# Learning Relative Similarity by Stochastic Dual Coordinate Ascent

**Pengcheng Wu[†], Yi Ding[†], Peilin Zhao[‡], Chunyan Miao[†], Steven C.H. Hoi[†]**
[†]School of Computer Engineering, Nanyang Technological University, 639798, Singapore
[‡]Department of Statistics, Rutgers University, Piscataway, NJ, 08854, USA
{pcwu,chhoi,ascymiao}@ntu.edu.sg, ding0077@e.ntu.edu.sg, peilin.zhao@rutgers.edu

**Abstract**

Learning relative similarity from pairwise instances is an important problem in machine learning and has a wide range of applications. Despite being studied for years, some existing methods solved by Stochastic Gradient Descent (SGD) techniques generally suffer from slow convergence. In this paper, we investigate the application of Stochastic Dual Coordinate Ascent (SDCA) technique to tackle the optimization task of relative similarity learning by extending from vector to matrix parameters. Theoretically, we prove the optimal linear convergence rate for the proposed SDCA algorithm, beating the well-known sublinear convergence rate by the previous best metric learning algorithms. Empirically, we conduct extensive experiments on both standard and large-scale data sets to validate the effectiveness of the proposed algorithm for retrieval tasks.

## Introduction

Similarity learning has attracted a significant amount of interests in machine learning community due to its great potential for real-world applications, including image retrieval and classification (Hoi et al. 2006), recommender systems, web search and information retrieval (Wu et al. 2011), etc. A variety of similarity/distance functions have been devised for solving challenges in different domains. The most commonly used examples include cosine similarity or Euclidean distance. The major limitation of such kinds of schemes is that they adopt a rigid similarity/distance function that is usually computed in the original feature space, which may not be optimal or sometimes could be computationally expensive. To overcome the limitation, recent years have witnessed the surge of studies for Distance Metric Learning (DML), which explores machine learning techniques to optimize flexible similarity/distance functions from training data (Yang 2006; Xing et al. 2003; Kwok and Tsang 2003; Yang et al. 2006; Wu et al. 2009a; 2009b; Hoi, Liu, and Chang 2010; Wu et al. 2013; Xia, Wu, and Hoi 2013).

One straightforward approach for similarity learning is to directly learn real-valued pairwise similarity or distance functions from training data that contains explicit similarity/distance values for every pairwise objects. However, in most real-world applications, it is often difficult or expensive to obtain the ground truth with precise numerical values for pairwise similarity/distance.

Instead of learning from explicit similarity/distance values, another more commonly used approach is to learn similarity/distance functions from pairwise relationship which indicates relative similarity of some pairs (Frome et al. 2007). Most previous studies have focused on learning the Mahalanobis distance (Globerson and Roweis 2005; Weinberger, Blitzer, and Saul 2005; Hoi et al. 2006; Xiang, Nie, and Zhang 2008) or the parametric similarity function in a bi-linear form (Chechik et al. 2010). Despite being studied extensively, these existing approaches often have slow convergence rate in theory, and usually suffer from high computational cost empirically, making them often scale poorly large-scale applications.

To tackle the above challenges, we present a new relative similarity learning scheme by extending the Stochastic Dual Coordinate Ascent (SDCA) technique (Shalev-Shwartz and Zhang 2013), a recently proposed promising optimization method that picks a coordinate to update uniformly at random. The proposed SDCA algorithm with uniformly random sampling is able to converge much faster than the existing algorithms. Besides, the proposed approach is computationally efficient as it follows an online learning setting and avoids computing all matrices before training and the expensive retraining costs, making it more scalable and suitable for large-scale machine learning tasks. More importantly, the proposed algorithm enjoys a solid theoretical guarantee in which it can be proved with a linear convergence rate, which is better than the typical sub-linear convergence rate of existing metric learning algorithms using stochastic gradient descent. Finally, we conduct an extensive set of experiments, in which our experimental results show that the proposed SDCA algorithm achieves the state-of-the-art performance when comparing with a family of existing metric learning algorithms.

The rest of this paper is organized as follows. We first review related work, and then present the formulations of the proposed method and its theoretical analysis; we further discuss our experimental results, and finally make the concluding remark at the end.

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

## Related Work

Similarity/distance metric learning has been extensively studied in machine learning community (Yang 2006). Most existing works for DML often focus on learning a Mahalanobis distance parameterized by a positive semidefinite matrix (Shalev-Shwartz, Singer, and Ng 2004; Shental et al. 2002; Schultz and Joachims 2003; Jin, Wang, and Zhou 2009). Inspired by its applications in the context of ranking, the work in (Weinberger, Blitzer, and Saul 2005) addresses the DML problem together with a large margin nearest-neighbor classifier. The study in (Globerson and Roweis 2005) formulated it in a supervised setting by adding positive constraints. The works by (Davis et al. 2007) and (Jain et al. 2008) proposed online metric learning algorithms based on LogDet-regularization with different loss functions. All these approaches focus on the symmetric format: given two images $p_1$ and $p_2$ they measure similarity through $(p_1\text{-}p_2)^\top M(p_1\text{-}p_2)$, where the matrix $M$ must be positive semidefinite. However, imposing the positive semidefinitiveness constraint often results in a computationally expensive optimization task, making it impractical for solving large-scale real applications.

