

On Geometric Alignment in Low Doubling Dimension

Hu Ding

School of Computer Science and Technology
School of Data Science
University of Science and Technology of China
He Fei, China, 230026
huding@ustc.edu.cn

Mingquan Ye

Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA, 48824
yemingqu@msu.edu

Abstract

In real-world, many problems can be formulated as the alignment between two geometric patterns. Previously, a great amount of research focus on the alignment of 2D or 3D patterns, especially in the field of computer vision. Recently, the alignment of geometric patterns in high dimension finds several novel applications, and has attracted more and more attentions. However, the research is still rather limited in terms of algorithms. To the best of our knowledge, most existing approaches for high dimensional alignment are just simple extensions of their counterparts for 2D and 3D cases, and often suffer from the issues such as high complexities. In this paper, we propose an effective framework to compress the high dimensional geometric patterns and approximately preserve the alignment quality. As a consequence, existing alignment approach can be applied to the compressed geometric patterns and thus the time complexity is significantly reduced. Our idea is inspired by the observation that high dimensional data often has a low intrinsic dimension. We adopt the widely used notion “doubling dimension” to measure the extents of our compression and the resulting approximation. Finally, we test our method on both random and real datasets; the experimental results reveal that running the alignment algorithm on compressed patterns can achieve similar qualities, comparing with the results on the original patterns, but the running times (including the times cost for compression) are substantially lower.

1 Introduction

Given two geometric patterns, the problem of alignment is to find their appropriate spatial positions so as to minimize the difference between them. In general, a geometric pattern is represented by a set of (weighted) points in the space, and their difference is often measured by some objective function. In particular, geometric alignment finds many applications in the field of computer vision, such as image retrieval, pattern recognition, fingerprint and facial shape alignment, etc (Cohen and Guibas 1999; Maltoni et al. 2009; Cao et al. 2014). For different applications, we may have different constraints for the alignment, e.g., we allow rigid transformations for fingerprint alignment. In addition, Earth Mover’s Distance (EMD) (Rubner, Tomasi, and Guibas

2000) has been widely adopted as the metric for measuring the difference of patterns in computer vision, where its major advantage over other measures is the robustness with respect to noise in practice. Besides the computer vision applications in 2D or 3D, recent research shows that a number of high dimensional problems can be solved by geometric alignments. We briefly introduce several interesting examples below.

(1) The research on natural language processing has revealed that different languages often share some similar structure at the word level (Youn et al. 2016); in particular, the recent study on word semantics embedding has also shown the existence of structural isomorphism across languages (Mikolov, Le, and Sutskever 2013), and further finds that EMD can serve as a good distance for languages or documents (Zhang et al. 2017; Kusner et al. 2015). Therefore, (Zhang et al. 2017) proposed to learn the transformation between different languages without any cross-lingual supervision, and the problem is reduced to minimizing the EMD via finding the optimal geometric alignment in high dimension. (2) A Protein-Protein Interaction (PPI) network is a graph representing the interactions among proteins. Given two PPI networks, finding their alignment is a fundamental bioinformatics problem for understanding the correspondences between different species (Malod-Dognin, Ban, and Pržulj 2017). However, most existing approaches require to solve the NP-hard subgraph isomorphism problem and often suffer from high computational complexities. To resolve this issue, (Liu et al. 2017) recently applied the geometric embedding techniques to develop a new framework based on geometric alignment in Euclidean space. (3) Other applications of high dimensional alignment include domain adaptation and indoor localization. Domain adaptation is an important problem in machine learning, where the goal is to predict the annotations of a given unlabeled dataset by determining the transformation (in the form of EMD) from a labeled dataset (Pan and Yang 2010); as the rapid development of wireless technology, indoor localization becomes rather important for locating a person or device inside a building. Recent studies show that they both can be well formulated as geometric alignments in high dimension. We refer the reader to (Courty et al. 2017) and (Yang, Wu, and Liu 2012) for more details.

