Mathematical Analysis on Information-Theoretic Metric Learning with Application to Supervised Learning

Jooyeon Choi, Chohong Min, and Byungjoon Lee

December 27, 2018

Abstract

This article presents a concrete mathematical analysis on Information-Theoretic Metric Learning(ITML)[4]. The analysis provides theoretical foundation for ITML, by supplying well-posedness, strong duality, and convergence. Our analysis suggests the correction of a typo in original ITML article [4] that may lead to the loss of accuracy in the metric learning. The necessity of this correction is confirmed by several numerical experiments on supervised learning.

1 Introduction

Many algorithms in machine learning depend on the setting of distance metric to measure similarities of data [9]. In the classification of data, K-Nearest Neighbor (KNN) [3] uses a metric to identify the nearest neighbors. One of the most popular algorithms in data clustering is K-Means algorithm [11] which is also dependent on the distance measurement between data.

The simplest distance metric to consider is Euclidean distance, which is a measurement to represent the distance between two points. Despite of its simplicity, Euclidean distance is often not suitable for distributed data due to the lack of information about correlation of data sets. Among many attempts to overcome this limitation, Mahalanobis distance [12] is one of the well-known distance metric. Mahalanobis distance of two points is defined by

$$d_{Mahal}\left(x_i, x_j\right) = \sqrt{\left(x_i - x_j\right)^T \Sigma^{-1} \left(x_i - x_j\right)}$$

where Σ is the covariance matrix of the data. This metric not only measures the distance between two points, but also reflects the correlation with given data sets. However, it is hard to obtain the true covariance of data in practice.

Tons of researches have been done to resolve this issue by learning a distance metric to approximate the covariance matrix Σ. The earliest attempt was the work of Xing et al. [14], where the Mahalanobis metric learning was conducted in a way that maximizing the sum of distances of between dissimilar pairs while keeping the sum of distances between similar pairs small. Weinberger et al. [13] proposed a metric learning method so called Large-Margin Nearest Neighbors (LMNN) based on a statistical learning on a pseudo-metric for KNN classification.

In this article, we focus on Information-Theoretic Metric Learning (ITML) suggested by Davis et al. [4], which has been one of the most efficient metric learning methods. Unlike previous works, ITML has no projection step on the positive semi-definite cone which is computationally expensive. The main point of their work is that a metric learning procedure can be seen as LogDet divergence regularization. The LogDet divergence is a Bregman matrix divergence generated by the convex function $\phi\left(X\right) = -\log\left|X\right|$, where X is a positive definite matrix. The Bregman divergence on positive definite matrices is defined as

$$D_{ld}\left(A, A_0\right) = \operatorname{tr}\left(AA_0^{-1}\right) - \log\left|AA_0^{-1}\right| - n \tag{1}$$

where n is the dimension of the input data. The formulation of ITML proposed in [4] is the following LogDet optimization problem:

$$\begin{aligned} \min_{A \succeq 0} \; & D_{ld}\left(A, A_0\right) + \gamma \cdot D_{ld}\left(\operatorname{diag}\left(\xi\right), \operatorname{diag}\left(\xi_0\right)\right) \\ \text{subject to } & \operatorname{tr}\left(A\left(x_i - x_j\right)\left(x_i - x_j\right)^T\right) \le \xi_{c(i,j)} \qquad (i,j) \in S \\ & \operatorname{tr}\left(A\left(x_i - x_j\right)\left(x_i - x_j\right)^T\right) \ge \xi_{c(i,j)} \qquad (i,j) \in D \end{aligned} \tag{2}$$

Here, S, D denote similar and dissimilar sets, respectively, and the slack variable ξ is introduced to guarantee the existence of a feasible solution A of (2).

In [4], authors solved the optimization problem (2) with the iterative method based on the Bregman projection, the Bregman iteration. This is an extension of the work of Kulis et al. [10]. The Bregman projection is simply performed by the following iterative procedure

$$A_{t+1} = A_t + \beta A_t \left(x_i - x_j\right) \left(x_i - x_j\right)^T A_t \tag{3}$$

where x_i and x_j are the constrained data and β is the projection parameter computed through the algorithm.

The main purpose of this article is to provide a mathematical analysis on the ITML algorithm. ITML has been one of the most applied algorithm in various fields of machine learning. Nevertheless, there have been no concrete analyses of the algorithm, especially on the Bregman iteration in the algorithm. Up to the current, ITML algorithm has been cited thousands of times and applied to numerous areas. To our best searches, a formal discussion of such analyses is still missing. It can be said that most of its users take just for granted the well-posedness and the basic convergence. Our aim is to furnish ITML algorithm and its wide-ranged applications with mathematical foundation. Our study reveals that there is a typo in ITML manuscript [4] that can lead to a serious flaw, and presents its correction.

An outline of the article as follows. In section 2, we present a brief explanation of the Bregman iteration. Section 3 provides a mathematical analysis on ITML and a correction to original ITML paper [4] based on this analysis. Several numerical experiments are implemented in section 4 to verify the necessity of a correction from section 3. The last section includes conclusions.

2 Bregman Iteration

Note that the formulation for ITML (2) is a constrained optimization problem as follow:

$$\begin{array}{ll} \text{Minimize} & f(x) \\ \text{Subject to} & x \in C_i, \forall i \in \{1, \ldots, m\} \end{array} \tag{4}$$

In this section, we will present a brief review of the Bregman iteration [2], which is one of the most successful algorithm in convex optimization.

Assume that the closed convex sets C_i for the constraints are given for $i \in \{1, \ldots, m\}$ and $R = \cap_{i=1}^{m} C_i$ is not empty. The key idea of Bregman iteration is to find extrema of $f(x)$ via the function $D : S \times S \to \mathbb{R}$ satisfying the following six conditions.

