$\Delta_k = \left(\frac{2^{k-1}}{N^2}, \frac{2^k}{N^2}\right] = (r_k^-, r_k^+],\ k = 1, \ldots, s\ .$$

Benefited from the first kind of range partition, there are only *finite s* segments during the current case, thus the theoretical analysis over one segment could be extended to the whole range by union bound. The decomposition of intervals help to get tighter results. In addition, it could introduce the variable $\|\hat{M}-M^*\|$ on the r.h.s. of the bound. For a particular $\|M-M^*\| \leq r$, the upper bound value r must lies in a certain segment $r \in \Delta_k$. In the following, we bound Eqs. 15 and 16 separately by the following two lemmas:

Lemma 1 *Given conditions in Theorem 1 and* $\frac{1}{N^2} \leq \|\Delta M\|_F \leq 2R$,

$$\sup_{\|\Delta M\| \leq \|\Delta \hat{M}\|} \langle \nabla \epsilon(M) - \nabla \epsilon(M^*) - [\nabla \epsilon_N(M) - \nabla \epsilon_N(M^*)], \Delta M \rangle \leq \frac{\beta A^2 \|\Delta \hat{M}\|^2}{\sqrt{N}} C_1. \tag{17}$$

where $C_1 = 16\sqrt{2} + 8\sqrt{2 \log s/\delta}$.

Lemma 2 *Given conditions in Theorem 1 and* $\frac{1}{N^2} \leq \|\Delta M\|_F \leq 2R$,

$$\sup_{\|\Delta M\| \leq \|\Delta \hat{M}\|} \langle \nabla \epsilon(M^*) - \nabla \epsilon_N(M^*), \Delta M \rangle \leq \frac{\|\Delta \hat{M}\|}{\sqrt{N}} \sqrt{\beta A^2 \epsilon(M^*)} C_2 + \frac{\|\Delta \hat{M}\| C_3}{N}, \tag{18}$$

where $C_2 = 8\sqrt{2} + 8\sqrt{\log(s/\delta)}$ *and* $C_3 = \frac{40B}{3} \log(s/\delta)$.

In summary, by Eqs. 17 and 18, we have

$$\begin{aligned} & F(\hat{M}) - F(M^*) + \frac{\lambda}{2} \|\hat{M} - M^*\|^2 \\ & \leq C_1 \|\hat{M} - M^*\|^2 \frac{\beta A^2}{\sqrt{N}} + \|\hat{M} - M^*\| \frac{C_2}{\sqrt{N}} \sqrt{\beta A^2 \epsilon(M^*)} + \|\hat{M} - M^*\| \frac{C_3}{N} \end{aligned} \tag{19}$$

For the first two terms, we have:

$$C_1 \|\hat{M} - M^*\|^2 \frac{\beta A^2}{\sqrt{N}} \leq C_1^2 \|\hat{M} - M^*\|^2 \frac{\beta^2 A^4}{\lambda N} + \frac{\lambda}{4} \|\hat{M} - M^*\|^2, \tag{20}$$

and

$$\|\hat{M} - M^*\| \frac{C_2}{\sqrt{N}} \sqrt{\beta A^2 \epsilon(M^*)} \leq \frac{C_2^2}{\lambda N} \beta A^2 \epsilon(M^*) + \frac{\lambda}{4} \|\hat{M} - M^*\|^2. \tag{21}$$

Plugging Eqs. 20 and 21 into Eq. 19, we have the first result:

$$\begin{aligned} & F(\hat{M}) - F(M^*) \leq C_1^2 \|\hat{M} - M^*\|^2 \frac{\beta^2 A^4}{\lambda N} + \frac{C_2^2}{\lambda N} \beta A^2 \epsilon(M^*) + \|\hat{M} - M^*\| \frac{C_3}{N} \\ & \leq \frac{4 C_1^2 R^2 \beta^2 A^4}{\lambda N} + \frac{C_2^2}{\lambda N} \beta A^2 \epsilon(M^*) + \frac{2RC_3}{N} \sim \mathcal{O}\left(\frac{1}{N}\right). \end{aligned}$$

In addition, for the last two terms in Eq. 19, we can bound by

$$\|\hat{M} - M^*\| \frac{C_2}{\sqrt{N}} \sqrt{\beta A^2 \epsilon(M^*)} \leq \frac{2 C_2^2 \beta A^2 \epsilon(M^*)}{\lambda N} + \frac{\lambda}{8} \|\hat{M} - M^*\|^2. \tag{22}$$

and

$$\|\hat{M} - M^*\| \frac{C_3}{N} \leq \frac{2 C_3^2}{\lambda N^2} + \frac{\lambda}{8} \|\hat{M} - M^*\|^2. \tag{23}$$

If we have the condition

$$C_1\|\hat{M}-M^*\|^2\frac{\beta A^2}{\sqrt{N}} \leq \frac{\lambda\|\hat{M}-M^*\|^2}{4} ,$$

which means

$$N \geq (\frac{4\beta A^2 C_1}{\lambda})^2 = \frac{16\beta^2 A^4 C_1^2}{\lambda^2} ,$$

with Eqs. 22 and 23, we have

$$F(\hat{M})-F(M^*) \leq \frac{2C_2^2\beta A^2\epsilon(M^*)}{\lambda N}+\frac{2C_3^2}{\lambda N^2} .$$

□

Proof of Lemma 1 To prove an upper bound of the Eq. 15, we consider a particular $\|M - M^*\| = \|\Delta M\| \leq r$. Therefore, it equals to prove

$$\begin{aligned}
&\sup_{\|\Delta M\|\leq r} \langle\nabla\epsilon(M)-\nabla\epsilon(M^*)-[\nabla\epsilon_N(M)-\nabla\epsilon_N(M^*)],\Delta M\rangle \\
\leq &\sup_{\|\Delta M\|\leq r_k^+} \langle\nabla\epsilon(M)-\nabla\epsilon(M^*)-[\nabla\epsilon_N(M)-\nabla\epsilon_N(M^*)],\Delta M\rangle \\
= &\sup_{\|\Delta M\|\leq r_k^+} g(\mathbf{z}_1,\ldots,\mathbf{z}_N) .
\end{aligned}$$

which can be further bounded by McDiarmid inequality (McDiarmid 1989). Bounded difference condition is required to use this inequality, which means the maximum absolute change of the objective value when a particular instance in the input is changed. For two set of datasets $\mathcal{D} = \{\mathbf{z}_1,\ldots,\mathbf{z}_k,\ldots,\mathbf{z}_N\}$ and $\mathcal{D}' = \{\mathbf{z}_1,\ldots,\mathbf{z}'_k,\ldots,\mathbf{z}_N\}$, the notation $(\mathbf{z}_i,\mathbf{z}_j)\sim\mathcal{D}$ means a set of $P = N(N-1)$ pairs of examples sampled from dataset $\mathcal{D}$. We have

$$\begin{aligned}
&\sup_{D,D'}|g(\mathbf{z}_1,\ldots,\mathbf{z}_k,\ldots,\mathbf{z}_N)-g(\mathbf{z}_1,\ldots,\mathbf{z}'_k,\ldots,\mathbf{z}_N)| \\
=&\sup_{D,D'}|\frac{1}{P}\sum_{(\mathbf{z}_i,\mathbf{z}_j)\sim\mathcal{D}}\langle\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_M^2(\mathbf{z}_i,\mathbf{z}_j)))-\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_{M^*}^2(\mathbf{z}_i,\mathbf{z}_j))),\Delta M\rangle \\
&-\frac{1}{P}\sum_{(\mathbf{z}_i,\mathbf{z}_j)\sim\mathcal{D}'}\langle\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_M^2(\mathbf{z}_i,\mathbf{z}_j)))-\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_{M^*}^2(\mathbf{z}_i,\mathbf{z}_j))),\Delta M\rangle| \\
=&\sup_{D,D'}|\frac{2}{P}\sum_{j\neq k}\langle\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_M^2(\mathbf{z}_k,\mathbf{z}_j)))-\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_{M^*}^2(\mathbf{z}_k,\mathbf{z}_j))),\Delta M\rangle \\
&-\frac{2}{P}\sum_{j\neq k}\langle\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_M^2(\mathbf{z}'_k,\mathbf{z}_j)))-\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_{M^*}^2(\mathbf{z}'_k,\mathbf{z}_j))),\Delta M\rangle| \\
\leq&\sup_{D,D'}\frac{2}{P}\sum_{j\neq k}\|\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_M^2(\mathbf{z}_k,\mathbf{z}_j)))-\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_{M^*}^2(\mathbf{z}_k,\mathbf{z}_j)))\|\|\Delta M\| \\
&+\sum_{j\neq k}\|\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_M^2(\mathbf{z}'_k,\mathbf{z}_j)))-\nabla\ell(q_{ij}(\gamma-\mathrm{Dis}_{M^*}^2(\mathbf{z}'_k,\mathbf{z}_j)))\|\|\Delta M\| \\
\leq&\frac{4}{N}\beta A^2\|\Delta M\|^2 \leq \frac{4\beta A^2}{N}(r_k^+)^2 .
\end{aligned}$$

Next step is to bound the expectation of g, which can be often dealt with by Rademacher Complexity. For distance metric learning, however, since the given pairs are no longer i.i.d., it is hard to use symmetrization technology and introduce Rademacher random variables. Benefited from the following lemma (Clémençon et al. 2008), expectation over sum of pairwise examples can be upper bounded by their sum of i.i.d. blocks.