Another popular similarity learning approach aims to optimize an unconstrained similarity function in a bilinear form, such as OASIS (Chechik et al. 2010). Specifically, given two images $p_1$ and $p_2$ they measure similarity by $p_1^\top M p_1$, where matrix $M$ is not required to be positive semi-definite. This kind of measurement is more efficient in real-world applications since it avoids enforcing positive semi-definite constraints when learning the similarity function. Unlike OASIS that uses online passive aggressive algorithms (Crammer et al. 2006), we explore the emerging Stochastic Dual Coordinate Ascent (SDCA) method (Shalev-Shwartz and Zhang 2013) for solving relative similarity learning problem.

In this work, we explore online optimization techniques to learn similarity functions from triplet constraint streams. Online learning works in a sequential fashion, which is efficient and scalable for large-scale applications (Hoi, Wang, and Zhao 2014; Rosenblatt 1958; Cesa-Bianchi and Lugosi 2006; Crammer et al. 2006; Dredze, Crammer, and Pereira 2008; Chechik et al. 2010; Zhao, Hoi, and Jin 2011). In this paper, we extend the SDCA method (Shalev-Shwartz and Zhang 2013) to tackle the optimization task of relative similarity learning in an online learning setting.

## Relative Similarity Learning

### Problem Formulation

Following (Chechik et al. 2010), we would study the problem of learning a relative similarity function $S$. Formally, let $\{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d | i \in [n]\}$ (where $[n] = \{1, \ldots, n\}$) be a set of instances, where the relevance between $\mathbf{x}_i$ and $\mathbf{x}_i^+$ is greater than that between $\mathbf{x}_i$ and $\mathbf{x}_i^-$, our goal is to learn a similarity function $S(\mathbf{x}, \mathbf{x}')$ that assigns higher similarity scores to more relevant instances, i.e.,

$$S(\mathbf{x}_i, \mathbf{x}_i^+) > S(\mathbf{x}_i, \mathbf{x}_i^-), \forall i \in [n]. \quad (1)$$

For the similarity function, we specifically adopt a parametric similarity function that has a bi-linear form,

$$S_M(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top M \mathbf{x}' \quad (2)$$

where $M \in \mathbb{R}^{d \times d}$. In order to learn the optimal parameter $M$, we introduce some loss function that measures its performance on the $i$-th triplet:

$$\ell(M; (\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)) = ([1 - S_M(\mathbf{x}_i, \mathbf{x}_i^+) + S_M(\mathbf{x}_i, \mathbf{x}_i^-)]_+)^2 \quad (3)$$

where $[\cdot]_+ = \max(0, \cdot)$. The above loss measures how much is the violation of the desired constraint $S_M(\mathbf{x}_i, \mathbf{x}_i^+) \geq S_M(\mathbf{x}_i, \mathbf{x}_i^-)$ by the similarity function defined by $M$.

With the above loss function, we formulate the relative similarity learning problem as a regularized optimization:

$$\min_{M \succeq 0} P(M) := \left[ \frac{1}{n} \sum_{i=1}^{n} \ell_i(M) + \frac{\lambda}{2} \|M\|^2 \right] \quad (4)$$

where $\ell_i(M) = \ell(M; (\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-))$, $\lambda$ is a regularization parameter, and $\|M\|^2 = \mathrm{trace}(M^\top M)$ is Frobenius norm.

In literature, different approaches have been proposed to tackle the similar optimization task in (4). For example, in (Chechik et al. 2010), the authors proposed OASIS — an online learning scheme to solve the problem (4) by applying the online Passive Aggressive (PA) learning (Crammer et al. 2006), which proved some mistake bound but did not give convergence rate. Another popular method to solve this problem is based on the Stochastic Gradient Descent (SGD) method (Zinkevich 2003; Zhang 2004), which usually achieves a sub-linear convergence rate.

### Algorithm

To tackle the optimization of relative similarity learning, we explore the application of the Stochastic Dual Coordinate Ascent (SDCA) method (Shalev-Shwartz and Zhang 2013), which guarantees a linear convergence rate. Unlike the existing study of SDCA which only handled vector parametric problems, we extend the SDCA method to handle more complicated problem with matrix-based parameters.

Specifically, the Dual Coordinate Ascent (DCA) method aims to solve the dual problem of (4) as follows:

$$\max_{\Theta = (\Theta_i)_{i=1}^n,} D(\Theta) := \left[ \frac{1}{n} \sum_{i=1}^{n} -\ell_i^*(-\Theta_i) - \frac{\lambda}{2} \Big\| \frac{1}{\lambda n} \sum_{i=1}^{n} \Theta_i \Big\|^2 \right] \quad (5)$$

where $\Theta_i \in \mathbb{R}^{d \times d}, i = 1, \ldots, n$ is a dual variable associated with every triplet instance $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$, $\ell_i^*(U) = \max_V [\langle U, V \rangle - \ell(V)]$, and $\langle U, V \rangle = trace(U^\top V)$.

It is difficult to directly solve the dual objective in (5) as there is a different dual variable associated with each example in the training set. The idea of stochastic DCA is to pick up a dual variable at each iteration, and then optimize the dual objective with respect to the single dual variable, while the rest of the dual variables are kept in tact.