Despite of the above studies in terms of applications, the

research on the algorithms is still rather limited and far from being satisfactory. Basically, we need to take into account of the high dimensionality and large number of points of the geometric patterns, simultaneously. In particular, as the developing of data acquisition techniques, data sizes increase fast and designing efficient algorithms for high dimensional alignment will be important for some applications. For example, due to the rapid progress of high-throughput sequencing technologies, biological data are growing exponentially (Yin et al. 2017). In fact, even for the 2D and 3D cases, solving the geometric alignment problem is quite challenging yet. For example, the well-known iterative closest point (ICP) method (Besl and McKay 1992) can only guarantee to obtain a local optimum. More previous works will be discussed in Section 1.2.

To the best of our knowledge, we are the first to consider developing efficient algorithmic framework for the geometric alignment problem in large-scale and high dimension. Our idea is inspired by the observation that many real-world datasets often manifest low intrinsic dimensions (Belkin 2003). For example, human handwriting images can be well embedded to some low dimensional manifold though the Euclidean dimension can be very high (Tenenbaum, De Silva, and Langford 2000). Following this observation, we consider to exploit the widely used notion, “doubling dimension” (Krauthgamer and Lee 2004; Talwar 2004; Karger and Ruhl 2002; Har-Peled and Mendel 2006; Dasgupta and Sinha 2013), to deal with the geometric alignment problem. Doubling dimension is particularly suitable to depict the data having low intrinsic dimension. We prove that the given geometric patterns with low doubling dimensions can be substantially compressed so as to save a large amount of running time when computing the alignment. More importantly, our compression approach is an independent step, and hence can serve as the preprocessing for various alignment methods.

The rest of the paper is organized as follows. We provide the definitions that are used throughout this paper in Section 1.1, and discuss some existing approaches and our main idea in Section 1.2. Then, we present our algorithm, analysis, and the time complexity in detail in Section 2 and 3. Finally, we study the practical performance of our proposed algorithm in Section 4.

1.1 Preliminaries

Before introducing the formal definition of geometric alignment, we need to define EMD and rigid transformation first.

Definition 1 (Earth Mover’s Distance (EMD)). *Let $A = \{a_1, a_2, \dots, a_{n_1}\}$ and $B = \{b_1, b_2, \dots, b_{n_2}\}$ be two sets of weighted points in \mathbb{R}^d with nonnegative weights α_i and β_j for each $a_i \in A$ and $b_j \in B$ respectively, and W_A and W_B be their respective total weights. The earth mover’s distance between A and B is $\mathcal{EMD}(A, B) =$*

$$\frac{1}{\min\{W_A, W_B\}} \min_F \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f_{ij} \|a_i - b_j\|^2, \quad (1)$$

where $F = \{f_{ij}\}$ is a feasible flow from A to B , i.e., each $f_{ij} \geq 0$, $\sum_{i=1}^{n_1} f_{ij} \leq \beta_j$, $\sum_{j=1}^{n_2} f_{ij} \leq \alpha_i$, and

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f_{ij} = \min\{W_A, W_B\}.$$

Intuitively, EMD can be viewed as the minimum transportation cost between A and B , where the weights of A and B are the “supplies” and “demands” respectively, and the cost of each edge connecting a pair of points from A to B is their “ground distance”. In general, the “ground distance” can be defined in various forms, and here we simply use the squared distance because it is widely adopted in practice.

Definition 2 (Rigid Transformation). *Let P be a set of points in \mathbb{R}^d . A rigid transformation \mathcal{T} on P is a transformation (i.e., rotation, translation, reflection, or their combination) which preserves the pairwise distances of the points in P .*

We consider rigid transformation for alignment, because it is very natural to interpret in real-world and has already been used by the aforementioned applications.

Definition 3 (Geometric Alignment). *Given two weighted point sets A and B as described in Definition 1, the problem of geometric alignment between A and B under rigid transformation is to determine a rigid transformation \mathcal{T} for B so as to minimize the earth mover’s distance $\mathcal{EMD}(A, \mathcal{T}(B))$.*

As previously mentioned, we consider to use doubling dimension to describe high dimensional data having low intrinsic dimension. We denote a metric space by (X, d_X) where d_X is the distance function of the set X . For instance, we can imagine that X is a set of points in a low dimensional manifold and d_X is simply Euclidean distance. For any $x \in X$ and $r \geq 0$, $Ball(x, r) = \{p \in X \mid d_X(x, p) \leq r\}$ indicates the ball of radius r around x (note that $Ball(x, r)$ is a subset of X).