Lemma 3 (Clémençon et al. 2008) *Let $q_\tau : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be real-valued functions indexed by $\tau \in T$ where T is some set. If $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are i.i.d. then for any convex nondecreasing function Ψ,*

$$\mathbb{E}\Psi\left(\sup_{\tau\in T}\frac{1}{P}\sum_{i\neq j}q_\tau(\mathbf{z}_i,\mathbf{z}_j)\right)\leq\mathbb{E}\Psi\left(\sup_{\tau\in T}\frac{1}{\lfloor\frac{N}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{N}{2}\rfloor}q_\tau(\mathbf{z}_i,\mathbf{z}_{i+\lfloor\frac{N}{2}\rfloor})\right),$$

assuming the supreme are measurable and the expected values exist.

In the following, we use the notation, $n = \lfloor\frac{N}{2}\rfloor$. $\mathbb{E}_{\mathbf{z}}[\cdot]$ and $\mathbb{E}_{\mathbf{z}'}[\cdot]$ mean the expectation is taken w.r.t. the data $\mathcal{D}$ and $\mathcal{D}'$, respectively.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left[\sup_{\|\Delta M\|\leq r_k^+}\langle\nabla\epsilon(M)-\nabla\epsilon(M^*)-[\nabla\epsilon_N(M)-\nabla\epsilon_N(M^*)],\Delta M\rangle\right]\\
&=\mathbb{E}_{\mathbf{z}}\left[\sup_{\|\Delta M\|\leq r_k^+}\langle\nabla\epsilon(M)-\nabla\epsilon(M^*)-\frac{1}{P}\sum_{i=1}^{N}\sum_{j\neq i}[\nabla\ell(M,\mathbf{z}_i,\mathbf{z}_j)-\nabla\ell(M^*,\mathbf{z}_i,\mathbf{z}_j)],\Delta M\rangle\right]\\
&=\mathbb{E}_{\mathbf{z}}\left[\sup_{\|\Delta M\|\leq r_k^+}\frac{1}{P}\sum_{i=1}^{N}\sum_{j\neq i}\langle\left(\nabla\epsilon(M)-\nabla\epsilon(M^*)\right)-\left(\nabla\ell(M,\mathbf{z}_i,\mathbf{z}_j)-\nabla\ell(M^*,\mathbf{z}_i,\mathbf{z}_j)\right),\Delta M\rangle\right]
\end{aligned}$$

Define a equivalent form of loss function as $\bar{\epsilon}_n(M) = \sum_{i=1}^{n}\ell(M,\mathbf{z}_i,\mathbf{z}_{i+n}) = \sum_{i=1}^{n}\ell(q_{i,i+n}(\gamma-\mathrm{Dis}_M^2(\mathbf{z}_i,\mathbf{z}_{i+n})))$, we can upper bound the above expectation by Lemma 3 with

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\left[\sup_{\|\Delta M\|\leq r_k^+}\langle\nabla\epsilon(M)-\nabla\epsilon(M^*)-[\nabla\epsilon_N(M)-\nabla\epsilon_N(M^*)],\Delta M\rangle\right]\\
&\leq\mathbb{E}_{\mathbf{z}}\left[\sup_{\|\Delta M\|\leq r_k^+}\frac{1}{n}\sum_{i=1}^{n}\langle\left(\nabla\epsilon(M)-\nabla\epsilon(M^*)\right)-\left(\nabla\ell(M,\mathbf{z}_i,\mathbf{z}_{i+n})-\nabla\ell(M^*,\mathbf{z}_i,\mathbf{z}_{i+n})\right),\Delta M\rangle\right]\\
&=\mathbb{E}_{\mathbf{z}}\left[\sup_{\|\Delta M\|\leq r_k^+}\frac{1}{n}\sum_{i=1}^{n}\langle\nabla\epsilon(M)-\nabla\epsilon(M^*),\Delta M\rangle-\langle\nabla\ell(M,\mathbf{z}_i,\mathbf{z}_{i+n})-\nabla\ell(M^*,\mathbf{z}_i,\mathbf{z}_{i+n}),\Delta M\rangle\right]\\
&=\mathbb{E}_{\mathbf{z}}\left[\sup_{\|\Delta M\|\leq r_k^+}\mathbb{E}_{\mathbf{z}'}[\langle\nabla\bar{\epsilon}_n(M)-\nabla\bar{\epsilon}_n(M^*),\Delta M\rangle]-\langle\nabla\bar{\epsilon}_n(M)-\nabla\bar{\epsilon}_n(M^*),\Delta M\rangle\right]\\
&\leq 2\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\leq r_k^+}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle\nabla\ell(M,\mathbf{z}_i,\mathbf{z}_{i+n})-\nabla\ell(M^*,\mathbf{z}_i,\mathbf{z}_{i+n}),\Delta M\rangle\right] \qquad (24)\\
&=\frac{2}{n}\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\leq r_k^+}\sum_{i=1}^{n}\sigma_i(\ell'(M,\mathbf{z}_i,\mathbf{z}_{i+n})-\nabla\ell'(M^*,\mathbf{z}_i,\mathbf{z}_{i+n}))\langle A_{i,i+n},\Delta M\rangle\right] \qquad (25)
\end{aligned}$$

In Eq. 24, we use Jensen's inequality to combine two expectations, and then introduce Rademacher random variables $\sigma_i \in \{-1, 1\}$ with equal probability for each block. Then in Eq. 25, it comes from that the σ_i and $\sigma_i q_{i,i+\frac{n}{2}}$ have the same distribution.

Here, the goal transform to bound the r.h.s. of Eq. 25, which is a multiplication of two related terms. It can be decoupled by the similar method in Zhang et al. (2017), we list the remaining steps here for completeness. Define $u_i = u_i(M) = \frac{1}{\sqrt{\beta}}(\ell'(M, \mathbf{z}_i, \mathbf{z}_{i+n}) - \ell'(M^*, \mathbf{z}_i, \mathbf{z}_{i+n})) \in [-\sqrt{\beta}Ar_k^+, \sqrt{\beta}Ar_k^+]$ and $v_i = v_i(M) = \sqrt{\beta}\langle A_{i,i+n}, \Delta M\rangle \in [-\sqrt{\beta}Ar_k^+, \sqrt{\beta}Ar_k^+]$. With $u_i v_i = \frac{1}{4}((u_i + v_i)^2 - (u_i - v_i)^2)$, we can get

$$\begin{aligned}
&\frac{2}{n}\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i(\ell'(M, \mathbf{z}_i, \mathbf{z}_{i+n}) - \nabla\ell'(M^*, \mathbf{z}_i, \mathbf{z}_{i+n}))\langle A_{i,i+n}, \Delta M\rangle\right] \\
&= \frac{2}{n}\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i u_i(M)v_i(M)\right] \\
&\le \frac{1}{2n}\left(\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i(u_i + v_i)^2\right] + \mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i(u_i - v_i)^2\right]\right).
\end{aligned}$$

With the following comparison lemma:

Lemma 4 (Meir and Zhang 2003) *Let $\{g_i(\theta)\}$ and $\{h_i(\theta)\}$ be sets of functions defined for all θ in some domain Θ. If for all i, θ, θ', $|g_i(\theta) - g_i(\theta')| \le |h_i(\theta) - h_i(\theta')|$, then*

$$\mathbb{E}_\sigma\left\{\sup_{\theta\in\Theta}\sum_{i=1}^{N}\sigma_i g_i(\theta)\right\} \le \mathbb{E}_\sigma\left\{\sup_{\theta\in\Theta}\sum_{i=1}^{N}\sigma_i h_i(\theta)\right\},$$

we can first remove the square on above equalities by the $2\sqrt{\beta}Ar_k^+$-Lipschitz of $(\cdot)^2$ in the range of $[-\sqrt{\beta}Ar_k^+, \sqrt{\beta}Ar_k^+]$:

$$\begin{aligned}
&\frac{1}{2n}\left(\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i(u_i + v_i)^2\right] + \mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i(u_i - v_i)^2\right]\right) \\
&\le \frac{\sqrt{\beta}Ar_k^+}{n}\left(\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i(u_i + v_i)\right] + \mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i(u_i - v_i)\right]\right) \\
&= \frac{2\sqrt{\beta}Ar_k^+}{n}\left(\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i u_i\right] + \mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i v_i\right]\right).
\end{aligned}$$