To solve the relative similarity learning problem, we extend the Stochastic Dual Coordinate Ascent (SDCA) method (Shalev-Shwartz and Zhang 2013). At each iteration, we choose one dual coordinate to optimize uniformly

at random, in which each dual variable is associated with some triplet instance $\{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$. Specifically, given a uniformly sampled triplet instance $\{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$, the $i$-th dual coordinate at each iteration can be optimized by:

$$\Delta\Theta_i^{t-1} = \arg\max_{\Delta\Theta_i} -\ell_i^*(-(\Theta_i^{t-1} + \Delta\Theta_i)) - \frac{\lambda n}{2}\|M^{t-1} + (\lambda n)^{-1}\Delta\Theta_i\|^2 \quad (6)$$

Given the loss function in (3), it is not difficult to show that it is $(2\|X_i\|^2)$-smooth (refer to Definition 1 below), where $X_i = \mathbf{x}_i(\mathbf{x}_i^+ - \mathbf{x}_i^-)^\top$, and its dual function is

$$\ell_i^*(-\Theta_i) = \begin{cases} -\alpha + \alpha^2/4 & \Theta_i = \alpha X_i, \alpha \geq 0 \\ \infty & \text{otherwise} \end{cases}$$

Thus, we can derive the closed-form solution of $\Delta\Theta_i$

$$\Delta\Theta_i = \max\left(\frac{1 - \mathbf{x}_i^\top M(\mathbf{x}_i^+ - \mathbf{x}_i^-) - \frac{\alpha_i}{2}}{\frac{1}{2} + (\lambda n)^{-1}\|X_i\|^2}, -\alpha_i\right) X_i.$$

The details of the proposed algorithm for relative similarity learning are summarized in Algorithm 1.

**Algorithm 1** SDCA: Stochastic Dual Coordinate Ascent for Relative Similarity Learning

**Input**: $\lambda > 0$, $\{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d | i \in [n]\}$
**Initialize**: $M = 0$, $\Theta_1, \ldots, \Theta_n = 0$, $\alpha_1, \ldots, \alpha_n = 0$
**for** $t = 1, \ldots, T$ **do**
  Uniformly sample a triplet instance $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$

$$\Delta\alpha_i^{t-1} = \max\left(\frac{1 - \mathbf{x}_i^\top M^{t-1}(\mathbf{x}_i^+ - \mathbf{x}_i^-) - \frac{\alpha_i^{t-1}}{2}}{\frac{1}{2} + (\lambda n)^{-1}\|X_i\|^2}, -\alpha_i^{t-1}\right)$$

  $\Delta\Theta_i^{t-1} = \Delta\alpha_i^{t-1} X_i$;
  $\alpha_i^t = \alpha_i^{t-1} + \Delta\alpha_i^{t-1}$;
  $\Theta_i^t = \Theta_i^{t-1} + \Delta\Theta_i^{t-1}$;
  $M^t = M^{t-1} + (\lambda n)^{-1}\Delta\Theta_i^{t-1}$;
**end for**
**Output (Average option):**
  Let $\bar{\Theta} = \frac{1}{T - T_0}\sum_{t=T_0+1}^{T} \Theta^{t-1}$
  Let $\bar{M} = M(\bar{\Theta}) = \frac{1}{T - T_0}\sum_{t=T_0+1}^{T} M^{t-1}$
  Return $\bar{M}$
**Output (Random option) :**
  Let $\bar{\Theta} = \Theta^t$ and $\bar{M} = M^t$ for uniformly random $t \in \{T_0 + 1, \ldots, T\}$
  Return $\bar{M}$

## Theoretical Analysis

We analyze the theoretical performance of Algorithm 1. For simplicity, we only consider $(1/\gamma)$-smooth loss functions in this paper, which is defined as follows.

**Definition 1.** *A function $\ell : \mathbb{R}^{d\times d} \to \mathbb{R}$ is $(1/\gamma)$-smooth if that for all $U, V \in \mathbb{R}^{d\times d}$, we have*

$$\ell(U) \leq \ell(V) + \langle \ell'(V), (U - V)\rangle + \frac{1}{2\gamma}\|U - V\|^2$$

*where $\ell'$ is the derivative of $\ell$. It is well-known that if $\ell$ is $(1/\gamma)$-smooth, then $\ell^*$ is $\gamma$-strongly convex, that is, for all $U, V \in \mathbb{R}^{d\times d}$ and $s \in [0, 1]$, we have*

$$\ell^*(sU + (1-s)V) \leq \ell^*(U) + (1-s)\big(\ell^*(V) - \frac{\gamma s\|U - V\|^2}{2}\big).$$

If we define the following

$$M(\Theta) = \frac{1}{\lambda n}\sum_{i=1}^{n} \Theta_i \quad (7)$$

then it is known that $M(\Theta^*) = M^*$, where $\Theta^*$ is an optimal solution of (5). It is also known that $P(M^*) = D(\Theta^*)$ which immediately implies that for all $M$ and $\Theta$, we have $P(M) \geq D(\Theta)$, and hence the duality gap defined as

$$P(M(\Theta)) - D(\Theta) \quad (8)$$

can be regarded as an upper bound of the primal sub-optimality $P(M(\Theta)) - P(M^*)$.