Definition 4 (Doubling Dimension). *The doubling dimension of a metric space (X, d_X) is the smallest number ρ , such that for any $x \in X$ and $r \geq 0$, $Ball(x, 2r)$ is always covered by the union of at most 2^ρ balls with radius r .*

Doubling dimension describes the expansion rate of (X, d_X) ; intuitively, we can imagine a set of points uniformly distributed inside a ρ -dimensional hypercube, where its doubling dimension is $O(\rho)$ but the Euclidean dimension can be very high. For a more general case, a manifold in high dimensional Euclidean space may have a very low doubling dimension, as many examples studied in machine learning (Belkin 2003). Unfortunately, as shown before (Laakso 2002), such low doubling dimensional metrics cannot always be embedded to low dimensional Euclidean spaces with low distortion in terms of Euclidean distance. Therefore, we need to design the techniques being able to manipulate the data in high dimensional Euclidean space directly.

1.2 Existing Results and Our Approach

If building a bipartite graph, where the two columns of vertices correspond to the points of A and B respectively and each edge connecting (a_i, b_j) has the weight $\|a_i - b_j\|^2$, we can see that computing EMD actually is a min-cost flow problem. Many min-cost flow algorithms have been developed in the past decades, such as Network simplex algo-

rithm (Ahuja, Magnanti, and Orlin 1993), a specialized version of the simplex algorithm. Since EMD is an instance of min-cost flow problem in Euclidean space and the geometric techniques (e.g., the geometric sparse spanner) are applicable, a number of faster algorithms have been proposed in the area of computational geometry (Agarwal et al. 2017; Cabello et al. 2008; Indyk 2007), however, most of them only work for low dimensional case. Several EMD algorithms with assumptions (or some modifications on the objective function of EMD) have been studied in practical areas (Pele and Werman 2009; Ling and Okada 2007; Benamou et al. 2015).

Computing the geometric alignment of A and B is more challenging, since we need to determine the rigid transformation and EMD flow simultaneously. Moreover, due to the flexibility of rigid transformations, we cannot apply the EMD embedding techniques (Indyk and Thaper 2003; Andoni et al. 2009) to relieve the challenges. For example, the embedding can only preserve the EMD between A and B ; however, since there are infinite number of possible rigid transformations \mathcal{T} for B (note that we do not know \mathcal{T} in advance), it is difficult to also preserve the EMD between A and $\mathcal{T}(B)$. In theory, (Cabello et al. 2008) presented a $(2 + \epsilon)$ -approximation solution for the 2D case, and later (Klein and Veltkamp 2005) achieved an $O(2^{d-1})$ -approximation in \mathbb{R}^d ; (Ding and Xu 2017) proposed a PTAS for constant dimension. However, these theoretical algorithms cannot be efficiently implemented when the dimensionality is not constant. (Ding and Xu 2017) also mentioned that any constant factor approximation needs a time complexity at least $n^{\Omega(d)}$ based on some reasonable assumption in the theory of computational complexity, where $n = \max\{|A|, |B|\}$. That is, it is unlikely to obtain a $(1 + \epsilon)$ -approximation within a practical running time, especially when n is very large.

In practice, (Cohen and Guibas 1999) proposed an alternating minimization approach for computing the geometric alignment of A and B . Several other approaches (Cornea et al. 2005; Todorovic and Ahuja 2008) based on graph matching are inappropriate to be extended for high dimensional alignment. In machine learning, a related topic is called “manifold alignment” (Ham, Lee, and Saul 2005; Wang, Krafft, and Mahadevan 2011); however, it usually has different settings and applications, and thus is out of the scope of this paper.

Because the approach of (Cohen and Guibas 1999) is closely related to our proposed algorithm, we introduce it with more details for the sake of completeness. Roughly speaking, their approach is similar to the Iterative Closest Point method (ICP) method (Besl and McKay 1992), where its each iteration alternatively updates the EMD flow and rigid transformation. Thus it converges to some local optimum eventually. To update the rigid transformation, we can apply Orthogonal Procrustes (OP) analysis (Schönemann 1966). The original OP analysis is only for unweighted point sets, and hence we need some significant modification for our problem.