I. $D(x, y) \geq 0$, $D(x, y) = 0$ if and only if $x = y$.

II. For any $y \in S$, $i \in T$, a point $x = P_i y \in C_i \cap S$ exists such that

$$D(x, y) = \min_{z \in C_i \cap S} D(z, x)$$

This point x is called the D-projection of the point y onto the set C_i.

III. For each $i \in T$, $y \in S$, the function $G(z) = D(z, y) - D(z, P_i y)$ is convex over $C_i \cap S$.

IV. A derivative $\frac{\partial D}{\partial x}(x, y)$ of the function $D(x, y)$ exists and $\frac{\partial D}{\partial x}(y, y) = 0$.

V. For each $z \in R \cap S$ and for every real number L, the set $T = \{x \in S | D(z, x) \leq L\}$ is compact.

VI. If $D(x^n, y^n) \to 0$, $y^n \to y^* \in \bar{S}$, and the set of elements of the series $\{x^n\}$ is compact, then $x^n \to y^*$.

Once the function $D(x, y)$ is chosen, the optimization problem (4) can be solved by the iterative process

$$x^{n+1} = \arg \min_{z \in C_i \cap S} D(z, x^n) \tag{5}$$

as was proposed by Bregman in [2]. We restate the convergence result of the iterative process (5) from [2] to be self-contained.

Lemma 1. *For any sequence of indices, we have the following:*
(1) The set of elements of the relaxation sequence $\{x^n\}$ is compact.
(2) For any $z \in S$, there exists $\lim_{n \to \infty} D(z, x^n)$.
(3) $D(x^{n+1}, x^n) \to 0$ when $n \to \infty$.

The universal choice for the function D is the one so called "Bregman distance". The Bregman distance corresponding to a convex function f at the point y is defined by

$$D(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \tag{6}$$

With the Bregman distance, the optimization problem (4) for the inequality constraint

$$\begin{array}{ll} \text{Minimize} & f(x) \\ \text{Subject to} & x \in R = \cap_{i=1}^{m} C_i = \left\{ x | Ax \geq b, x \in \bar{S} \right\} \end{array} \tag{7}$$

was proved to be convergent provided the function D satisfies the following additional two conditions.

VII. The function $D(x,y)$ is defined when $x \in \bar{S}$,and if $y^n \to y^* \in S$, then $D(y^*, y^n) \to 0$.

VIII. The D-projection of any point x belonging to the interior of the set S onto the set $\{x|Ax = b\}$ also belong to the interior of S.

We finalize this section with restatement of the convergence result of the problem (7) from [2].

Theorem 1. *Assume that Bregman distance $D(x,y)$ satisfies the conditions I-VIII. Then, the sequence $\{x^n\}$ obtained as a result of applying KKT conditions on (7) converges to the point x^*, which is a solution of the problem (7).*

3 Mathematical analysis on ITML

In this section, we provide a concrete mathematical analysis on ITML algorithm. There are three parts in the analysis. The first part checks the well-posedness of the optimization, the second part discusses the strong duality of the optimization, and the third one presents the convergence analysis of the Bregman iteration.

3.1 Well-posedness of optimization

ITML algorithm solves the following minimization problem with linear constraints.

$$\begin{array}{rl} \text{Given } & A_0 \in \left(S_+^n\right)^o,\ \xi_0 \in \left(D_+^m\right)^o \\ & \{v_1, \cdots, v_m\} \in \mathbb{R}^n \\ & \{\delta_1, \cdots, \delta_m\} \in \{\pm 1\},\ \text{and } \gamma \in \mathbb{R}_+, \\ \text{minimize } & f(A,\xi) := D(A, A_0) + \gamma D(\xi, \xi_0) \\ \text{subject to } & A - A_0 \in S_+^n \text{ and } \xi - \xi_0 \in D_+^m \\ & (A,\xi) \in C_i = \left\{ (A,\xi) \,|\, \left(\left\langle A, v_i v_i^T \right\rangle - \xi_i \right) \delta_i \leq 0 \right\},\ \forall i \in \{1, \cdots, m\}. \end{array} \tag{8}$$

Lemma 2. *Let R be the set of (A,ξ) satisfying the constraints, then R is nonempty, convex and closed.*

Proof. The choice of $A = A_0$ and ξ with $\xi_i = \left\langle A_0, v_i v_i^T \right\rangle$, $\forall i$ satisfies the constraints, and C_i is nonempty. S_+^n and D_+^m are closed convex, and so are their affine translations. Each linearly constrained set is closed and convex. Since R is the intersection of closed and convex sets, R is closed and convex. □

Lemma 3. *For $A \in S_+^n$, $-\log|A| = -\sum_{i=1}^n \log \lambda_i$ and $\frac{\partial}{\partial A}\left[-\log|A|\right] = -A^{-1}$, where $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of A.*

Proof. See page 641 of [1]. □

Lemma 4. *$f(A,\xi)$ is convex in the domain $\left(A_0 + S_+^n\right) \times \left(\xi_0 + D_+^m\right)$. Furthermore, $f(A,\xi)$ is strictly convex in the interior and takes value $+\infty$ on the boundary.*

Proof. When (A,ξ) is on the boundary, either $A - A_0 \in \partial S_+^n$ or $\xi - \xi_0 \in \partial D_+^m$, which implies that either $A - A_0$ or $\xi - \xi_0$ has a zero eigenvalue. By Lemma 3, $f(A,\xi) = +\infty$ in either case.