We first make some assumptions on the loss function without loss of generality: 1) $\ell_i(M) \geq 0$ for any $i \in [n]$ and $M \in \mathbb{R}^{d\times d}$; 2) $\ell_i(0) \leq 1$ for any $i \in [n]$.

Given these assumptions, we first present the following lemma to facilitate our proof of the convergence rate for the proposed algorithm. This lemma generally gives an upper bound for the dual ascent at the t-th step based on the duality gap at the (t-1)-th step plus a variance for the subgradients.

**Lemma 1.** *Assume $\ell_i^*$ is $\gamma$-strongly convex (where $\gamma$ can be zero). Then, for any iteration $t$ and any $s \in [0, 1]$, we have*

$$\mathbb{E}[D(\Theta^t) - D(\Theta^{t-1})] \geq \frac{s}{n}\mathbb{E}[P(M^{t-1}) - D(\Theta^{t-1})] - \frac{s}{2\lambda n^2}G^t \quad (9)$$

*where $G^t = \frac{1}{n}\sum_{i=1}^{n}(s - \gamma(1-s)\lambda n)\,\mathbb{E}\|U_i - \Theta_i\|^2$, and $-U_i^{t-1} \in \partial\ell_i(M^{t-1})$.*

*Proof.* Since only the $i$-th element of $\Theta_i$ is updated, the improvement in the dual objective can be written as

$$n[D(\Theta^t) - D(\Theta^{t-1})] = \underbrace{\left(-\ell_i^*(-\Theta^t) - \frac{\lambda n}{2}\|M^t\|^2\right)}_{A} - \underbrace{\left(-\ell_i^*(-\Theta^{t-1}) - \frac{\lambda n}{2}\|M^{t-1}\|^2\right)}_{B}$$

By the definition of the update, for all $s \in [0, 1]$ we have

$$\begin{aligned} A &= \max_{\Delta\Theta_i} -\ell_i^*(-(\Theta_i^{t-1} + \Delta\Theta_i)) - \frac{\lambda n}{2}\|M^{t-1} + (\lambda n)^{-1}\Delta\Theta_i\|^2 \\ &\geq -\ell_i^*(-(\Theta_i^{t-1} + s(U_i^{t-1} - \Theta_i^{t-1}))) - \frac{\lambda n}{2}\|M^{t-1} + (\lambda n)^{-1}s(U_i^{t-1} - \Theta_i^{t-1})\|^2 \quad (10) \end{aligned}$$

From now on, we omit the superscripts and subscripts. Since $\ell^*$ is $\gamma$-strongly convex, we have that

$$\ell^*(-(\Theta + s(U - \Theta))) = \ell^*(s(-U) + (1-s)(-\Theta)) \leq s\ell^*(-U) + (1-s)\ell^*(-\Theta) - \frac{\gamma}{2}s(1-s)\|U - \Theta\|^2 \quad (11)$$

Combining this with (10) and rearranging the terms gives

$$\begin{aligned} A &\geq -s\ell^*(-U) - (1-s)\ell^*(-\Theta) + \\ &\quad \frac{\gamma}{2}s(1-s)\|U-\Theta\|^2 - \frac{\lambda n}{2}\|M + (\lambda n)^{-1}s(U-\Theta)\|^2 \\ &= -s\ell^*(-U) - (1-s)\ell^*(-\Theta) + \frac{\gamma}{2}s(1-s)\|U-\Theta\|^2 \\ &\quad -\frac{\lambda n}{2}\|M\|^2 - s\langle M, U-\Theta\rangle - \frac{s^2}{2\lambda n}\|U-\Theta\|^2 \\ &= \underbrace{-s(\ell^*(-U) + \langle M, U\rangle)}_{s\ell(M)} + \underbrace{(-\ell^*(-\Theta) - \frac{\lambda n}{2}\|M\|^2)}_{B} + \\ &\quad s(\ell^*(-\Theta) + \langle M, \Theta\rangle) + \frac{s}{2}\left(\gamma(1-s) - \frac{s}{\lambda n}\right)\|U-\Theta\|^2 \end{aligned}$$

where we used $-U \in \partial\ell(M)$ which yields $\ell^*(-U) = -\langle U, M\rangle - \ell(M)$. Therefore,

$$\begin{aligned} A - B \geq s[&\ell(M) + \ell^*(-\Theta) + \langle M, \Theta\rangle + \\ &\left(\frac{\gamma(1-s)}{2} - \frac{s}{2\lambda n}\right)\|U-\Theta\|^2] \end{aligned}$$

Next note that

$$\begin{aligned} P(M) - D(\Theta) &= \frac{1}{n}\sum_{i=1}^n \ell_i(M) + \frac{\lambda}{2}\langle M, M\rangle - \\ &\quad \left(-\frac{1}{n}\sum_{i=1}^n \ell^*(-\Theta_i) - \frac{\lambda}{2}\langle M, M\rangle\right) \\ &= \frac{1}{n}\sum_{i=1}^n (\ell_i(M) + \ell_i^*(-\Theta_i) + \langle M, \Theta_i\rangle) \end{aligned}$$