Suppose that the EMD flow $F = \{f_{ij}\}$ is fixed and the rigid transformation is waiting to update in the current stage.

We imagine two new sets of weighted points

$$\hat{A} = \{a_1^1, a_1^2, \dots, a_{n_1}^{n_2}; a_2^1, a_2^2, \dots, a_{n_2}^{n_2}; \dots; a_{n_1}^1, a_{n_1}^2, \dots, a_{n_1}^{n_2}\}; \quad (2)$$

$$\hat{B} = \{b_1^1, b_2^1, \dots, b_{n_2}^1; b_2^2, b_2^2, \dots, b_{n_2}^2; \dots; b_{n_1}^1, b_{n_1}^1, \dots, b_{n_2}^1\}, \quad (3)$$

where each a_i^j (resp., b_j^i) has the weight f_{ij} and the same spatial position of a_i (resp., b_j). With a slight abuse of notations, we also use a_i^j (resp., b_j^i) to denote the corresponding d -dimensional column vector in the following description. First, we take a translation vector \vec{v} such that the weighted mean points of \hat{A} and $\hat{B} + \vec{v}$ coincide with each other (this can be easily proved, due to the fact that the objective function uses squared distance (Cohen and Guibas 1999)). Second, by OP analysis, we compute an orthogonal matrix \mathcal{R} for $\hat{B} + \vec{v}$ to minimize its weighted L_2^2 difference to \hat{A} . For this purpose, we generate two $d \times (n_1 n_2)$ matrices M_A and M_B , where each point of \hat{A} (resp., $\hat{B} + \vec{v}$) corresponds to an individual column of M_A (resp., M_B); for example, a point $a_i^j \in \hat{A}$ (resp., $b_j^i + \vec{v} \in \hat{B} + \vec{v}$) corresponds to a column $\sqrt{f_{ij}} a_i^j$ (resp., $\sqrt{f_{ij}} (b_j^i + \vec{v})$) in M_A (resp., M_B). Let the SVD of $M_A \times M_B^T$ be $U \Sigma V^T$, and the optimal orthogonal matrix \mathcal{R} should be UV^T through OP analysis. Actually we do not need to really construct the large matrices M_A and M_B , since many of the columns are identical. Instead, we can compute the multiplication $M_A \times M_B^T$ in $O(n_1 n_2 d + \min\{n_1, n_2\} \cdot d^2)$ time (see Lemma 3 in Appendix). Therefore, the time complexity for obtaining the optimal \mathcal{R} is $O(n_1 n_2 d + \min\{n_1, n_2\} \cdot d^2 + d^3)$.

Proposition 1. *Each iteration of the approach of (Cohen and Guibas 1999) takes $\Gamma(n_1, n_2, d) + O(n_1 n_2 d + \min\{n_1, n_2\} \cdot d^2 + d^3)$ time, where $\Gamma(n_1, n_2, d)$ indicates the time complexity of the EMD algorithm it adopts. In practice, we usually assume $n_1, n_2 = O(n)$ with some $n \geq d$, and then the complexity can be simplified to be $\Gamma(n, d) + O(n^2 d)$.*

The bottleneck is that the algorithm needs to repeatedly compute the EMD and transformation, especially when n and d are large (usually $\Gamma(n, d) = \Omega(n^2 d)$). Based on the property of low doubling dimension, we construct a pair of compressed point sets to replace the original A and B , and run the same algorithm on the compressed data instead. As a consequence, the running time is reduced significantly. Note that our compression step is **independent of** the approach (Cohen and Guibas 1999); actually, any alignment method with the same objective function in Definition 3 can benefit from our compression idea. Recently, (Nasser, Jubran, and Feldman 2015) proposed a core-set based compression approach to speed up the computation of alignment. However, their method requires to know the correspondences between the point sets in advance and therefore it is not suitable to handle EMD; moreover, their compression achieves the advantage only when d is small.