By the β-smooth of $\ell'(\cdot)$, we have the property of $u_i(M)$ that

$$\begin{aligned}
&|u_i(M) - u_i(M')| \le \frac{1}{\sqrt{\beta}}|\ell'(M, \mathbf{z}_i, \mathbf{z}_{i+n}) - \ell'(M', \mathbf{z}_i, \mathbf{z}_{i+n})| \le \sqrt{\beta}\|M - M'\| \\
&= \sqrt{\beta}|\langle A_{i,i+n}, \Delta M\rangle - \langle A_{i,i+n}, M' - M^*\rangle| = |v_i(M) - v_i(M')|.
\end{aligned}$$

Thus, by comparison Lemma 4, we have

$$\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i u_i\right] \le \mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i v_i\right].$$

For the r.h.s., it can be bounded by:

$$\mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i v_i\right] = \mathbb{E}_{\mathbf{z},\sigma}\left[\sup_{\|\Delta M\|\le r_k^+}\sum_{i=1}^{n}\sigma_i\sqrt{\beta}\langle A_{i,i+n}, \Delta M\rangle\right]$$

$$\le \sqrt{\beta}r_k^+\mathbb{E}_{\mathbf{z},\sigma}\left[\|\sum_{i=1}^{n}\sigma_i A_{i,i+n}\|\right] \le \sqrt{\beta}r_k^+ A\left\|\sqrt{\mathbb{E}_{\mathbf{z},\sigma}(\sum_{i=1}^{n}\sigma_i)^2}\right\| \le r_k^+ A\sqrt{\beta}\sqrt{n}\,. \quad (26)$$

The last inequality Eq. 26 comes the concave property of the $\sqrt{\cdot}$ function, then the i.i.d. property of Rademacher random variable σ_i. Thus, we can get

$$\mathbb{E}_{\mathbf{z}}\left[\sup_{\|\Delta M\|\le r_k^+}\langle\nabla\epsilon(M) - \nabla\epsilon(M^*) - [\nabla\epsilon_N(M) - \nabla\epsilon_N(M^*)], \Delta M\rangle\right]$$

$$= \frac{2}{n}\left(\frac{2r_k^+ A\sqrt{\beta}}{n}\right)(r_k^+ A\sqrt{\beta}\sqrt{n}) = \frac{4(r_k^+)^2\beta A^2}{\sqrt{n}} = \frac{4\sqrt{2}(r_k^+)^2\beta A^2}{\sqrt{N}}\,. \quad (27)$$

Combine above results in McDiarmid inequality, we have

$$\sup_{\|\Delta M\|\le r_k^+}\langle\nabla\epsilon(M) - \nabla\epsilon(M^*) - [\nabla\epsilon_N(M) - \nabla\epsilon_N(M^*)], \Delta M\rangle$$

$$\le \frac{4\sqrt{2}(r_k^+)^2\beta A^2}{\sqrt{N}} + 2(r_k^+)^2\beta A^2\sqrt{\frac{2\log 1/\delta}{N}} = \frac{\beta A^2(r_k^+)^2}{\sqrt{N}}(4\sqrt{2} + 2\sqrt{2\log 1/\delta})\,.$$

Then, with $r_k^+ \le 2r$, and combine results of s segments together with union bound, we have:

$$\sup_{\|\Delta M\|\le r}\langle\nabla\epsilon(M) - \nabla\epsilon(M^*) - [\nabla\epsilon_N(M) - \nabla\epsilon_N(M^*)], \Delta M\rangle \quad (28)$$

$$\le \frac{4\sqrt{2}(r_k^+)^2\beta A^2}{\sqrt{N}} + 2(r_k^+)^2\beta A^2\sqrt{\frac{2\log s/\delta}{N}} = \frac{\beta A^2 r^2}{\sqrt{N}}C_1\,. \quad (29)$$

where $C_1 = 16\sqrt{2} + 8\sqrt{2\log s/\delta}$. □

Proof of Lemma 2 To bound the term $\sup_{\|\Delta M\|\le\|\hat{M}-M^*\|}\langle\nabla\epsilon(M^*) - \nabla\epsilon_N(M^*), \Delta M\rangle$, we also consider the impact of segmentation and bound $\sup_{\|\Delta M\|\le r_k^+}\langle\nabla\epsilon(M^*) - \nabla\epsilon_N(M^*), \Delta M\rangle$. We can use the following Bosquet type inequality:

Theorem 3 (Rejchel 2015) *Assume $\mathcal{G}$ is a subset of family of functions $\{g : \mathcal{Z}\times\mathcal{Z}\to\mathbb{R}\}$ that are uniformly bounded by a constant G. Denote a U-Statistic with kernel g by $U_n(g) = \frac{1}{N(N-1)}\sum_{i\ne j} g(\mathbf{z}_i, \mathbf{z}_j)$ and its expectation by $U(g) = \mathbb{E}g(\mathbf{z}_1, \mathbf{z}_2)$. Let the $W = \sup_{g\in\mathcal{G}}|U_n(g) - U(g)|$ be the supreme of a centered U-process $U_n(g)$ and $\Sigma^2 = \sup_{g\in\mathcal{G}}\mathrm{Var}[g(\mathbf{z}_1, \mathbf{z}_2)]$. Define a simpler form of the empirical process*

$$T = \sup_{g\in\mathcal{G}}\frac{1}{\lfloor\frac{N}{2}\rfloor}|\sum_{i=1}^{\lfloor\frac{N}{2}\rfloor} g(\mathbf{z}_i, \mathbf{z}_{i+\lfloor\frac{N}{2}\rfloor}) - \mathbb{E}g(\mathbf{z}_i, \mathbf{z}_{i+\lfloor\frac{N}{2}\rfloor})|\,.$$

Then, for every non-negative confidence δ, with probability at least $1-\delta$ we have:

$$W \le \mathbb{E}T + \sqrt{\frac{2\log(1/\delta)\Sigma^2}{\lfloor\frac{N}{2}\rfloor}} + \frac{10G\log(1/\delta)}{3\lfloor\frac{N}{2}\rfloor}\,. \quad (30)$$

We can set $g(\mathbf{z}_i, \mathbf{z}_j) = \langle \nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_j), \Delta M \rangle$. The upper bound of g can be easily obtained by

$$\begin{aligned} G = \sup |g(\mathbf{z}_i, \mathbf{z}_j)| &= \sup_{\|\Delta M\| \le r_k^+} |\langle \nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_j), \Delta M \rangle| \\ &\le r_k^+ \|\nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_j)\| \le B r_k^+ . \end{aligned}$$

For the variance term Σ^2, we have

$$\begin{aligned} \Sigma^2 &= \sup_{g \in \mathcal{G}} \mathrm{Var}[g(\mathbf{z}_1, \mathbf{z}_2)] \le \sup_{g \in \mathcal{G}} \mathbb{E}[g(\mathbf{z}_1, \mathbf{z}_2)^2] \\ &= \sup_{\|\Delta M\| \le r_k^+} \mathbb{E}[\langle \nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_j), \Delta M \rangle^2] \\ &\le \sup_{\|\Delta M\| \le r_k^+} \mathbb{E}[\|\nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_j)\|^2 \|\Delta M\|^2] \\ &\le (r_k^+)^2 \mathbb{E}[\|\nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_j)\|^2] \\ &\le 4(r_k^+)^2 \beta A^2 \mathbb{E}[\ell(M^*, \mathbf{z}_i, \mathbf{z}_j)] = 4(r_k^+)^2 \beta A^2 \epsilon(M^*) . \end{aligned} \tag{31}$$

The inequality in Eq. 31 comes from the βA^2-smooth property of $\nabla \ell(\cdot)$ function. For the expectation term, we have:

$$\begin{aligned} \mathbb{E}[T] &= \mathbb{E}_{\mathbf{z}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} | \sum_{i=1}^{n} g(\mathbf{z}_i, \mathbf{z}_{i+n}) - \mathbb{E}_{\mathbf{z}'} g(\mathbf{z}'_i, \mathbf{z}'_{i+n}) | \right] \\ &\le \mathbb{E}_{\mathbf{z}, \mathbf{z}'} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} | \sum_{i=1}^{n} g(\mathbf{z}_i, \mathbf{z}_{i+n}) - g(\mathbf{z}'_i, \mathbf{z}'_{i+n}) | \right] \\ &= \frac{2}{n} \mathbb{E}_{\mathbf{z}, \sigma} \left[\sup_{g \in \mathcal{G}} | \sum_{i=1}^{n} \sigma_i g(\mathbf{z}_i, \mathbf{z}_{i+n}) | \right] \\ &= \frac{2}{n} \mathbb{E}_{\mathbf{z}, \sigma} \left[\sup_{\|\Delta M\| \le r_k^+} | \sum_{i=1}^{n} \sigma_i \langle \nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_{i+n}), \Delta M \rangle | \right] \\ &\le \frac{2}{n} \mathbb{E}_{\mathbf{z}, \sigma} \left[\sup_{\|\Delta M\| \le r_k^+} \| \sum_{i=1}^{n} \sigma_i \nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_{i+n}) \| \|\Delta M\| \right] \\ &\le \frac{2 r_k^+}{n} \mathbb{E}_{\mathbf{z}, \sigma} \left[\| \sum_{i=1}^{n} \sigma_i \nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_{i+n}) \| \right] \\ &\le \frac{2 r_k^+}{n} \sqrt{\mathbb{E}_{\mathbf{z}, \sigma} \left[\| \sum_{i=1}^{n} \sigma_i \nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_{i+n}) \|^2 \right]} \\ &\le \frac{2 r_k^+}{n} \sqrt{\sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}} \left[\|\nabla \ell(M^*, \mathbf{z}_i, \mathbf{z}_{i+n})\|^2 \right]} \\ &\le \frac{2 r_k^+}{n} \sqrt{4 n \beta A^2 \epsilon(M^*)} = \frac{4 r_k^+}{\sqrt{n}} \sqrt{\beta A^2 \epsilon(M^*)} . \end{aligned}$$