Therefore, if we take expectation of $(A-B)$ with respect to the choice of $i$, we will obtain that

$$\begin{aligned} &\frac{1}{s}\mathbb{E}[A-B] = \frac{1}{n}\sum_{i=1}^n [\ell_i(M) + \ell_i^*(-\Theta_i) + \\ &\langle M, \Theta_i\rangle + \left(\frac{\gamma(1-s)}{2} - \frac{s}{2\lambda n}\right)\|U_i - \Theta_i\|^2] \end{aligned}$$

which indicates

$$\frac{n}{s}\mathbb{E}[D(\Theta^t) - D(\Theta^{t-1})] \geq \mathbb{E}[P(M^{t-1}) - D(\Theta^{t-1})] - \frac{G^t}{2\lambda n}$$

Multiplying both sides of the above inequalities will conclude the proof. □

One disadvantage of Lemma 1 is that it contains a term $G^t$ which is hard to estimate. To solve this issue, we would try to estimate one lower bound of the right hand side of (9), which is independent of $G^t$ and as large as possible. Specifically, we propose to estimate as follows:

$$\begin{aligned} &\max_{s\in[0,1]} \frac{s}{n}\mathbb{E}[P(M^{t-1}) - D(\Theta^{t-1})] - \frac{s}{2\lambda n^2}G^t \\ &\geq \max_{s\in[0,\frac{\lambda n\gamma}{1+\lambda n\gamma}]} \frac{s}{n}\mathbb{E}[P(M^{t-1}) - D(\Theta^{t-1})] - \frac{s}{2\lambda n^2}G^t \\ &\geq \max_{s\in[0,\frac{\lambda n\gamma}{1+\lambda n\gamma}]} \frac{s}{n}\mathbb{E}[P(M^{t-1}) - D(\Theta^{t-1})] = \\ &\frac{s^*}{n}\mathbb{E}[P(M^{t-1}) - D(\Theta^{t-1})] \end{aligned}$$

where the last inequality used $G^t \leq 0$, and $s^* = \frac{\lambda n\gamma}{1+\lambda n\gamma}$.

Now, we will analyze the convergence rate of the proposed algorithm. Specifically, for the expected duality gap of $\mathbb{E}[P(M^T) - D(\Theta^T)]$, we have the following theorem.

**Theorem 1.** *Assume $\ell_i(\cdot)$ is $(1/\gamma)$-smooth $\forall i \in [n]$. To obtain an expected duality gap of $\mathbb{E}[P(M^T) - D(\Theta^T)] \leq \epsilon_P$, it suffices to have a total number of iterations of*

$$T \geq (n + \frac{1}{\lambda\gamma})\log\left((n + \frac{1}{\lambda\gamma})\frac{1}{\epsilon_P}\right) \tag{12}$$

*Moreover, to obtain an expected duality gap of $\mathbb{E}[P(\bar{M}) - D(\bar{\Theta})] \leq \epsilon_P$, it suffices to have a total number of iterations of $T > T_0$ where*

$$T_0 \geq (n + \frac{1}{\lambda\gamma})\log\left((n + \frac{1}{\lambda\gamma})\frac{1}{(T-T_0)\epsilon_P}\right) \tag{13}$$

*Proof.* Since $\ell_i$ is $(1/\gamma)$-smooth, its dual $\gamma$-strongly convex. Then according to Lemma 1, if we set $s^* = \frac{\lambda n\gamma}{1+\lambda n\gamma}$ we have

$$\mathbb{E}[D(\Theta^t) - D(\Theta^{t-1})] \geq \frac{s^*}{n}\mathbb{E}[P(M^{t-1}) - D(\Theta^{t-1})] \tag{14}$$

since $G^t \leq 0$. Furthermore since

$$\epsilon_D^{t-1} := D(\Theta^*) - D(\Theta^{t-1}) \leq P(M^{t-1}) - D(\Theta^{t-1}) \tag{15}$$

where $\Theta^*$ is the optimal solution for the dual problem, and $D(\Theta^t) - D(\Theta^{t-1}) = \epsilon_D^{t-1} - \epsilon_D^t$, we obtain that

$$\mathbb{E}[\epsilon_D^t] \leq (1 - \frac{s^*}{n})\mathbb{E}[\epsilon_D^{t-1}] \leq (1 - \frac{s^*}{n})^t\mathbb{E}[\epsilon_D^0] \tag{16}$$

In addition, since $P(0) = \frac{1}{n}\ell_i(0) \leq 1$ and

$$\begin{aligned} D(0) &= \frac{1}{n}\sum_{i=1}^n -\ell_i^*(0) = \frac{1}{n}\sum_{i=1}^n -\max_M(0 - \ell_i(M)) = \\ &\quad \frac{1}{n}\sum_{i=1}^n \min_M \ell_i(M) \geq 0 \end{aligned} \tag{17}$$

we have $\epsilon_D^0 \leq P(0) - D(0) \leq 1$. Combining this with inequality (16), we obtain

$$\mathbb{E}[\epsilon_D^t] le (1 - \frac{s^*}{n})^t \leq \exp(-\frac{s^* t}{n}) = \exp(-\frac{\lambda\gamma t}{1+\lambda\gamma n})$$