Take any (A_1,ξ_1) and (A_2,ξ_2) from the domain. For $\lambda \in (0,1)$, consider the inequality of the convex condition,

$$(1-\lambda) f(A_1,\xi_1) + \lambda f(A_2,\xi_2) \geq f\left((1-\lambda)A_1 + \lambda A_2, (1-\lambda)\xi_1 + \lambda \xi_2\right).$$

When (A_1,ξ_1) or (A_2,ξ_2) is on the boundary, LHS becomes $+\infty$, and the inequality holds. Otherwise, both are inside.

As shown in page 74 of [1], $-\ln|\cdot|$ is strictly convex interior of S_+^n, and so is $D(\cdot, A_0)$ in $\left(A_0 + S_+^n\right)^o$. In the similar manner, $D(\cdot, A_0)$ is strictly convex in $\left(\xi_0 + D_+^m\right)^o$, and so is $f(A,\xi)$ in the interior domain. □

Lemma 5. *$f(A,\xi)$ is coercive in $\left(A_0+S_+^n\right)\times\left(\xi_0+D_+^m\right)$.*

Proof. When $\parallel (A,\xi)\parallel=\sqrt{\parallel A\parallel^2+\parallel\xi\parallel^2}\to+\infty$, either $\parallel A\parallel\to+\infty$ or $\parallel\xi\parallel\to+\infty$. We show that $D(A,A_0)$ is coercive in $A_0+S_+^n$. The other case can be similarly dealt with

$$D(A,A_0)=-\log|A|+\log|A_0|+\left\langle A-A_0,A_0^{-1}\right\rangle$$

Let $A=\lambda_1v_1v_1^T+\cdots+\lambda_nv_nv_n^T$, $\lambda_i\geq0$, $v_i\cdot v_j=\delta_{ij}$ and $A_0=\mu_1w_1w_1^T+\cdots+\mu_nw_nw_n^T$, $\mu_i\geq0$, $w_i\cdot w_j=\delta_{ij}$.

Since $\sum_{j=1}^n(v_i\cdot w_j)^2=1$, we have

$$\begin{aligned}
D(A,A_0)&=-\sum_{i=1}^n\log\lambda_i+\sum_{j=1}^n\log\mu_j-n+\sum_{i=1}^n\sum_{j=1}^n\frac{\lambda_i}{\mu_j}(v_i\cdot w_j)^2\\
&=-\sum_{i=1}^n\log\lambda_i\sum_{j=1}^n(v_i\cdot w_j)^2+\sum_{j=1}^n\log\mu_j\sum_{j=1}^n(v_i\cdot w_j)^2-n+\sum_{i=1}^n\sum_{j=1}^n\frac{\lambda_i}{\mu_j}(v_i\cdot w_j)^2\\
&=\sum_{i=1}^n\sum_{j=1}^n\left(-\log\lambda_i+\log\mu_j+\frac{\lambda_i}{\mu_j}-1\right)(v_i\cdot w_j)^2\\
&=\sum_{i=1}^n\left(\sum_{j=1}^n\left(\frac{\lambda_i}{\mu_j}-\log\frac{\lambda_i}{\mu_j}-1\right)(v_i\cdot w_j)^2\right)
\end{aligned}$$

Consider $x-\ln x-1\geq\frac{1}{2}(\ln x)^2$, if $x>1$.

As $\parallel A\parallel=\sqrt{\langle A,A\rangle}=\sqrt{\lambda_1+\cdots\lambda_n}\to\infty$, we have $\max(\lambda_1,\cdots,\lambda_n)\to\infty$. This implies that

$$\begin{aligned}
D(A,A_0)&\geq\sum_{j=1}^n\left(\frac{\lambda_i}{\mu_j}-\log\frac{\lambda_i}{\mu_j}-1\right)(v_i\cdot w_j)^2,\ \text{for some } i,\ \frac{\lambda_i}{\mu_j}>1\\
&\geq\frac{1}{2}\sum_{j=1}^n\left(\log\frac{\lambda_i}{\mu_j}\right)^2(v_i\cdot w_j)^2\\
&\geq\frac{1}{2}\left(\log\frac{\lambda_i}{\mu_{\max}}\right)^2=\max_i\left(\log\frac{\lambda_i}{\mu_{\max}}\right)^2=\left(\log\frac{\max_i\lambda_i}{\mu_{\max}}\right)\to\infty
\end{aligned}$$

□

Theorem 2. *ITML optimization described by (8) has a unique solution.*

Proof. By Lemma 2 and Lemma 4, ITML optimization is a proper convex optimization problem. Since $f(A,\xi)$ is strictly convex inside and finite only inside, the minimum point is unique. By Lemma 5, a minimum point exists by Proposition VI.2.2 in [6]. □

3.2 Strong Duality

ITML optimization has m number of linear constraints, but Bregman iteration solves the optimization with a single linear constraint that is iteratively chosen from the m constraints. In this section, we analyze the duality of the optimization with a single linear constraint.

$$\inf_{x\in S\cap H}D\left(x,x^k\right)=\inf_{x\in S}\sup_{\alpha\in\mathbb{R}_+}\phi\left(x,x^k,\alpha\right)$$

Here, $H=\left\{x=(A,\xi)\,|\delta\left(\left\langle A,v_iv_i^T\right\rangle-\xi_i\right)\leq0\right\}$, and the Lagrangian is denoted by

$$\phi(x,x^n;\alpha)=D\left(x,x^k\right)+\alpha\delta\left(\left\langle A,v_iv_i^T\right\rangle-\xi_i\right).$$

The weak duality, which holds by default, is the inequality,

$$\inf_{x\in S}\sup_{\alpha\in\mathbb{R}_+}\phi(x,x^n;\alpha)\geq\sup_{\alpha\in\mathbb{R}_+}\inf_{x\in S}\phi(x,x^n;\alpha).$$

The strong duality, which may not hold in general, is the equality in the above. A consequence of the strong duality is the extremal condition between primal and dual optimizations, the so called Karush-Kuhn-Tucker (KKT) condition.