Putting all components together, we can get

$$\begin{aligned}
&\sup_{\|\Delta M\| \le \|\hat{M}-M^*\|} \langle \nabla \epsilon(M^*) - \nabla \epsilon_N(M^*), \Delta M \rangle \\
&\le \frac{4r_k^+}{\sqrt{n}} \sqrt{\beta A^2 \epsilon(M^*)} + \sqrt{\frac{2\log(1/\delta) 4 (r_k^+)^2 \beta A^2 \epsilon(M^*)}{n}} + \frac{10 B r_k^+ \log(1/\delta)}{3n} \\
&= \frac{4\sqrt{2} r_k^+}{\sqrt{N}} \sqrt{\beta A^2 \epsilon(M^*)} + 4 r_k^+ \sqrt{\frac{\log(1/\delta) \beta A^2 \epsilon(M^*)}{N}} + \frac{20 B r_k^+ \log(1/\delta)}{3N} .
\end{aligned}$$

Combining s segments together, we have:

$$\begin{aligned}
&\sup_{\|\Delta M\| \le r} \langle \nabla \epsilon(M^*) - \nabla \epsilon_N(M^*), \Delta M \rangle \\
&\le \frac{8\sqrt{2} r}{\sqrt{N}} \sqrt{\beta A^2 \epsilon(M^*)} + 8r \sqrt{\frac{\log(s/\delta) \beta A^2 \epsilon(M^*)}{N}} + \frac{40 B r \log(s/\delta)}{3N} \\
&= \frac{r}{\sqrt{N}} \sqrt{\beta A^2 \epsilon(M^*)} C_2 + \frac{r C_3}{N} , \qquad (32)
\end{aligned}$$

where $C_2 = 8\sqrt{2} + 8\sqrt{\log(s/\delta)}$ and $C_3 = \frac{40B}{3} \log(s/\delta)$. □

7 Experiments

In this section, we validate the generalization convergence rate of a distance metric learning problem proposed in this paper on a two-class synthetic dataset. Various properties of the proposed theory can be observed from the experimental results.

Square loss is used to implement $\ell(\cdot)$, and (squared) Frobenius norm $\|M\|_F^2$ serves as the regularizer. Thus, the objective possesses both the strongly convex and smooth requirements. Therefore, the empirical objective of a metric M in this case is:

$$\hat{F}(M) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (q_{ij}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_i, \mathbf{x}_j)) - 1)^2 + \lambda \|M\|_F^2 . \qquad (33)$$

While the expected objective of a metric M can be formulated as:

$$F(M) = \mathbb{E}_{\mathbf{z}_1, \mathbf{z}_2}[(q_{12}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_1, \mathbf{x}_2)) - 1)^2] + \lambda \|M\|_F^2 .$$

This loss is equivalent to promote the distance of a pair to the predefined value $\gamma - q_{ij}$. For notation simplicity, we use λ as the non-negative parameter to weight the importance of the regularizer. The larger the value of λ, the more proportion of strong convexity of the regularized objective.

The synthetic dataset is generated as follows. Instances from both classes are sampled based on normal distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, respectively. Given the number of training instances, we divide examples equally into two classes (therefore, two classes with equal prior). The side information q_{ij} can be generated based on the class difference of sampled instances. For clarity of notations, we assume $\mu_1 = \mathbf{1}$ and $\mu_2 = -\mathbf{1}$, where $\mathbf{1}$ is the vector with all elements equal one. Then, we assume both covariances of the two classes are equal to ηI, where $\eta > 0$ is a nonnegative parameter to tune the proportion of overlap between these two classes and I is the identity matrix. Threshold value γ is set as 2. Since

the true distribution of examples is known and fixed beforehand, the true risk of a particular metric M can be computed analytically[2]:

$$F(M) = 5 - 8\eta \mathrm{Tr}(M) + 8\eta^2 \mathrm{Tr}(MM) + 4\eta^2(\mathrm{Tr}(M))^2 - 12\mathbf{1}^\top M\mathbf{1} + 16\eta\mathbf{1}^\top MM\mathbf{1} + 8(\mathbf{1}^\top M\mathbf{1})^2 + 8\eta \mathrm{Tr}(M)\mathbf{1}^\top M\mathbf{1} + \lambda\|M\|_F^2 . \quad (34)$$

Considering the computational burden, the number of training examples N changes from 2 to 4000 and we constrain M to be a diagonal matrix $M = diag(\mathbf{m})$, where $diag(\cdot)$ transforms a vector $\mathbf{m} \in \mathbb{R}^d$ to a diagonal matrix.[3] Based on the diagonal property, the closed form of expected objective in Eq. 34 can be simplified as:

$$\begin{aligned} F(M) &= 5 - 8\eta\mathbf{1}^\top\mathbf{m} + 8\eta^2\mathbf{m}^\top\mathbf{m} + 4\eta^2(\mathbf{1}^\top\mathbf{m})^2 - 12\mathbf{1}^\top\mathbf{m} \\ &\quad + 16\eta\mathbf{m}^\top\mathbf{m} + 8(\mathbf{1}^\top\mathbf{m})^2 + 8\eta(\mathbf{1}^\top\mathbf{m})^2 + \lambda\|\mathbf{m}\|_2^2 . \\ &= \epsilon(\mathbf{m}) + \lambda\|\mathbf{m}\|_2^2 \end{aligned} \quad (35)$$

The empirical objective in Eq. 33 equals

$$\begin{aligned} \hat{F}(M) &= \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} ((\gamma - q_{ij}) - (\mathbf{x}_i - \mathbf{x}_j) diag(\mathbf{m})(\mathbf{x}_i - \mathbf{x}_j))^2 + \lambda\|\mathbf{m}\|_2^2 \\ &= \underbrace{\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} ((\gamma - q_{ij}) - ((\mathbf{x}_i - \mathbf{x}_j) \odot (\mathbf{x}_i - \mathbf{x}_j))^\top \mathbf{m})^2}_{\epsilon_N(\mathbf{m})} + \lambda\|\mathbf{m}\|_2^2 . \end{aligned}$$

$\odot$ is the element-wise product. In other words, the empirical optimal solution can be obtained by solving the ridge regression problem with pairwise instance extension $(\mathbf{x}_i - \mathbf{x}_j) \odot (\mathbf{x}_i - \mathbf{x}_j) \in \mathbb{R}^d$ by a closed form solution.

Remark 9 We check the strongly convexity of the expected regularized objective $F(M)$ by observing its Hessian matrix w.r.t. the diagonal part $\mathbf{m}$ based on Eq 35:

$$\frac{\partial^2 F(\mathbf{m})}{\partial^2 \mathbf{m}} = (16\eta^2 + 32\eta + \lambda)I + (8\eta^2 + 16\eta + 16)\mathbf{1}\mathbf{1}^\top . \quad (36)$$

The second term of Eq. 36 is positive semidefinite, while the first term has all eigen-values equal $(16\eta^2 + 32\eta + \lambda) > 0$. So the Hessian matrix is positive definite, and the expected objective is strongly convex. For our specific expected objective $F(\mathbf{m})$, its strongly convex proportion depends on both the value of λ and η.

When focusing on the loss term $\epsilon_N(\mathbf{m})$, with no regularizer, it is not necessary strongly convex since the instance matrix is not always full row rank. For its expectation $\epsilon(\mathbf{m})$, however, has positive definite Hessian and minimum eigenvalue at least $8\eta^2 + 16\eta$, such that satisfying the strongly convex expectation objective as in Theorem 1.

Remark 10 The value of λ used in the regularizer usually is tuned with the different number of instances. In the paper, we analyze the behavior of excess risk for distance metric learning objective with fixed regularizer $\Omega(M)$, the fixed balance parameter λ in this case. In addition,

[2] The derivation of a general case can be found in the appendix.

[3] Since only the symmetric property of the learned metric is used in our Theorem 1, to be consistent, we do not impose the PSD constraint on the learned metric here.

we can expect there exists a specified λ (or the true value of regularization parameter λ^*), and the result shows how fast the regularized objective convergence with this pre-defined parameter setting.