2 The Algorithm and Analysis

Our idea starts from the widely studied k -center clustering. Given an integer $k \geq 1$ and a point set P in some metric

space, k -center clustering is to partition P into k clusters and cover each cluster by an individual ball, such that the maximum radius of the balls is minimized. (Gonzalez 1985) presented an elegant 2-approximation algorithm, where the radius of each resulting ball (i.e., cluster) is at most two times the optimum. Initially, it selects an arbitrary point, say c_1 , from the input P and sets $S = \{c_1\}$; then it iteratively selects a new point which has the largest distance to S among the points of P and adds it to S , until $|S| = k$ (the distance between a point q and S is defined as $\min\{\|q - p\| \mid p \in S\}$); suppose $S = \{c_1, \dots, c_k\}$, and then P is covered by the k balls $Ball(c_1, r), \dots, Ball(c_k, r)$ with $r \leq \min\{\|c_i - c_j\| \mid 1 \leq i \neq j \leq k\}$. It is able to prove that r is at most two times the optimal radius of the given instance. Using the property of doubling dimension, we have the following lemma.

Lemma 1. *Let P be a point set in \mathbb{R}^d with the doubling dimension $\rho \ll d$. The diameter of P is denoted by Δ , i.e., $\Delta = \max\{\|p - q\| \mid p, q \in P\}$. Given a small parameter $\epsilon > 0$, if one runs the k -center clustering algorithm of Gonzalez by setting $k = (\frac{2}{\epsilon})^\rho$, the radius of each resulting ball is at most $\epsilon\Delta$.*

Proof. Let S be the set of k points by Gonzalez's algorithm, and the resulting radius be r . We also define the aspect ratio of S as the ratio between the maximum and minimum pairwise distances in S . Then, it is easy to see that the aspect ratio of S is at most Δ/r . Now, we need the following Claim 1 from (Krauthgamer and Lee 2004; Talwar 2004). Actually, the claim can be obtained by recursively applying the definition of doubling dimension.

Claim 1. *Let (X, d_X) be a metric space with the doubling dimension ρ , and $Y \subset X$. If the aspect ratio of Y is upper bounded by some positive value α , then $|Y| \leq 2^{\rho \lceil \log_2 \alpha \rceil}$.*

Replacing X and Y by P and S respectively in the above claim, we have

$$|S| \leq 2^{\rho \lceil \log_2 \Delta/r \rceil} \leq 2^{\rho(1 + \log_2 \Delta/r)}. \quad (4)$$

Since $|S| = (2/\epsilon)^\rho$, (4) implies $r \leq \epsilon\Delta$. \square

Let A and B be the two given point sets in Definition 3, and the maximum of their diameters be Δ . We also assume that they both have the doubling dimension at most ρ . Our idea for compressing A and B is as follows. As described in Lemma 1, we set $k = (\frac{2}{\epsilon})^\rho$ and run Gonzalez's algorithm on A and B respectively. We denote by $S_A = \{c_1^A, \dots, c_k^A\}$ and $S_B = \{c_1^B, \dots, c_k^B\}$ the obtained sets of k -cluster centers. For each cluster center c_j^A (resp., c_j^B), we assign a weight that is equal to the total weights of the points in the corresponding cluster. As a consequence, we obtain a new instance (S_A, S_B) for geometric alignment. It is easy to know that the total weights of S_A (resp., S_B) is equal to W_A (resp., W_B). The following theorem shows that we can achieve an approximate solution for the instance (A, B) by solving the alignment of (S_A, S_B) .