Lemma 6.

$$\inf_{x\in S}\sup_{\alpha\in\mathbb{R}_+}\phi\left(x,x^n;\alpha\right)=\min_{x\in S}\sup_{\alpha\in\mathbb{R}_+}\phi\left(x,x^n;\alpha\right)=\sup_{\alpha\in\mathbb{R}_+}\inf_{x\in S}\phi\left(x,x^n;\alpha\right)$$

Proof. When $x=(A,\xi)\in\partial S$, either A or ξ has a zero eigenvalue, $-\log|A|-\gamma\log|\xi|=+\infty$. Consequently $\phi\left(x,x^n;\alpha\right)=+\infty$ on ∂S, and the infimum is attained inside S. The result follows from the coercivity of $D\left(x,x^k\right)$ in Lemma 5 and Proposition IV.2.3 in [6]. □

Lemma 7. $\sup_{\alpha\in\mathbb{R}_+}\inf_{x\in S}\phi\left(x,x^k;\alpha\right)=\max_{\alpha\in\mathbb{R}_+}\inf_{x\in S}\phi\left(x,x^k,\alpha\right)$

Proof. Let $g\left(\alpha\right)=\inf_{x\in S}\phi\left(x,x^k;\alpha\right)$. It is enough to show that g is coercive in each of the following cases. Then the result follows again from Proposition IV.2.3 in [6].

(a) Case $\delta=-1$

Take $x=(x,\xi)=\left(A^k+v_iv_i^T,\xi^k\right)$. Then $\phi\left(x,x^k;\alpha\right)=-\alpha\left\|v_i\right\|^4+(\text{const w.r.t.}\alpha)$ and $g(\alpha)\leq-\alpha\left\|v_i\right\|^4$. Thus $g(\alpha)\to-\infty$ as $\alpha\to\infty$.

(b) Case $\delta=1$

Take $x=(x,\xi)=\left(A^k,\xi^k+e_i\right)$. Then $\phi\left(x,x^k;\alpha\right)=-\alpha+(\text{const w.r.t.}\alpha)$. Thus $g(\alpha)\to-\infty$ as $\alpha\to\infty$. □

Theorem 3. *(KKT condition) There exists a unique solution $x=(A,\xi)$ for $\inf D\left(x,x^k\right)$ subject to $x\in S\cap H$, and x is characterized by*

$$\begin{cases}\alpha &= \min\left(0,\frac{\gamma}{1+\gamma}\delta_i\left(\frac{1}{\xi_i^k}-\frac{1}{p}\right)\right)\\ A &= A^k-\frac{\alpha\delta_i}{1+\alpha\delta_ip}A^kv_iv_i^TA^k\\ \xi_i &= \frac{\xi_i^k\gamma}{\gamma-\alpha\delta_i\xi_i^k}\end{cases}$$

where $p=\left\langle A^k,v_iv_i^T\right\rangle$ and $\xi_j=\xi_j^k$ when $j\neq i$.

Proof. By the above two lemmas and Proposition VI.1.2 in [6], there exists a saddle point $x=(A,\xi)$. Being strictly convex in S, the saddle point is unique. By the extremal condition of Proposition VI.1.6 in [6], we have

$$\begin{cases}-A^{-1}+A_k^{-1}+\alpha\delta_iv_iv_i^T &= 0\\ \left(-\xi^{-1}+\xi_k^{-1}-\frac{\alpha\delta_i}{\gamma}e_i\right)\cdot\alpha &= 0\end{cases}$$

when $\alpha\neq0$ or $\alpha>0$,

$$\begin{cases}A^{-1} &= \left(A^k\right)^{-1}+\alpha\delta_iv_iv_i^T\\ \xi^{-1} &= \left(\xi^k\right)^{-1}-\frac{\alpha\delta_i}{\gamma}e_i\end{cases}\quad\text{or}$$

By Sherman-Morrison formula, we have

$$A=A^k-\frac{\alpha\delta_i}{1+\alpha\delta_ip}A^kv_iv_i^TA^k\text{ , and}$$
$$\xi_i=\frac{\gamma\xi_i^k}{\gamma-\alpha\delta_i\xi_i^k}.$$

Finally, from $\left\langle A,v_iv_i^T\right\rangle-\xi_i=0$, we obtain $\alpha=\frac{\gamma}{1+\gamma}\cdot\delta_i\left(\frac{1}{\xi_i^k}-\frac{1}{p}\right)$. □

3.3 Convergence of Bregman iteration

In this subsection, we consider the application of Bregman iteration to solving ITML optimization. As reviewed in Section 2, Bregman iteration was proved to be convergent if the Bregman distance D_f generated by an objective function $f(x)$ satisfies conditions I-VI and VII-VIII.

Let $S=\left(S_+^n\right)^o\times\left(D_+^m\right)^o$ and $f\left(x\right)=D_{ld}\left(A,A_0\right)+\gamma\cdot D_{ld}\left(\mathrm{diag}\left(\xi\right),\mathrm{diag}\left(\xi_0\right)\right)$, where $x=(A,\xi)$. We begin with computing the Bregman distance D_f generated by $f(x)$.