For a fixed N, $\frac{N}{2}$ examples are randomly sampled from two normal distributions respectively, and $N(N-1)$ pairs are generated to train the distance metric M. This procedure is repeated 30 times, and in each trial, benefited from the square form of the objective, the empirical optimal solution $\hat{M}$ can be found in a closed form. The optimal $F(M^*)$ is approximated by the minimum value of all computed $F(\hat{M})$. The same strategy is also used to compute or estimate $\epsilon(\hat{M})$ and $\epsilon(M^*)$.[4]

Mean values and error bars of excess risk w.r.t. variations on the training number of examples with different parameter settings are plotted. To validate the convergence rate of excess risk, we times it with different parameters proportion to the number of training instances, namely, $(F(\hat{M})-F(M^*))\times\sqrt{N}$, $(F(\hat{M})-F(M^*))\times N$, and $(F(\hat{M})-F(M^*))\times N^2$. If the plot has a descending trend, it means that its convergence ratio is lower than the one over the corresponding scale. When there appears an ascending trend, one over the corresponding scale is a higher estimation. If the plot approaches a constant at last, it is the corresponding ratio reveal the right convergence rate. The same strategy is also used for $(\epsilon(\hat{M})-\epsilon(M^*))$. In the legends, "$d$" indicates the dimensionality of instances, η representing the mess level of each class, and "λ" reveals the impact of strongly convex regularizer additionally introduced by the regularizer except for the loss function itself.

We first investigate the change of excess risk in the low dimension case $d=2$, and the results are in the Fig. 1. Each row in Fig. 1 represents the change of excess risk w.r.t. the number of training instances in a particular learning scenario (with different parameters for dataset generation or metric learning objective, i.e., η and λ). The first and third rows focus on the excess risk change over the loss term. Four columns correspond to the $\sqrt{N}$ (blue), N (green), N^2 (red), and 1 (pink) times the excess risk $F(\hat{M})-F(M^*)$ (resp. $\epsilon(\hat{M})-\epsilon(M^*)$), respectively. The darker color in plots shows the mean value of estimated excess risk over all trials, while the lighter color shows the variance. This appearance style is also used for the following experiments. For the change trend of the objective excess risk $F(\hat{M})-F(M^*)$, it will converge to zero given enough training examples as shown in the fourth column. These results verify the fact that the metric $\hat{M}$ learned from the empirical objective approach to the one M^* learned from the true distribution given more and more examples. For its convergence rate, we can find that in figures (f), (n) and (v), $(F(\hat{M})-F(M^*))\times N$ approaches to a constant very quickly, i.e., with a small number of training instances, which validate that there is a basic $\mathcal{O}\left(\frac{1}{N}\right)$ convergence rate for the whole objective as in Theorem 1. As shown in the Theorem 1, small value of $\epsilon(M^*)$ is also a necessary condition to get the faster $\mathcal{O}\left(\frac{1}{N^2}\right)$ convergence rate. Since the hypothesis space is small in this low dimensional metric learning case, it may be difficult to find a suitable M^* to make the value $\epsilon(M^*)$ small enough, so only the $\mathcal{O}\left(\frac{1}{N}\right)$ can be obtained. The convergence can obtain a non-apparent improvement as in plot (s) when η is very small and λ is relatively large. The plot in (s) levels out at last, which means that in this easier case with lower noise, $\epsilon(M^*)$ is relatively small, and the convergence rate of excess risk may be faster than $\mathcal{O}\left(\frac{1}{N}\right)$. The first and third rows in Fig. 1 show the convergence property of the loss part $\epsilon(\hat{M})-\epsilon(M^*)$. It is noteworthy that the empirical solution $\hat{M}$ is solved from the regularized objective, as in most metric learning implementations, and measured by the expected loss function. Like the change of objective excess risk, it will converge to zero when the number of training instances is large. From

[4] In the following context, we use metric M and its diagonal vector $\mathbf{m}$ exchangeably.

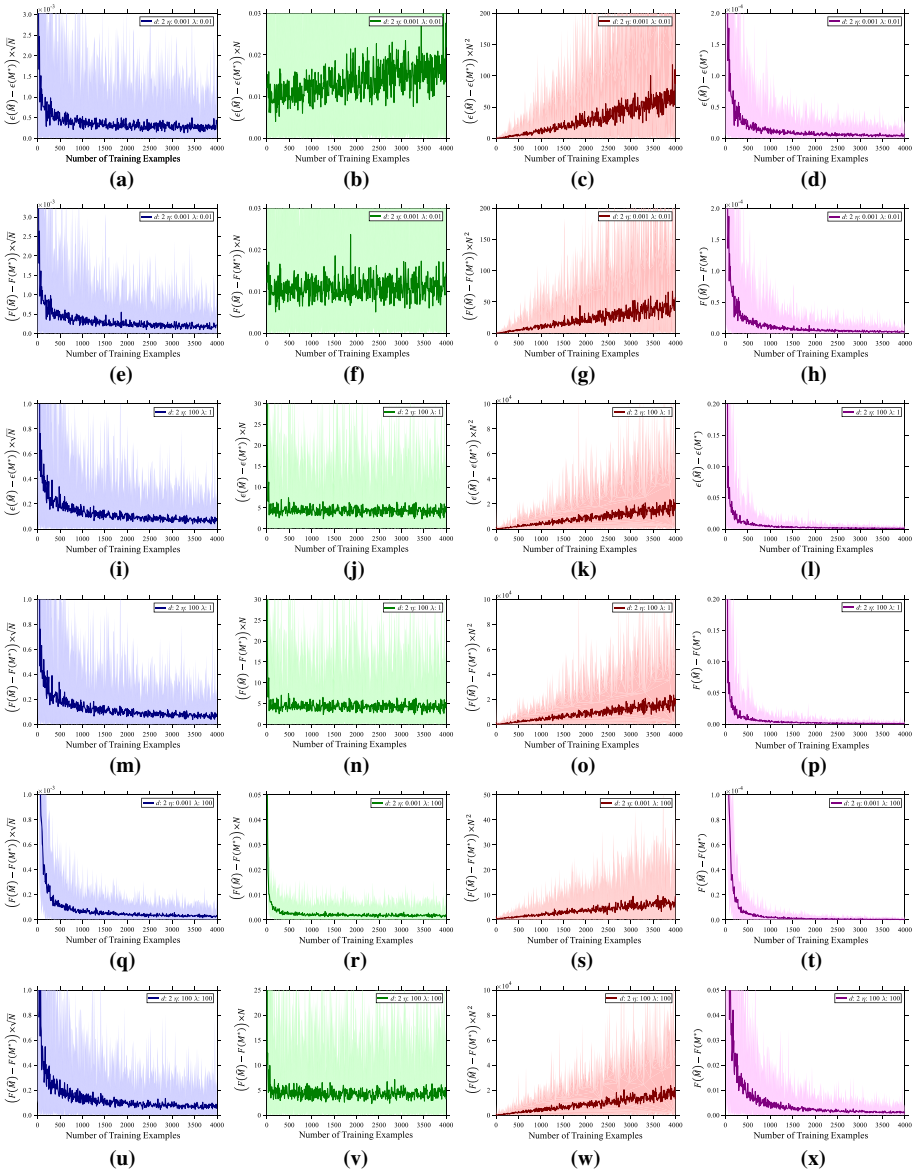

Fig. 1 Results on four sets of synthetic datasets. Each row of figures represents the change of excess risk w.r.t. the number of training instances in a particular learning scenario (with different parameters for dataset generation or metric learning objective). The first and third rows focus on the excess risk change over the loss term. Four columns correspond to the $\sqrt{N}$ (blue), N (green), N^2 (red), and 1 (pink) times the excess risk $F(\hat{M}) - F(M^*)$ (resp. $\epsilon(\hat{M}) - \epsilon(M^*)$), respectively (Color figure online)

(a)-(b), the $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ convergence rate can be discovered, as we proved in Remark 2. But in (j), the convergence of excess risk for loss part has the order $\mathcal{O}\left(\frac{1}{N}\right)$. The improvement of the convergence may come from the strongly convexity property of loss part itself introduced by