Theorem 1. *Suppose $\epsilon > 0$ is a small parameter in Lemma 1. Given any $c \geq 1$, let \tilde{T} be*

a rigid transformation yielding c -approximation for minimizing $\mathcal{EMD}(S_A, \mathcal{T}(S_B))$ in Definition 3. Then, $\mathcal{EMD}(A, \tilde{T}(B))$

$$\begin{aligned} &\leq c(1 + 2\epsilon)^2 \cdot \min_{\mathcal{T}} \mathcal{EMD}(A, \mathcal{T}(B)) \\ &\quad + 2\epsilon(c + 1 + 2c\epsilon)(1 + 2\epsilon)\Delta^2 \\ &= c(1 + O(\epsilon)) \cdot \min_{\mathcal{T}} \mathcal{EMD}(A, \mathcal{T}(B)) \\ &\quad + 2\epsilon(1 + O(\epsilon))(c + 1)\Delta^2. \end{aligned} \quad (5)$$

Proof. First, we denote by \mathcal{T}_{opt} the optimal rigid transformation achieving $\min_{\mathcal{T}} \mathcal{EMD}(A, \mathcal{T}(B))$. Since \tilde{T} yields c -approximation for minimizing $\mathcal{EMD}(S_A, \mathcal{T}(S_B))$, we have $\mathcal{EMD}(S_A, \tilde{T}(S_B))$

$$\begin{aligned} &\leq c \cdot \min_{\mathcal{T}} \mathcal{EMD}(S_A, \mathcal{T}(S_B)) \\ &\leq c \cdot \mathcal{EMD}(S_A, \mathcal{T}_{opt}(S_B)). \end{aligned} \quad (6)$$

Recall that each point c_j^A (resp., c_j^B) has the weight equal to the total weights of the points in the corresponding cluster. For instance, if the cluster contains $\{a_{j(1)}, a_{j(2)}, \dots, a_{j(h)}\}$, the weight of c_j^A should be $\sum_{l=1}^h \alpha_{j(l)}$; actually, we can view c_j^A as h overlapping points $\{a'_{j(1)}, a'_{j(2)}, \dots, a'_{j(h)}\}$ with each $a'_{j(l)}$ having the weight $\alpha_{j(l)}$. Therefore, for the sake of convenience, we use another representation for S_A and S_B in our proof below:

$$S_A = \{a'_1, \dots, a'_{n_1}\} \text{ and } S_B = \{b'_1, \dots, b'_{n_2}\}, \quad (7)$$

where each a'_j (resp., b'_j) has the weight α_j (resp., β_j). Note that S_A and S_B only have k distinct positions respectively in the space. Moreover, due to Lemma 1, we know that $\|a'_i - a_i\|, \|b'_j - b_j\| \leq \epsilon\Delta$ for any $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$, and these bounds are invariant under any rigid transformation in the space. Consequently, for any pair (i, j) and rigid transformation \mathcal{T} , we have $\|a_i - \mathcal{T}(b_j)\|^2$

$$\begin{aligned} &\leq (\|a_i - a'_i\| + \|a'_i - \mathcal{T}(b'_j)\| \\ &\quad + \|\mathcal{T}(b'_j) - \mathcal{T}(b_j)\|)^2 \\ &\leq (\|a'_i - \mathcal{T}(b'_j)\| + 2\epsilon\Delta)^2 \\ &= \|a'_i - \mathcal{T}(b'_j)\|^2 + 4\epsilon\Delta\|a'_i - \mathcal{T}(b'_j)\| + 4\epsilon^2\Delta^2 \\ &\leq \|a'_i - \mathcal{T}(b'_j)\|^2 + 2\epsilon(\Delta^2 + \|a'_i - \mathcal{T}(b'_j)\|^2) \\ &\quad + 4\epsilon^2\Delta^2 \\ &= (1 + 2\epsilon)\|a'_i - \mathcal{T}(b'_j)\|^2 + (2\epsilon + 4\epsilon^2)\Delta^2 \end{aligned} \quad (8)$$

by applying triangle inequality. Using exactly the same idea, we have $\|a'_i - \mathcal{T}(b'_j)\|^2$

$$\leq (1 + 2\epsilon)\|a_i - \mathcal{T}(b_j)\|^2 + (2\epsilon + 4\epsilon^2)\Delta^2. \quad (9)$$