Lemma 8. *$D_f\left(x,y\right)=D_{ld}\left(A,B\right)+\gamma D_{ld}\left(diag\left(\xi\right),diag\left(\zeta\right)\right)$, when $x=(A,\xi)$ and $y=(B,\zeta)$.*

Proof. By Lemma 3, $\nabla f(y) = \left(-B^{-1} + A_0^{-1}, \gamma\left(-\zeta^{-1} + \xi_0^{-1}\right)\right)$. Then

$$\begin{aligned}
D_f(x,y) &= f(x) - f(y) - \langle x - y, \nabla f(y)\rangle \\
&= D_{ld}(A, A_0) + \gamma D_{ld}(\mathrm{diag}(\xi), \mathrm{diag}(\xi_0)) - D_{ld}(B, A_0) - \gamma D_{ld}(\mathrm{diag}(\zeta), \mathrm{diag}(\xi_0)) \\
&\quad - \left\langle A - B, -B^{-1} + A_0^{-1}\right\rangle - \gamma\left\langle \mathrm{diag}(\xi) - \mathrm{diag}(\zeta), -\mathrm{diag}(\zeta)^{-1} + \mathrm{diag}(\xi_0)^{-1}\right\rangle \\
&= -\log|A| + \log|B| + \left\langle A - B, B^{-1}\right\rangle \\
&\quad + \gamma\left(-\log|\mathrm{diag}(\xi)| + \log|\mathrm{diag}(\zeta)| + \left\langle \mathrm{diag}(\xi) - \mathrm{diag}(\zeta), \mathrm{diag}(\zeta)^{-1}\right\rangle\right) \\
&= D_{ld}(A, B) + \gamma D_{ld}(\mathrm{diag}(\xi), \mathrm{diag}(\zeta)).
\end{aligned}$$

□

Hence, we prove that the Bregman distance D_f in lemma 8 satisfies conditions I-VI' and VII'-VIII.

Lemma 9. *D_f satisfies the conditions I-IV.*

Proof. By Lemma 4 and Lemma 5, $-\log|\cdot|$ is strictly convex and continuously differentiable in $\left(S_+^n\right)^o$ and in $\left(D_+^m\right)^o$, and so is $f(x)$ in S. By the argument in page 206 of [2], conditions I-IV are satisfied. □

Lemma 10. *D_f satisfies the condition V.*

Proof. Condition V holds true when $D(x,y)$ is coercive with respect to y. $D(x,y) = D_{ld}(A,B) + \gamma D_{ld}(\mathrm{diag}(\xi), \mathrm{diag}(\zeta))$, where $x = (A, \xi)$ and $y = (B, \zeta)$. It is enough to show that $D_{ld}(A,B)$ is coercive w.r.t. B. Let $A = \sum_{i=1}^n \mu_i v_i v_i^T$ and $B = \sum_{i=1}^n \lambda_i w_i w_i^T$.

$$\begin{aligned}
D_{ld}(A,B) &= -\log|A| + \log|B| + \left\langle A, B^{-1}\right\rangle - n \\
&= -\sum_{i=1}^n \log(\mu_i) + \sum_{j=1}^n \log(\lambda_j) + \sum_i \sum_j \frac{\mu_i}{\lambda_j}(v_i \cdot w_j)^2 - n \\
&= \sum_{i=1}^n \sum_{j=1}^n \left(\frac{\mu_i}{\lambda_j} - \log\frac{\mu_i}{\lambda_j} - 1\right)(v_i \cdot w_j)^2
\end{aligned}$$

Note that $x - \ln x - 1 \geq \frac{1}{4}\left(\frac{1}{x} - 1\right)$ if $0 < x < 1$. As $\| B \| = \sqrt{\langle B, B\rangle} = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_n^2 \to \infty$, we have $\max(\lambda_1, \cdots, \lambda_n) \to \infty$. This implies that

$$\begin{aligned}
D_{ld}(A,B) &= \sum_{i=1}^n \sum_{j=1}^n \left(\frac{\mu_i}{\lambda_j} - \log\frac{\mu_i}{\lambda_j} - 1\right)(v_i \cdot w_j)^2 \\
&\geq \sum_{i=1}^n \left(\frac{\mu_i}{\lambda_j} - \log\frac{\mu_i}{\lambda_j} - 1\right)(v_i \cdot w_j)^2 \quad \text{for some } j \\
&\geq \frac{1}{4}\sum_{i=1}^n \left(\frac{\lambda_j}{\mu_i} - 1\right)(v_i \cdot w_j)^2 \\
&\geq \frac{1}{4}\left(\frac{\lambda_{\max}}{\mu_i} - 1\right) \\
&\geq \frac{1}{4}\max_i \left(\frac{\lambda_{\max}}{\mu_i} - 1\right) \to \infty
\end{aligned}$$

□

Lemma 11. *D_f satisfies the condition VI.*

Proof. Let $x^k = \left(A^k, \xi^k\right)$ and $y^k = \left(B^k, \zeta^k\right)$ for each k, and $y^* = (B^*, \zeta^*)$. Using the orthogonal diagonalization, we have

$$\begin{aligned}
A^k &= \lambda_1^k v_1^k \left(v_1^k\right)^T + \cdots + \lambda_n^k v_n^k \left(v_n^k\right)^T \\
B^k &= \mu_1^k w_1^k \left(w_1^k\right)^T + \cdots + \mu_n^k w_n^k \left(w_n^k\right)^T,
\end{aligned}$$

where the eigenvalues are listed in the increasing order. Since $D\left(x^k, y^k\right) = D_{ld}\left(A^k, B^k\right) + \gamma D_{ld}\left(\xi^k, \zeta^k\right) \to 0$, we also have $D_{ld}\left(A^k, B^k\right) \to 0$.

$$D_{ld}\left(A^k, B^k\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\lambda_i^k}{\mu_j^k} - \log\left(\frac{\lambda_i^k}{\mu_j^k}\right) - 1\right)\left(v_i^k \cdot w_j^k\right)^2$$

Step I : we show that $\lambda_1^k \to \mu_1^*$.