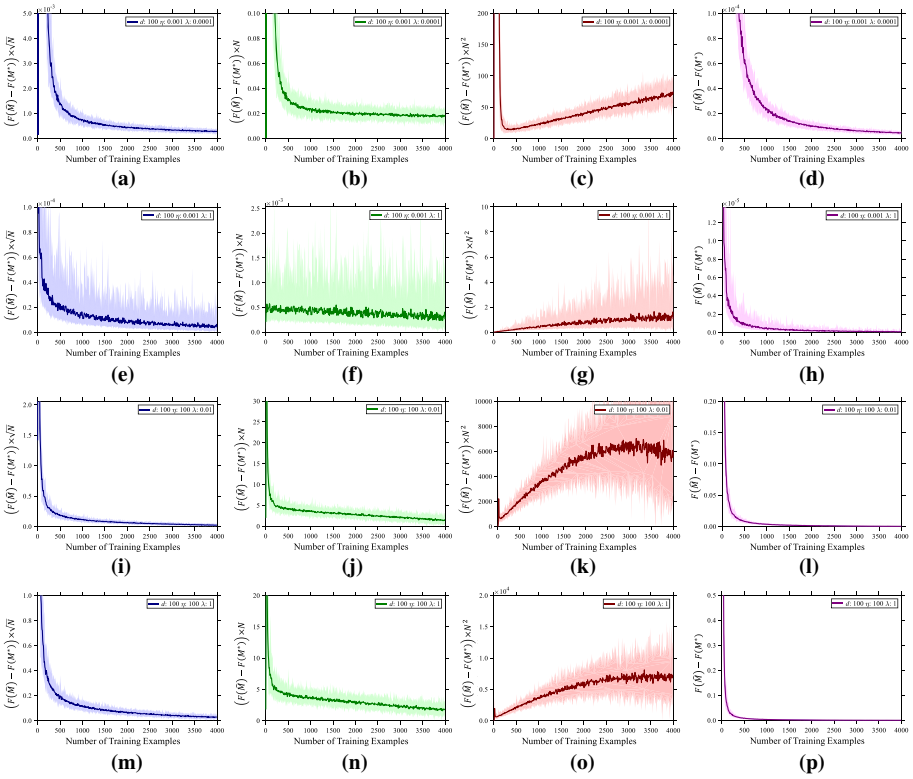

Fig. 2 Excess risk results in the case dimensionality $d = 100$. Each row of figures represents the change of excess risk w.r.t. the number of training instances in a particular learning scenario (with different parameters for dataset generation or metric learning objective). Four columns correspond to the $\sqrt{N}$ (blue), N (green), N^2 (red), and 1 (pink) times the excess risk $F(\hat{M}) - F(M^*)$, respectively (Color figure online)

η, as shown in Remark 9. We neglect the change of excess risk $\epsilon(\hat{M}) - \epsilon(M^*)$ in last two cases, i.e., $\eta = 0.001, \lambda = 100$ and $\eta = 100, \lambda = 100$. For the first case, $\epsilon(\hat{M}) - \epsilon(M^*)$ diverges since a large value of λ masks all properties of the distance measure loss counterpart, and the learned metric does not have small loss values. For the later case, the excess risk of loss is very similar to the excess risk of the objective.

The convergence results of the expected objective in the higher dimension case $d = 100$ are shown in Fig. 2. First two rows are the low noise case ($\eta = 0.001$), while in the latter two rows $\eta = 100$. According to the results of Remark 9, when the expected objective is far from strongly convexity, i.e., both η and λ are too small, there is only a $\mathcal{O}\left(\frac{1}{N}\right)$ rate as in plot (b). While in other cases as in plots (g), (k), and (o), the strongly convex objective accelerates the convergence rate of the excess risk a lot. It validates that for the *non-i.i.d.* pairwise distance metric learning setting, when N is large enough, the convergence rate of $F(\hat{M}) - F(M^*)$ can obtain the even faster order of $\mathcal{O}\left(\frac{1}{N^2}\right)$. These results are consistent with our theoretical analysis in Theorem 1. Comparing (g) with (k) and (o), it is notable that in later two cases the convergence rate of excess risk obtain the faster $\mathcal{O}\left(\frac{1}{N^2}\right)$ rate earlier (the plots become a constant trend with a smaller number of instances). This phenomenon results from the fact that η influences more the strongly convexity of the expected objective than λ with given

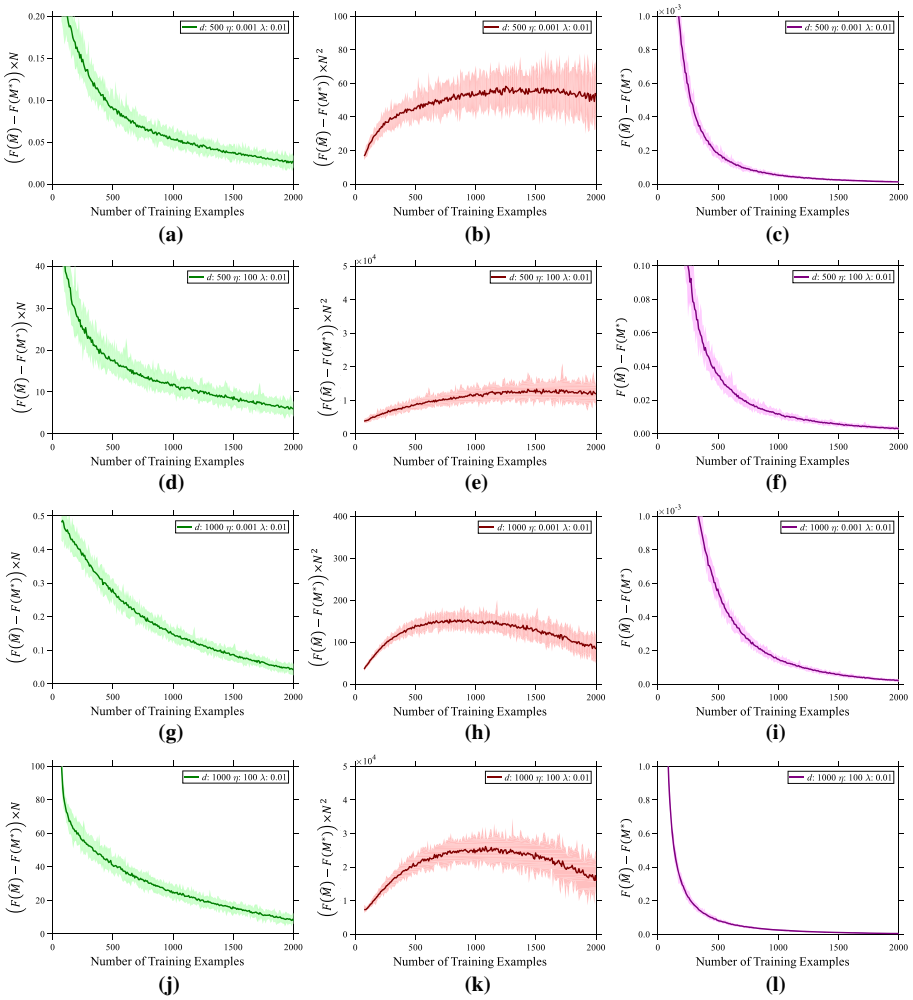

Fig. 3 Excess risk results when dimension $d = 500, 1000$. Each row of figures represents the change of excess risk w.r.t. the number of training instances in a particular learning scenario (with different parameters for dataset generation or metric learning objective). Three columns correspond to the N (green), N^2 (red), and 1 (pink) times the excess risk $F(\hat{M}) - F(M^*)$, respectively (Color figure online)

sets of values as in Remark 9. In addition, the threshold value of N to achieve the faster convergence rates $\mathcal{O}\left(\frac{1}{N^2}\right)$ may be *smaller than* the theoretical value in Eq. 6.

For higher dimension scenario, i.e., dimension $d = 500$ and $d = 1000$, we show the change of the expected objective function values in Fig. 3. Considering the computational burden, the maximum number of training instances is set to 2000. We neglect the results of with $\sqrt{N}$ scale since in this higher dimensional case the convergence rate of the expected objective function is obviously faster than $\mathcal{O}\left(\frac{1}{N}\right)$. In different settings, the expected objective function values converge at last. Based on (b), (e), (h), (k) in Fig. 3, the rate of convergence achieves the faster rate with order $\mathcal{O}\left(\frac{1}{N^2}\right)$. This is consistent with our theoretical analysis since when dimensionality is higher enough, it is easier to get low expected loss function

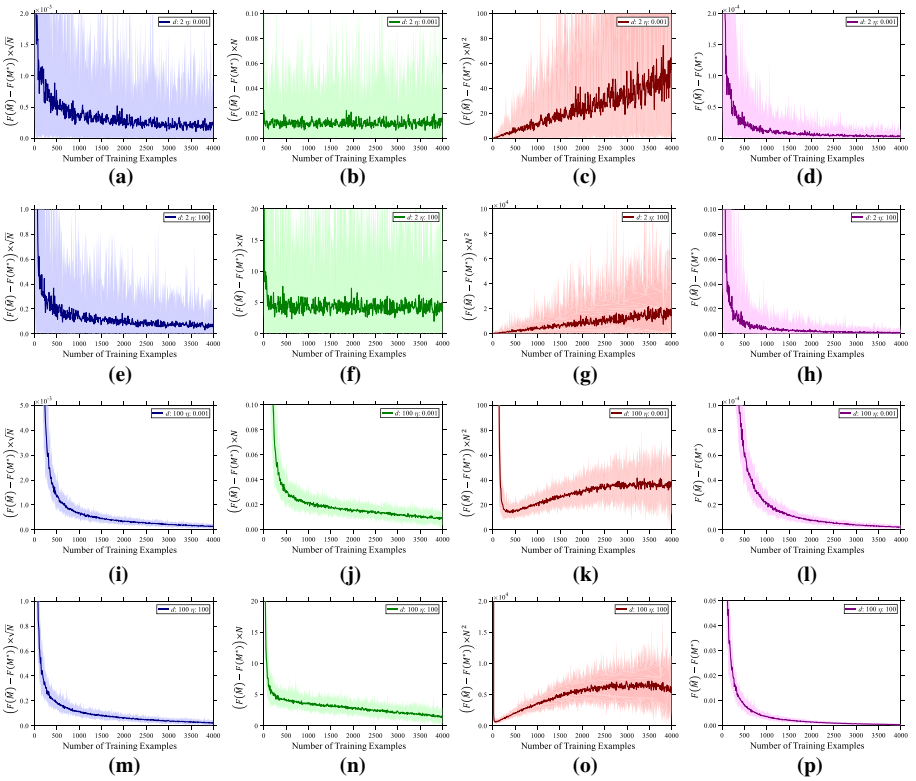

Fig. 4 Excess risk results when there is *no regularizer* (loss part is the objective). Each row of figures represents the change of excess risk w.r.t. the number of training instances in a particular learning scenario (with different parameters for dataset generation or metric learning objective). Four columns correspond to the $\sqrt{N}$ (blue), N (green), N^2 (red), and 1 (pink) times the excess risk $F(\hat{M}) - F(M^*)$ (resp. $\epsilon(\hat{M}) - \epsilon(M^*)$), respectively (Color figure online)