Based on Definition 1, we denote by $\tilde{F} = \{\tilde{f}_{ij}\}$ the induced flow of $\mathcal{EMD}(S_A, \tilde{T}(S_B))$ (using the representations (7) for S_A and S_B). Then (8) directly implies that

$$\mathcal{EMD}(A, \tilde{\mathcal{T}}(B))$$

$$\begin{aligned} &\leq \frac{1}{\min\{W_A, W_B\}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \tilde{f}_{ij} \|a_i - \tilde{\mathcal{T}}(b_j)\|^2 \\ &\leq \frac{1+2\epsilon}{\min\{W_A, W_B\}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \tilde{f}_{ij} \|a'_i - \tilde{\mathcal{T}}(b'_j)\|^2 \\ &\quad + (2\epsilon + 4\epsilon^2)\Delta^2 \\ &= (1+2\epsilon)\mathcal{EMD}(S_A, \tilde{\mathcal{T}}(S_B)) \\ &\quad + (2\epsilon + 4\epsilon^2)\Delta^2. \end{aligned} \quad (10)$$

By the similar idea (replacing $\tilde{\mathcal{T}}$ by \mathcal{T}_{opt} , and exchanging the roles of (A, B) and (S_A, S_B)), (9) directly implies that $\mathcal{EMD}(S_A, \mathcal{T}_{opt}(S_B))$

$$\begin{aligned} &\leq (1+2\epsilon)\mathcal{EMD}(A, \mathcal{T}_{opt}(B)) \\ &\quad + (2\epsilon + 4\epsilon^2)\Delta^2. \end{aligned} \quad (11)$$

Combining (6), (10), and (11), we have $\mathcal{EMD}(A, \tilde{\mathcal{T}}(B))$

$$\begin{aligned} &\leq (1+2\epsilon)\mathcal{EMD}(S_A, \tilde{\mathcal{T}}(S_B)) \\ &\quad + (2\epsilon + 4\epsilon^2)\Delta^2 \\ &\leq (1+2\epsilon) \cdot c \cdot \mathcal{EMD}(S_A, \mathcal{T}_{opt}(S_B)) \\ &\quad + (2\epsilon + 4\epsilon^2)\Delta^2 \\ &\leq c(1+2\epsilon)^2 \cdot \mathcal{EMD}(A, \mathcal{T}_{opt}(B)) \\ &\quad + 2\epsilon(c+1+2c\epsilon)(1+2\epsilon)\Delta^2, \end{aligned} \quad (12)$$

and the proof is completed. \square

When ϵ is small enough, Theorem 1 shows that $\mathcal{EMD}(A, \tilde{\mathcal{T}}(B)) \approx c \cdot \mathcal{EMD}(A, \mathcal{T}_{opt}(B))$. That is, $\tilde{\mathcal{T}}$, the solution of (S_A, S_B) , achieves roughly the same performance on (A, B) . Consequently, we propose the approximation algorithm for geometric alignment (see Algorithm 1). We would like to emphasize that though we use the algorithm from (Cohen and Guibas 1999) in Step 3, Theorem 1 is an independent result; that is, any alignment method with the same objective function in Definition 3 can benefit from Theorem 1.

Algorithm 1 Geometric Alignment

- 1: Given $\epsilon \in (0, 1)$ and set $k = (2/\epsilon)^\rho$.
 - 2: Run Gonzalez's k -center clustering algorithm on A and B , and obtain the sets of cluster centers S_A and S_B respectively.
 - 3: Apply the existing alignment algorithm, e.g., (Cohen and Guibas 1999), on (S_A, S_B) .
 - 4: Obtain the rigid transformation \mathcal{T} from Step 3, and compute the corresponding EMD flow between A and $\mathcal{T}(B)$.
 - 5: **Output:** A rigid transformation \mathcal{T} of B and the EMD flow between A and $\mathcal{T}(B)$.
-

3 The Time Complexity

We analyze the time complexity of Algorithm 1 and consider step 2-4 separately. To simplify our description, we use n to denote $\max\{n_1, n_2\}$. In step 3, we suppose that the iterative approach (Cohen and Guibas 1999) takes $\lambda \geq 1$ rounds.

Step 2. A straightforward implementation of Gonzalez's algorithm is selecting the k cluster centers iteratively where the resulting running time is $O(knd)$. Several faster implementations for the high dimensional case with low doubling dimension have been studied before; their idea is to maintain some data structures to reduce the amortized complexity of each iteration. We refer the reader to (Har-Peled and Mendel 2006) for more details.