According to the above inequality, by setting

$$t \geq (n + \frac{1}{\lambda\gamma})\log(1/\epsilon_D)$$

we will get $\mathbb{E}[\epsilon_D^t] \leq \epsilon_D$. Furthermore, according to inequality (14),

$$\mathbb{E}[P(M^t) - D(\Theta^t)] \leq \frac{n}{s^*}\mathbb{E}[\epsilon_D^t - \epsilon_D^{t+1}] \leq \frac{n}{s^*}\mathbb{E}[\epsilon_D^t], \tag{18}$$

by setting

$$t \geq (n + \frac{1}{\lambda\gamma})\log\left((n + \frac{1}{\lambda\gamma})\frac{1}{\epsilon_P}\right),$$

we will have $\mathbb{E}[\epsilon_D^t] \leq \frac{s^*}{n}\epsilon_P$ and $\mathbb{E}[P(M^t) - D(\Theta^t)] \leq \epsilon_P$.

Summing inequality (18) over $t = T_0, \ldots, T-1$ leads to

$$\mathbb{E}\left[\frac{1}{T-T_0}\sum_{t=T_0}^{T-1}\left(P(M^t) - D(\Theta^t)\right)\right] \leq \frac{n}{s^*(T-T_0)}\mathbb{E}[D(\Theta^T) - D(\Theta^{T_0})]$$

Now, if we choose $\bar{M}, \bar{\Theta}$ to be either the average matrix or a randomly chosen matrix over $t \in \{T_0+1, \ldots, T\}$, then the above implies

$$\mathbb{E}[P(\bar{M}) - D(\bar{\Theta})] \leq \frac{n}{s(T-T_0)}\mathbb{E}[D(\Theta^T) - D(\Theta^{T_0})] \leq \frac{n}{s(T-T_0)}\mathbb{E}[\epsilon_D^{T_0}]$$

It follows that in order to obtain a result of $\mathbb{E}[P(\bar{M}) - D(\bar{\Theta})] \leq \epsilon_P$, we only need to have

$$\mathbb{E}[\epsilon_D^{T_0}] \leq \frac{s(T-T_0)}{n}\epsilon_P = \frac{T-T_0}{n+1/\lambda\gamma}\epsilon_P$$

This implies the second part of the theorem. □

*Remark.* If we choose $T = 2T_0$, and assume that $T_0 \geq n + 1/(\lambda\gamma)$, then the second part of the above theorem indicates a requirement of $T_0 \geq (n + \frac{1}{\lambda\gamma})\log(\frac{1}{\epsilon_P})$, which is slightly smaller than the first part of the above theorem, when $\epsilon_P$ is relative large.

## Experiments

In this section, we conduct comprehensive experiments on different datasets to evaluate the efficacy of our proposed algorithms for relative similarity learning.

### Experimental Testbed and Setup

We first conduct experiments of similarity/distance metric learning on five standard machine learning datasets publicly available at LIBSVM [1], as shown in Table 1.

Table 1: Details of Machine Learning Datasets.

| Dataset. | Source | Class | Feature | Size |
|---|---|---|---|---|
| Vehicle | Statlog | 4 | 18 | 846 |
| Vowel | UCI | 11 | 10 | 990 |
| Letter | Statlog | 26 | 16 | 20000 |
| Svmguide4 | CWH03a | 6 | 6 | 612 |
| Segment | Statlog | 7 | 19 | 1937 |

For each dataset in Table 1, data instances from each class were split into training set (70%) and test set (30%). We adopt cross-validation to choose parameters for all algorithms, in which models were learned on 80% of the training set and validated on the rest 20%. The parameters set by cross validation include: the $\lambda$ parameter for SDCA ($\lambda \in \{0.0025, 0.005, 0.01\}$), and the $\eta$ parameter for ITML and LEGO ($\eta \in \{0.01, 0.125, 0.5\}$). Results reported below were achieved by choosing the best value of the parameter by cross validation. To adapt to image retrieval task, we also carried out experiments on Caltech256 dataset. To examine the scalability, we also test the algorithms on a large-scale image dataset.

To obtain side information in the form of triplets for learning similarity function, we generate a triplet instance by randomly sampling two instances sharing the same class and another one instance from any other different class. In total, we provide 10K triplet instances for each standard data set, 100K triplets for Caltech256 Dataset and 500K triplets for the large-scale experiment. We evaluate different algorithms fairly and adopt the standard mean Average Precision (mAP) to evaluate the retrieval performance.

### Comparison Algorithms

We compared the following list of algorithms:

- Eucl: Baseline method using the standard Euclidean distance in feature space.
- RCA: Relevance Component Analysis that learns a linear projection from equivalent constraints (Bar-Hillel et al. 2005).
- LMNN: Largest Margin Nearest Neighbor (Weinberger, Blitzer, and Saul 2005) in which k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin.
- ITML: Information Theoretic Metric Learning with the goal that minimizes the differential relative entropy between two multivariate Gaussians under constraints on the distance function. (Davis et al. 2007).
- LEGO: Similar to online ITML with a different loss function to ensure the positive definiteness for the learned matrix (Jain et al. 2008).
- OASIS. A bilinear similarity learning approach based on online Passive Aggressive algorithm using triplet instances (Chechik et al. 2010).
- SDCA: The proposed similarity learning algorithm.