value in such a large hypothesis space. From these numerical results, when d and N are large enough, it seems that the convergence rate could be even higher than $\mathcal{O}\left(\frac{1}{N^2}\right)$ since there exists a decreasing trend of the expected convergence rate especially in (h). The counter-intuitive phenomenon may result from the exponentially expanded hypothesis space or the sparsity in high dimensionality problems. Some new theoretical analysis with different conditions may be derived to explain the results (Fig. 4).

At last we consider the case there is *no regularizer* $\Omega(M)$, and the whole objective has only the loss part, i.e., $F(M) = \epsilon(M)$. Different from the previous cases where $\hat{M}$ and M^* is computed or estimated from the empirical objective with regularizer, here the corresponding solutions of metric are computed and estimated based just on the loss function. First two rows show the case with dimensionality equal 2, and last two rows are the cases $d = 100$. It can be clearly found the excess risk of the objective (only contains loss term) can achieve fast convergence rate $\mathcal{O}\left(\frac{1}{N}\right)$ in low dimension while achieving faster rate with order $\mathcal{O}\left(\frac{1}{N^2}\right)$ in high dimension scenario. On the one hand, as pointed out in Remark 9, even the empirical square loss is not strongly convex, its strongly convex expected objective is sufficed to use our Theorem 1. Therefore, the faster convergence rate over loss function is achieved compared with previous results (Guo and Ying 2014; Bellet et al. 2015; Verma and Branson 2015; Cao

et al. 2016). On the other hand, since in the high dimensional space it is easier to find M^* with lower $\epsilon(M^*)$, we can explain the faster rate in (k) and (o) using the results of the Theorem 1.

In summary, our experiments on synthetic datasets validate the result in Theorem 1, that the distance metric learning method will achieve fast convergence rate for smooth loss function and strongly convex objective.

8 Conclusion and discussion

Distance metric learning is widely used in various machine learning fields and helps to improve the performance of similarity/distance based methods a lot. With strongly convex and smooth properties for the objective function and the loss respectively, we prove that the generalization ability of distance metric learning problem, in particular, the excess risk, can achieve faster convergence rate than previous results. This result also validates that the distance metric learning problem, although dealing with non-i.i.d. training inputs, has the same good property as traditional classification tasks concerning i.i.d. examples. We also give discussions on the relatedness of our analysis with previous implementations of metric learning methods. Similar proof techniques can also be applied to the ranking task while improving its excess risk convergence rate as well.

More interesting investigations could be extended from our results. First, comparing the example number condition in Eq. 6 and the results in Figs. 1 and 2, the required number of examples to get a faster rate may be smaller than the number computed by Eq. 6. In addition, since the distance metric learning objective uses the pairwise information, it implicitly uses more information from examples than the i.i.d. classification models. Therefore, it is reasonable to guess there exists a condition with a smaller number of examples but achieves the faster convergence rates of excess risk. Second, although the properties of objective terms are considered in the paper, the properties of the metric is not stressed, e.g., the low-rank, sparse, or other complex structure induced by the regularizer. It is a promising future work to analyze the possibility of having better rates with structural metrics.

Acknowledgements Funding was provided by National Key R&D Program of China (2018YFB1004300), NSFC (Grant No. 61773198, 61632004).

Appendix

In the appendix, we derive the analytic form of generalization error for distance metric learning with a square loss when applying to examples generated from normal distributions, which helps analyze various properties of our theorem.

Given a two-class datasets with instances generated by $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, the distance metric learning objective with square loss is defined as:

$$\begin{aligned} F(M) &= \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[(q_{12}(\gamma - \mathrm{Dis}_M^2(\mathbf{x}_1, \mathbf{x}_2)) - 1)^2] + \Omega(M) \ . \\ &= \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[((\gamma - q_{12}) - \mathrm{Dis}_M^2(\mathbf{x}_1, \mathbf{x}_2))^2] + \Omega(M) \ . \end{aligned} \tag{37}$$

Since $\mathbf{z}_1$ and $\mathbf{z}_2$ are sampled independently from latent distribution, we set an equal prior distribution for both classes, i.e., $\Pr(y_1 = 1) = \Pr(y_1 = 2) = \frac{1}{2}$. The expected square loss value in Eq. 37 can be decomposed as

$$\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[((\gamma - q_{12}) - \mathrm{Dis}_M^2(\mathbf{x}_1, \mathbf{x}_2))^2]$$

$$= \frac{1}{4}\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[((\gamma - q_{12}) - \mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2))^2 \mid y_1 = 1, y_2 = 1]$$
$$+ \frac{1}{4}\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[((\gamma - q_{12}) - \mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2))^2 \mid y_1 = 2, y_2 = 2]$$
$$+ \frac{1}{2}\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[((\gamma - q_{12}) - \mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2))^2 \mid y_1 = 1, y_2 = 2] \tag{38}$$

For a pair of instance $(\mathbf{x}_1, \mathbf{x}_2)$, when they come from the same class, e.g., class 1, their difference vector come from the distribution $\mathbf{m}_{12} = \mathbf{x}_1 - \mathbf{x}_2 \sim \mathcal{N}(0, 2\Sigma_1)$. Thus, their expected distance value:

$$\mathbb{E}[\mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2)] = \mathbb{E}[(\mathbf{x}_1 - \mathbf{x}_2)^\top M(\mathbf{x}_1 - \mathbf{x}_2)] = \mathbb{E}[\mathbf{m}_{12}^\top M\mathbf{m}_{12}] = 2\mathrm{Tr}(M\Sigma_1) \ .$$

In addition, the expected value of their squared distance is:

$$\mathbb{E}[\mathrm{Dis}^4_M(\mathbf{x}_1, \mathbf{x}_2)] = \mathbb{E}[\mathbf{m}_{12}^\top M\mathbf{m}_{12}\mathbf{m}_{12}^\top M\mathbf{m}_{12}] = 8\mathrm{Tr}(M\Sigma_1 M\Sigma_1) + 4(\mathrm{Tr}(M\Sigma_1))^2 \ .$$

Therefore, the expected loss value for same class instances can be computed by:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[((\gamma - 1) - \mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2))^2] \\
&= \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[((\gamma - 1)^2 - 2(\gamma - 1)\mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2)) + \mathrm{Dis}^4_M(\mathbf{x}_1, \mathbf{x}_2))] \\
&= (\gamma - 1)^2 - 2(\gamma - 1)\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[\mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2))] + \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[\mathrm{Dis}^4_M(\mathbf{x}_1, \mathbf{x}_2)] \\
&= (\gamma - 1)^2 - 4(\gamma - 1)\mathrm{Tr}(M\Sigma_1) + 8\mathrm{Tr}(M\Sigma_1 M\Sigma_1) + 4(\mathrm{Tr}(M\Sigma_1))^2 \ .
\end{aligned} \tag{39}$$

When $(\mathbf{x}_1, \mathbf{x}_2)$ comes from the first and second classes respectively, $\mathbf{m}_{ij} \sim \mathcal{N}(\mu_1 - \mu_2, \Sigma_1 + \Sigma_2)$. Using similar derivations, we have the expected loss:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[((\gamma - 1) - \mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2))^2] \\
&= (\gamma + 1)^2 - 2(\gamma + 1)\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[\mathrm{Dis}^2_M(\mathbf{x}_1, \mathbf{x}_2))] + \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[\mathrm{Dis}^4_M(\mathbf{x}_1, \mathbf{x}_2)] \\
&= (\gamma + 1)^2 - 2(\gamma + 1)(\mathrm{Tr}(M(\Sigma_1 + \Sigma_2)) + (\mu_1 - \mu_2)^\top M(\mu_1 - \mu_2)) \\
&\quad + 2\mathrm{Tr}(M(\Sigma_1 + \Sigma_2)M(\Sigma_1 + \Sigma_2)) + 4(\mu_1 - \mu_2)^\top M(\Sigma_1 + \Sigma_2)M(\mu_1 - \mu_2) \\
&\quad + (\mathrm{Tr}(M(\Sigma_1 + \Sigma_2)) + (\mu_1 - \mu_2)^\top M(\mu_1 - \mu_2))^2 \ .
\end{aligned} \tag{40}$$

Plugging Eqs. 39 and 40 into Eq. 38, we can get the expected objective. When we set $\gamma = 2$, $\mu_1 = \mathbf{1}$, $\mu_2 = -\mathbf{1}$, and $\Sigma_1 = \Sigma_2 = \eta I$, the expected objective can be simplified as:

$$\begin{aligned}
F(M) = 5 - 8\eta\mathrm{Tr}(M) + 8\eta^2\mathrm{Tr}(MM) + 4\eta^2(\mathrm{Tr}(M))^2 - 12\mathbf{1}^\top M\mathbf{1} \\
+ 16\eta\mathbf{1}^\top MM\mathbf{1} + 8(\mathbf{1}^\top M\mathbf{1})^2 + 8\eta\mathrm{Tr}(M)\mathbf{1}^\top M\mathbf{1} + \Omega(M) \ .
\end{aligned}$$

References

Agarwal, S., & Niyogi, P. (2009). Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, *10*(Feb), 441–474.