Otherwise, for each small $\epsilon > 0$, there are infinitely many x^k's such that either $\lambda_1^k \in (-\infty, \mu_1^* - \epsilon)$, or $\lambda_1^k \in (\mu_1^* + \epsilon, \mu_2^* - \epsilon)$. Then

$$\begin{aligned}
\lim_{k\to\infty} D_{ld}\left(A^k, B^k\right) &\geq \lim_{k\to\infty}\sum_{j=1}^{n}\left(\frac{\lambda_1^k}{\mu_j^k} - \log\left(\frac{\lambda_1^k}{\mu_j^k}\right) - 1\right)\left(v_1^k \cdot w_j^k\right)^2 \\
&= \sum_{j=1}^{n}\left(\frac{\lambda_1^k}{\mu_j^*} - \log\left(\frac{\lambda_1^k}{\mu_j^*}\right) - 1\right)\left(v_1^k \cdot w_j^*\right)^2 \\
&\geq \frac{(\log\epsilon)^2}{4} \cdot 1 > 0,
\end{aligned}$$

which contradicts the assumption that $D_{ld}\left(A^k, B^k\right) \to 0$.

Step II : we show that $v_1^k \to w_1^*$, the first eigenvector. We may assume $\mu_1^* < \mu_2^*$, otherwise w_2^* can be taken as the first eigenvector.

Otherwise, for each small $\epsilon > 0$, there are infinitely many x^k's such that $\sum_{j=2}^{n}\left(v_1^k \cdot w_j^*\right)^2 > \epsilon$. Then

$$\begin{aligned}
\lim_{k\to\infty} D_{ld}\left(A^k, B^k\right) &\geq \lim_{k\to\infty}\sum_{j=1}^{n}\left(\frac{\mu_1^*}{\mu_j^k} - \log\left(\frac{\mu_1^*}{\mu_j^k}\right) - 1\right)\left(v_1^k \cdot w_j^k\right)^2 \\
&\geq \left(\frac{\mu_1^*}{\mu_2^*} - \log\left(\frac{\mu_1^*}{\mu_2^*}\right) - 1\right)\epsilon > 0,
\end{aligned}$$

which contradicts the assumption that $D_{ld}\left(A^k, B^k\right) \to 0$.

Step III :

we showed that the argument is true for the first eigenpair. We can recursively apply the argument to the next eigenpairs to show that $A^k \to B^*$. The case of $D_{ld}\left(\xi^k, \zeta^k\right)$ can be similarly treated. □

Lemma 12. *D_f satisfies the condition VII.*

Proof. Let $y^k = \left(B^k, \zeta^k\right)$ for each k, and $y^* = (B^*, \zeta^*)$. Using the orthogonal diagonalization, we have

$$\begin{aligned}
B^* &= \mu_1^* w_1^* \left(w_1^*\right)^T + \cdots + \mu_n^* w_n^* \left(w_n^*\right)^T, \\
B^k &= \mu_1^k w_1^k \left(w_1^k\right)^T + \cdots + \mu_n^k w_n^k \left(w_n^k\right)^T,
\end{aligned}$$

where the eigenvalues are listed in the increasing order. Since $y^k \to y^*$, we may assume $\mu_i^k \to \mu_i^*$ and $w_i^k \to w_i^*$ for each $i = 1, \cdots, n$. By Lemma 9 in [10], $\left\{y^k\right\}$ and y^* shares the same range space, so that $\left(B^*\right)^{-1}$ can be defined in the pseudo inverse on the range space, where $\mu^* \neq 0$. Then

$$\begin{aligned}
\lim_{k\to\infty} D_{ld}\left(B^k, B^*\right) &= \lim_{k\to\infty}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\mu_i^k}{\mu_j^*} - \log\left(\frac{\mu_i^k}{\mu_j^*}\right) - 1\right)\left(w_i^k \cdot w_j^*\right)^2 \\
&= \sum_{i=1}^{n}\left(\frac{\mu_i^*}{\mu_i^*} - \log\left(\frac{\mu_i^*}{\mu_i^*}\right) - 1\right) = 0.
\end{aligned}$$

The case of $D_{ld}\left(\zeta^k, \zeta^*\right)$ can be similarly treated. □

Lemma 13. *D_f satisfies the condition VIII.*

Proof. Let $H_i = \left\{(A,\xi) \,|\, \langle A, v_i v_i^T\rangle - \xi_i = 0\right\}$. For each $x = (A,\xi) \in \left(S_+^n\right)^o \times \left(D_+^m\right)^o$, let $(B,\zeta) = \arg\min_{z\in H_i} D_f(z,x)$ be the D-projection of (A,ξ) on H_i. Then (B,ζ) is characterized by the KKT condition.

$$\begin{cases} \left(-B^{-1}+A^{-1}, \gamma\left(-\zeta^{-1}+\xi^{-1}\right)\right) & = \lambda\left(v_i v_i^T, -e_i\right) \\ \left\langle B, v_i v_i^T\right\rangle - \zeta_i & = 0 \end{cases}$$

As the proof in the theorem 3, we can utilize the Sherman-Morrison to obtain

$$\lambda = \frac{\gamma}{1+\gamma}\left(\frac{1}{\left\langle A, v_i v_i^T\right\rangle} - \frac{1}{\xi_i}\right)$$
$$B = A + \frac{\lambda}{1-\lambda p} A v_i v_i^T A$$

where $p = \left\langle A, v_i v_i^T\right\rangle$. For any w, we have

$$\begin{aligned} \left\langle B, ww^T\right\rangle &= \left\langle A, ww^T\right\rangle + \frac{\lambda}{1-\lambda p}\left\langle A v_i v_i^T A, ww^T\right\rangle \\ &= \|w\|_A^2 + \frac{\lambda\left(v_i \cdot_A w\right)^2}{1-\lambda\|v_i\|_A^2} \\ &= \frac{\|w\|_A^2 - \lambda\left(\|w\|_A^2\|v_i\|_A^2 - \left(v_i \cdot_A w\right)^2\right)}{1-\lambda\|v_i\|_A^2} \geq 0 \end{aligned}$$