### Evaluation on the Standard Datasets

We first compare the performance of the proposed algorithm with six other approaches for similarity search on the standard data sets as shown in Table 1. All the experiments were conducted by fixing 5 different random seeds for each data set, and all the results were reported by averaging over 5 runs. The accuracy results represented by mean Average Precision (mAP) are shown in Table 2, from which we observe that all learning approaches achieve better performance than the Euclidean baseline, showing the ability of distance metric learning. On all standard datasets, the proposed SCDA algorithm outperforms all other algorithms.

Table 2: Evaluation of mAP on standard datasets.

| Alg. | vehicle | vowel | letter | svmguide4 | segment |
|---|---|---|---|---|---|
| Eucl | 0.3697 | 0.2920 | 0.2135 | 0.2088 | 0.6645 |
| RCA | 0.4039 | 0.3184 | 0.2297 | 0.2181 | 0.6781 |
| LMNN | 0.3935 | 0.3144 | 0.2411 | 0.2127 | 0.6722 |
| ITML | 0.4142 | 0.3074 | 0.2255 | 0.2236 | 0.6895 |
| LEGO | 0.4415 | 0.3143 | 0.2238 | 0.2186 | 0.6857 |
| OASIS | 0.5318 | 0.3383 | 0.2531 | 0.2295 | 0.6970 |
| **SDCA** | **0.5955** | **0.3564** | **0.2806** | **0.2731** | **0.7468** |

[1] http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/

### Evaluation on the Caltech256 Dataset

We also evaluate the proposed algorithm in an image similarity task with the well-known Caltech256 data set. To better show the direct comparison results with the previous literature, we use the same 50 classes to generate 100K triplet instances and extract same image representation fashion with (Chechik et al. 2010), which utilized *bag-of-local-descriptors* with a 1000-sized codebook.

Table 3: Evaluation of mAP performance on "Caltech256".

| Eucl | RCA | ITML | LEGO | OASIS | **SDCA** |
|---|---|---|---|---|---|
| 0.0924 | 0.0957 | 0.1009 | 0.1021 | 0.1155 | **0.1237** |

In this experiment, we exclude LMNN due to its extremely high computational cost. From Table 3, we observe that the proposed SDCA algorithm also outperforms other approaches on image retrieval task. Then, we show the convergence rate described by accumulative loss divided by the number of iterations below. Since the loss functions for RCA, LMNN, ITML and LEGO algorithms are not represented in the format of bilinear similarity, we just show the loss convergence behaviors between OASIS and SDCA.

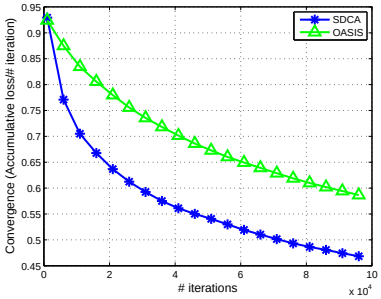


Figure 1: Average loss convergence comparison.

From Figure 1, we observe that using the same bilinear loss function, the proposed SDCA algorithm achieved a faster convergence rate than OASIS. After learning about 100k iterations, the proposed algorithm can converge to about 0.46, while OASIS can only converge to about 0.58. Therefore, our experiment validates that the proposed SDCA algorithm has a fast convergence rate as indicated by our theoretical analysis. To further confirm SDCA's fast convergence rate, we also evaluate the online mistake rates by both OASIS and SDCA shown in Figure 2. From the comparison in Figure 1, we observe that the decreasing trend of mistake rate is consistent with that of average loss convergence rate, which again validates that SDCA is able achieve a faster convergence and lower online mistake rate than OASIS.

### Evaluation on the Large-scale Dataset

To evaluate the scalability of the proposed algorithm, we apply the proposed algorithms on a large-scale image retrieval dataset named "ImageCLEF+Flickr", which includes the public dataset "ImageCLEF" [2] as the ground-truth and extra one-million images crawled from Flickr as the background. We generate 500K triplets in training stage and extract 297-dimensional global feature descriptors to represent images including color histogram and moments (81 dimensions), edge direction histogram (37 dimensions), Gabor wavelets transformation (120 dimensions) and local binary pattern(59 dimensions). The whole dataset is split into 60% for training and 40% for testing. 20% of the training set are utilized for cross validation, and 10% of the testing set are used as queries to evaluate mAP performance.

[2]http://www.imageclef.org/

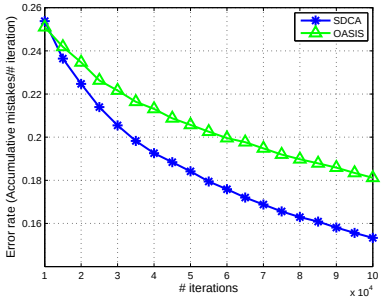


Figure 2: Online mistake rate comparison.

Table 4: Evaluation of mAP on "ImageCLEF+Flickr".

| Eucl | RCA | ITML | LEGO | OASIS | **SDCA** |
|---|---|---|---|---|---|
| 0.3305 | 0.3701 | 0.4019 | 0.3801 | 0.4088 | **0.4378** |

Table 4 shows the mAP performance of the six algorithms. From the results, we can again observe that the proposed algorithm consistently outperforms the state-of-the-art methods on this large-scale dataset. These encouraging results validate that the proposed SDCA algorithm not only enjoys an attractive convergence rate in theory, and empirically but also achieves the state-of-the-art accuracy on different-scale data sets, making it a practical solution for large-scale applications.