Bartlett, P. L., Bousquet, O., & Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, *33*(4), 1497–1537.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, *3*(Nov), 463–482.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*(1), 183–202.

Bellet, A., & Habrard, A. (2015). Robustness and generalization for metric learning. *Neurocomputing*, *151*, 259–267.

Bellet, A., Habrard, A., & Sebban, M. (2012). Similarity learning for provably accurate sparse linear classification. In *Proceedings of the 29th international conference on machine learning, Edinburgh, Scotland* (pp. 1871–1878).

Bellet, A., Habrard, A., & Sebban, M. (2015). *Metric learning. Synthesis lectures on artificial intelligence and machine learning*. Rafael: Morgan & Claypool Publishers.

Bian, W., & Tao, D. (2011). Learning a distance metric by empirical loss minimization. In *Proceedings of the 22nd international joint conference on artificial intelligence Barcelona, Spain* (pp. 1186–1191).

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*(Mar), 499–526.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Cao, Q., Guo, Z., & Ying, Y. (2016). Generalization bounds for metric and similarity learning. *Machine Learning*, *102*(1), 115–132.

Changpinyo, S., Liu, K., & Sha, F. (2013). Similarity component analysis. *Advances in neural information processing systems* (Vol. 26, pp. 1511–1519). Cambridge: MIT Press.

Chechik, G., Sharma, V., Shalit, U., & Bengio, S. (2010). Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, *11*, 1109–1135.

Clémençon, S., Lugosi, G., & Vayatis, N. (2008). Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, *36*(2), 844–874.

Davis, J. V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th international conference on machine learning, Corvalis, OR* (pp. 209–216).

Do, H., Kalousis, A., Wang, J., & Woznica, A. (2012). A metric learning perspective of SVM: on the relation of LMNN and SVM. In *Proceedings of the 15th international conference on artificial intelligence and statistics*, April 21–23, 2012, La Palma, Canary Islands (pp 308–317).

Frome, A., Singer, Y., & Malik, J. (2007). Image retrieval and classification using local distance functions. *Advances in neural information processing systems* (Vol. 19, pp. 417–424). Cambridge, MA: MIT Press.

Guo, Z., & Ying, Y. (2014). Guaranteed classification via regularized similarity learning. *Neural Computation*, *26*(3), 497–522.

Hsieh, C. K., Yang, L., Cui, Y., Lin, T. Y., Belongie, S. J., & Estrin, D. (2017). Collaborative metric learning. In *Proceedings of the 26th international conference on World Wide Web, Perth, Australia* (pp. 193–201).

Huang, K., Ying, Y., & Campbell, C. (2009). Gsml: A unified framework for sparse metric learning. In *Proceedings of the 9th IEEE international conference on data mining, Miami, FL* (pp. 189–198).

Hwang, S. J., Grauman, K., & Sha, F. (2013). Analogy-preserving semantic embedding for visual object categorization. In *Proceedings of the 30th international conference on machine learning, Atlanta, GA* (pp. 639–647).

Jin, R., Wang, S., & Zhou, Y. (2010). Regularized distance metric learning: Theory and algorithm. *Advances in neural information processing systems* (Vol. 23, pp. 862–870). Cambridge, MA: MIT Press.

Kulis, B. (2012). Metric learning: A survey. *Foundations and Trends in Machine Learning*, *5*(4), 287–364.

Law, M. T., Thome, N., & Cord, M. (2016a). Learning a distance metric from relative comparisons between quadruplets of images. *International Journal of Computer Vision*, *121*(1), 65–94.

Law, M. T., Yu, Y., Cord, M., & Xing, E. P. (2016b). Closed-form training of mahalanobis distance for supervised clustering. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition, Las Vegas, NV* (pp. 3909–3917)

Lim, D., Lanckriet, G., & McFee, B. (2013). Robust structural metric learning. In *Proceedings of the 30th international conference on machine learning, Atlanta, GA* (pp. 615–623).

Mason, B., Jain, L., & Nowak, R. D. (2017). Learning low-dimensional metrics. *Advances in neural information processing systems* (Vol. 30, pp. 4142–4150). Cambridge: MIT Press.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in combinatorics*, *141*(1), 148–188.

McFee, B., Lanckriet, G. R. (2010). Metric learning to rank. In *Proceedings of the 27th international conference on machine learning, Haifa, Israel* (pp. 775–782).

Meir, R., & Zhang, T. (2003). Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, *4*(Oct), 839–860.

Park, M., Jitkrittum, W., Qamar, A., Szabó, Z., Buesing, L., & Sahani, M. (2015). Bayesian manifold learning: The locally linear latent variable model (LL-LVM). *Advances in neural information processing systems* (Vol. 28, pp. 154–162). Cambridge: MIT Press.

Perrot, M., & Habrard, A. (2015). A theoretical analysis of metric hypothesis transfer learning. In *Proceedings of the 32nd international conference on machine learning, Lille, France* (pp. 1708–1717).

Perrot, M., Habrard, A., Muselet, D., & Sebban, M. (2014). Modeling perceptual color differences by local metric learning. In *European conference on computer vision, Springer* (pp. 96–111).

Qian, Q., Jin, R., Zhu, S., Lin, Y. (2015). Fine-grained visual categorization via multi-stage metric learning. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition, Boston, MA* (pp. 3716–3724).

Qian, Q., Jin, R., Yi, J., Zhang, L., & Zhu, S. (2013). Efficient distance metric learning by adaptive sampling and mini-batch stochastic gradient descent (sgd). *Machine Learning*, *99*(3), 353–372.

Rejchel, W. (2012). On ranking and generalization bounds. *Journal of Machine Learning Research*, *13*(May), 1373–1392.

Rejchel, W. (2015). Fast rates for ranking with large families. *Neurocomputing*, *168*, 1104–1110.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge: Cambridge University Press.

Shalev-Shwartz, S., Singer, Y., & Ng, A. Y. (2004). Online and batch learning of pseudo-metrics. In *Proceedings of the 21st international conference on machine learning, Alberta, Canada* (pp. 94–102).

Srebro, N., Sridharan, K., & Tewari, A. (2010). Smoothness, low noise and fast rates. *Advances in neural information processing systems* (pp. 2199–2207). Cambridge: MIT Press.

Sridharan, K., Shalev-Shwartz, S., & Srebro, N. (2009). Fast rates for regularized objectives. *Advances in neural information processing systems* (pp. 1545–1552). Cambridge: MIT Press.

Verma, N., & Branson, K. (2015). Sample complexity of learning mahalanobis distance metrics. *Advances in neural information processing systems* (Vol. 28, pp. 2584–2592). Cambridge: MIT Press.

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems* (Vol. 18, pp. 1473–1480). Cambridge, MA: MIT Press.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*, 207–244.

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (2003). Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems* (Vol. 15, pp. 505–512). Cambridge, MA: MIT Press.

Ye. H. J., Zhan, D. C., Si, X. M., & Jiang, Y. (2016a) Learning feature aware metric. In *Proceedings of The 8th Asian conference on machine learning, Hamilton, New Zealand* (pp 286–301).

Ye, H. J., Zhan, D. C., Si, X. M., Jiang, Y., & Zhou, Z. H. (2016b). What makes objects similar: A unified multi-metric learning approach. *Advances in neural information processing systems* (Vol. 29, pp. 1235–1243). Cambridge: MIT Press.

Ying, Y., Huang, K., & Campbell, C. (2009). Sparse metric learning via smooth optimization. *Advances in neural information processing systems* (Vol. 22, pp. 2214–2222). Cambridge: MIT Press.

Zhan, D. C., Li, M., Li, Y. F., & Zhou, Z. H. (2009). Learning instance specific distances using metric propagation. In *Proceedings of the 26th international conference on machine learning, Montreal, Canada* (pp. 1225–1232).

Zhang, L., Yang, T., & Jin, R. (2017). Empirical risk minimization for stochastic convex optimization: $O(1/n)$- and $O(1/n^2)$-type of risk bounds. In *Proceedings of the 30th conference on learning theory, Amsterdam, The Netherlands* (pp. 1954–1979).