Hence, B is symmetric positive definite which proves our lemma. □

Theorem 4. *Let $x^k = \left(A^k, \xi^k\right)$ be generated from ITML algorithm, then x^k converges to the unique solution of optimization (2).*

Proof. Lemma 9~13 proved that D_f satisfies all the required conditions I~VIII. Then the convergence follows from Theorem 1. □

3.4 Correction on ITML algorithm

ITML algorithm, briefly speaking, is an application of Bregman iteration on the constrained optimization. Figure 1 shows the ITML algorithm that was posted in [4]. Through our mathematical analysis, we found out a typo in the posted algorithm, and put forth a correction in Figure 1. The authors of [4] should have noticed the typo, since their programming code open to public (http://www.cs.utexas.edu/users/pjain/itml/) uses the ITML algorithm with correction. The following section shows that the typo may lead to a great loss of performance.

To our best searches, however, a correction of the typo has not been reported.

4 Numerical results

In this section, several numerical experiments are implemented to verify the necessity of the correction suggested in the previous section. We have tested with three different examples including popular applications of ITML, supervised learning. For the comparison purpose, we implement a corrected ITML algorithm with several value of γ in each example and compare with the case of $\gamma = 1$ which corresponds to the original ITML algorithm.

4.1 Simple 2D example

For the first example, we take a simple two-dimensional example. Given four points $\{(-1,0),(0,-1),(4,3),(2,4)\}$, let $\left\{(-1,0)_{=x_1},(0,-1)_{=x_2}\right\}$ and $\left\{(4,3)_{=x_3},(2,4)_{=x_4}\right\}$ be similar pairs. Let us set $\xi_0 = 0.1$ when $\delta = 1$ and $\xi_0 = 0.9$ when $\delta = -1$. To simply check the constraints, denote two vectors v_1 and v_2 as $v_1 = (x_1 - x_2)$ and $v_2 = (x_3 - x_4)$. With this example, we only compare outputs of two algorithms and satisfaction of constraints.

Testing with two different γ, $\gamma = 0.1$ and $\gamma = 0.01$, we can observe that outputs A are different. Table 1 shows the results computed with the corrected ITML and the original ITML. If we take a large value of γ,such as $\gamma = 10000$, then the output of original ITML gets smaller to be zero. Thus the original ITML with typo is more dependent to given value of γ and it affects to the result.

Input: X: input $d \times n$ matrix, S: set of similar pairs
D: set of dissimilar pairs, u, l: distance thresholds
A_0: input Mahalanobis matrix, γ: slack parameter,
c: constraint index function
Output: A: output Mahalanobis matrix

1. $A \leftarrow A_0,\ \lambda_{ij} \leftarrow 0\ \forall i,j$
2. $\xi_{c(i,j)} \leftarrow u$ for $(i,j) \in S$, otherwise $\xi_{c(i,j)} \leftarrow l$
3. Repeat
   3.1. Pick a constraint $(i,j) \in S$ or $(i,j) \in D$
   3.2. $p \leftarrow (x_i - x_j)^T A (x_i - x_j)$
   3.3. $\delta \leftarrow 1$ if $(i,j) \in S$, -1 otherwise
   3.4. $\alpha \leftarrow \min(\lambda_{ij}, \frac{\delta}{2}\left(\frac{1}{p} - \frac{\gamma}{\xi_{c(i,j)}}\right))$
   3.5. $\beta \leftarrow \gamma\alpha/(1 - \delta\alpha p)$
   3.6. $\xi_{c(i,j)} \leftarrow \gamma\xi_{c(i,j)}/(\gamma + \delta\alpha\xi_{c(i,j)})$
   3.7. $\lambda_{ij} \leftarrow \lambda_{ij} - \alpha$
   3.8. $A \leftarrow A + \beta A(x_i - x_j)(x_i - x_j)^T A$
4. **Until** convergence

Input: X: input $d \times n$ matrix, S: set of similar pairs
D: set of dissimilar pairs, u, l: distance thresholds
A_0: input Mahalanobis matrix, γ: slack parameter,
c: constraint index function
Output: A: output Mahalanobis matrix

1. $A \leftarrow A_0,\ \lambda_{ij} \leftarrow 0\ \forall i,j$
2. $\xi_{c(i,j)} \leftarrow u$ for $(i,j) \in S$, otherwise $\xi_{c(i,j)} \leftarrow l$
3. Repeat
   3.1. Pick a constraint $(i,j) \in S$ or $(i,j) \in D$
   3.2. $p \leftarrow (x_i - x_j)^T A (x_i - x_j)$
   3.3. $\delta \leftarrow 1$ if $(i,j) \in S$, -1 otherwise
   3.4. $\alpha \leftarrow \min(\lambda_{ij}, \frac{\delta\cdot\gamma}{1+\gamma}\left(\frac{1}{p} - \frac{1}{\xi_{c(i,j)}}\right))$
   3.5. $\beta \leftarrow \gamma\alpha/(1 - \delta\alpha p)$
   3.6. $\xi_{c(i,j)} \leftarrow \gamma\xi_{c(i,j)}/(\gamma + \delta\alpha\xi_{c(i,j)})$
   3.7. $\lambda_{ij} \leftarrow \lambda_{ij} - \alpha$
   3.8. $A \leftarrow A + \beta A(x_i - x_j)(x_i - x_j)^T A$
4. **Until** convergence

Figure 1: ITML algorithm posted in the paper [4] (left) and corrected in this manuscript (right). A typo and its correction are put into boxes for the comparison.