## Conclusions

This paper presented a novel scheme for relative similarity learning by extending the emerging Stochastic Dual Coordinate Ascent (SDCA) technique for online optimization of a bi-linear similarity function. In contrast to many existing solutions for similarity/distance metric learning that often have sub-linear convergence rates, we show that the proposed SDCA algorithm achieves a linear convergence rate in theory. We conducted an extensive set of experiments by comparing a number of state-of-the-art similarity/ditance learning techniques, in which the encouraging results showed that the proposed SDCA algorithm is able to achieve the state-of-the-art performance on various benchmark datasets, validating the effectiveness of the proposed technique for relative similarity learning. Future work can explore more other real-world applications of the proposed technique.

## Acknowledgments

The first three authors contributed equally. This research was supported in part by MOE Tier 1 Grant (RG33/11), and Interactive and Digital Media Programme Office, National Research Foundation hosted at Media Development Authority of Singapore (Grant No.: MDA/IDM/2012/8/8-2 VOL 01).

## References


Bar-Hillel, A.; Hertz, T.; Shental, N.; and Weinshall, D. 2005. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research* 6:937–965.

Cesa-Bianchi, N., and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge University Press.

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

Crammer, K.; Dekel, O.; Keshet, J.; Shalev-Shwartz, S.; and Singer, Y. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research* 7:551–585.

Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *ICML*, 209–216.

Dredze, M.; Crammer, K.; and Pereira, F. 2008. Confidence-weighted linear classification. In *ICML*, 264–271.

Frome, A.; Singer, Y.; Sha, F.; and Malik, J. 2007. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 1–8.

Globerson, A., and Roweis, S. T. 2005. Metric learning by collapsing classes. In *NIPS*.

Hoi, S. C.; Liu, W.; Lyu, M. R.; and Ma, W.-Y. 2006. Learning distance metrics with contextual constraints for image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Hoi, S. C.; Liu, W.; and Chang, S.-F. 2010. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)* 6(3):18.

Hoi, S. C.; Wang, J.; and Zhao, P. 2014. Libol: A library for online learning algorithms. *Journal of Machine Learning Research* 15:495–499.

Jain, P.; Kulis, B.; Dhillon, I. S.; and Grauman, K. 2008. Online metric learning and fast similarity search. In *NIPS*, 761–768.

Jin, R.; Wang, S.; and Zhou, Y. 2009. Regularized distance metric learning: Theory and algorithm. In *Advances in neural information processing systems*, 862–870.

Kwok, J. T., and Tsang, I. W. 2003. Learning with idealized kernels. In Fawcett, T., and Mishra, N., eds., *Proceedings of the 20th international conference on Machine learning*, ICML '03, 400–407. AAAI Press.

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65(6):386.

Schultz, M., and Joachims, T. 2003. Learning a distance metric from relative comparisons. In *NIPS*.

Shalev-Shwartz, S., and Zhang, T. 2013. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research* 14(1):567–599.

Shalev-Shwartz, S.; Singer, Y.; and Ng, A. Y. 2004. Online and batch learning of pseudo-metrics. In *ICML*.

Shental, N.; Hertz, T.; Weinshall, D.; and Pavel, M. 2002. Adjustment learning and relevant component analysis. In *ECCV (4)*, 776–792.

Weinberger, K. Q.; Blitzer, J.; and Saul, L. K. 2005. Distance metric learning for large margin nearest neighbor classification. In *NIPS*.

Wu, L.; Hoi, S. C.; Jin, R.; Zhu, J.; and Yu, N. 2009a. Distance metric learning from uncertain side information with application to automated photo tagging. In *Proceedings of the 17th ACM international conference on Multimedia*, 135–144. ACM.

Wu, L.; Jin, R.; Hoi, S. C.; Zhu, J.; and Yu, N. 2009b. Learning bregman distance functions and its application for semi-supervised clustering. In *NIPS*, 2089–2097.

Wu, P.; Hoi, S. C.-H.; Zhao, P.; and He, Y. 2011. Mining social images with distance metric learning for automated image tagging. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 197–206. ACM.

Wu, P.; Hoi, S. C.; Xia, H.; Zhao, P.; Wang, D.; and Miao, C. 2013. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, 153–162. ACM.

Xia, H.; Wu, P.; and Hoi, S. C. 2013. Online multi-modal distance learning for scalable multimedia retrieval. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 455–464. ACM.

Xiang, S.; Nie, F.; and Zhang, C. 2008. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition* 41:3600–3612.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. J. 2003. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, NIPS '02, 505–512.

Yang, L.; Jin, R.; Sukthankar, R.; and Liu, Y. 2006. An efficient algorithm for local distance metric learning. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*, AAAI 06', 543–548. AAAI Press.

Yang, L. 2006. Distance metric learning: A comprehensive survey.

Zhang, T. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, 116. ACM.

Zhao, P.; Hoi, S. C. H.; and Jin, R. 2011. Double updating online learning. *Journal of Machine Learning Research* 12:1587–1615.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*.