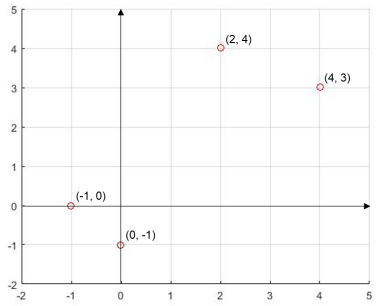

Figure 2: Given dataset $\{(-1,0),(0,-1),(4,3),(2,4)\}$ with $\{(-1,0),(0,-1)\} \in S$ and $\{(4,3),(2,4)\} \in S$.

4.2 KNN classification with Iris dataset

To perform an application of ITML, we take Iris data set [5] for KNN classification. We give fixed 6 pairs for constraints, then set $\xi_0 = 0.5$ when $\delta = 1$ and $\xi_0 = 2$ when $\delta = -1$. For given $\gamma \in [0.001, 100]$, we can easily find that the performance of two algorithms are different. The results are shown in Table 2.

4.3 Face identification

The goal of face identification is to determine well whether two images of faces are the same person or not. In the paper [7], they considered face recognition with the metric learning. Hence face recognition one of the most popular applications for metric learning. We tried to perform similar work with ITML metric learning, using Labeled Faces in the Wild (LFW) data set [8]. There are more than 13,000 images of faces in the LFW data set. We took 1288 samples with 7 classes from the data set that has at least 70 pictures of the same person. For the feature extraction, we applied Principal Component Analysis method (PCA). For ITML metric learning, given the number of constraints is 200. Finally we performed kNN method for classification. We changed γ with various range, and the following results are notable cases. As we can check in the Table 3, there are big differences in the accuracy in both cases.

$\gamma = 0.1$	Original ITML (typo)	Corrected ITML
Output: A	$A = \begin{pmatrix} 0.4667 & 0.2667 \\ 0.2667 & 0.8667 \end{pmatrix}$	$A = \begin{pmatrix} 0.3492 & 0.3318 \\ 0.3318 & 0.6985 \end{pmatrix}$
ξ	$\xi = \begin{pmatrix} 0.1000 \\ 0.1667 \end{pmatrix}$	$\xi = \begin{pmatrix} 0.3841 \\ 0.7682 \end{pmatrix}$
constraints	$\begin{matrix} \langle A, v_1 v_1^T \rangle - \xi_1 & = 0.700000 \geq 0 \\ \langle A, v_2 v_2^T \rangle - \xi_2 & = 1.500000 \geq 0 \end{matrix}$	$\begin{matrix} \langle A, v_1 v_1^T \rangle - \xi_1 & = -0.000001 \leq 0 \\ \langle A, v_2 v_2^T \rangle - \xi_2 & = 0.000000 \leq 0 \end{matrix}$
$\gamma = 0.01$	Original ITML (typo)	Corrected ITML
Output: A	$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$A = \begin{pmatrix} 0.7091 & 0.1703 \\ 0.1703 & 0.8821 \end{pmatrix}$
ξ	$\xi = \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix}$	$\xi = \begin{pmatrix} 1.2507 \\ 3.0374 \end{pmatrix}$
constraints	$\begin{matrix} \langle A, v_1 v_1^T \rangle - \xi_1 & = 1.900000 \geq 0 \\ \langle A, v_2 v_2^T \rangle - \xi_2 & = 4.900000 \geq 0 \end{matrix}$	$\begin{matrix} \langle A, v_1 v_1^T \rangle - \xi_1 & = -0.000023 \leq 0 \\ \langle A, v_2 v_2^T \rangle - \xi_2 & = 0.000000 \leq 0 \end{matrix}$

Table 1: When $\gamma = 0.1$, as shown the above table, constraints are not satisfied in the case of original ITML. In the case of $\gamma = 0.01$, the original ITML algorithm with typo does not work.

accuracy (%)	Original ITML (typo)	Corrected ITML
$\gamma = 0.1$	88.0	91.3
$\gamma = 0.01$	87.3	91.3
$\gamma = 0.001$	87.3	90.7
$\gamma = 10$	90.7	92.0
$\gamma = 50$	85.3	91.3
$\gamma = 100$	82.0	91.3

Table 2: Accuracy of Iris dataset with several value of γ.

5 Conclusion

We provide a mathematical analysis on ITML. The mathematical analysis supports the theoretical foundation of ITML, by supplying well-posedness, strong duality, and convergence. Furthermore we corrected a typo in ITML. Empirical results were presented to show that the typo may mislead to a great loss of performance.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[3] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[4] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[5] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[6] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*, volume 28. Siam, 1999.

$\gamma = 0.01$	Original ITML (typo)	Corrected ITML
accuracy (%)	65.4 ± 0.6	82.9 ± 0.4
$\gamma = 10$	Original ITML (typo)	Corrected ITML
accuracy (%)	61.3 ± 0.7	70.6 ± 0.5

Table 3: Accuracy of face recognition with LFW when $\gamma = 0.01$ and $\gamma = 10$

[7] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV 2009-International Conference on Computer Vision*, pages 498–505. IEEE, 2009.

[8] Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[9] Brian Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

[10] Brian Kulis, Mátyás Sustik, and Inderjit Dhillon. Learning low-rank kernel matrices. In *Proceedings of the 23rd international conference on Machine learning*, pages 505–512. ACM, 2006.

[11] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[12] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[13] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[14] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003.