Step 3. Since we run the algorithm (Cohen and Guibas 1999) on the smaller instance (S_A, S_B) instead of (A, B) , we know that the complexity of Step 3 is $O(\lambda(\Gamma(k, d) + k^2d + kd^2 + d^3))$ by Proposition 1.

Step 4. We need to compute the transformed $\mathcal{T}(B)$ first and then $\mathcal{EMD}(A, \mathcal{T}(B))$. Note that the transformation \mathcal{T} is not off-the-shelf, because it is the combination of a sequence of rigid transformations from the iterative approach (Cohen and Guibas 1999) in Step 3. Since it takes λ rounds, \mathcal{T} should be the multiplication of λ rigid transformations. We use $(\mathcal{R}_l, \vec{v}_l)$ to denote the orthogonal matrix and translation vector obtained in the l -th round for $1 \leq l \leq \lambda$. We can update B round by round: starting from $l = 1$, update B to be $\mathcal{R}_l B + \vec{v}_l$ in each round; the whole time complexity will be $O(\lambda nd^2)$. In fact, we have a more efficient way by computing \mathcal{T} first before transforming B .

Lemma 2. Let (\mathcal{R}, \vec{v}) be the orthogonal matrix and translation vector of \mathcal{T} . Then

$$\begin{aligned} \mathcal{R} &= \Pi_{i=1}^\lambda \mathcal{R}_i, \\ \vec{v} &= (\Pi_{i=2}^\lambda \mathcal{R}_i) \vec{v}_1 + (\Pi_{i=3}^\lambda \mathcal{R}_i) \vec{v}_2 + \\ &\quad \cdots + \mathcal{R}_\lambda \vec{v}_{\lambda-1} + \vec{v}_\lambda, \end{aligned} \quad (13)$$

and $\mathcal{T}(B)$ can be obtained in $O(\lambda d^3 + nd^2)$ time.

Proof. The equations (13) can be easily verified by simple calculations. In addition, we can recursively compute the multiplications $\Pi_{i=i}^\lambda \mathcal{R}_i$ for $i = \lambda, \lambda - 1, \dots, 1$. Consequently, the orthogonal matrix \mathcal{R} and translation vector \vec{v} can be obtained in $O(\lambda d^3)$ time. In addition, the complexity for computing $\mathcal{T}(B) = \mathcal{R}B + \vec{v}$ is $O(nd^2)$. \square

Lemma 2 provides a complexity significantly lower than the previous $O(\lambda nd^2)$ (usually n is much larger than d in practice). After obtaining $\mathcal{T}(B)$, we can compute $\mathcal{EMD}(A, \mathcal{T}(B))$ in $\Gamma(n, d)$ time. **Note** that the complexity $\Gamma(n, d)$ usually is $\Omega(n^2d)$, which dominates the complexity of Step 2 and the second term nd^2 in the complexity by Lemma 2. As a consequence, we have the following theorem for the running time.

Theorem 2. Suppose $n = \max\{n_1, n_2\} \geq d$ and the algorithm of (Cohen and Guibas 1999) takes $\lambda \geq 1$ rounds. The running time of Algorithm 1 is $O(\lambda(\Gamma(k, d) + k^2d + kd^2 + d^3)) + \Gamma(n, d)$.

Table 1: Random dataset: EMDs and running times of different compression levels.

γ	1/50	1/40	1/30	1/20	1/10	1
EMD	0.948	0.946	0.945	0.943	0.941	0.933
Time (s)	48.7	54.2	61.0	80.6	144.6	1418.2

Table 2: Linguistic datasets: EMDs and running times of different compression levels.

γ	1/50	1/40	1/30	1/20	1/10	1
sp-en EMD	0.989	0.969	0.955	0.931	0.892	0.820
Time (s)	3.4	3.6	3.7	3.9	5.3	100.5
it-en EMD	0.983	0.966	0.951	0.935	0.899	0.847
Time (s)	5.4	5.9	6.1	6.5	8.8	162.6
ja-ch EMD	0.991	0.975	0.962	0.941	0.910	0.836
Time (s)	1.6	2.1	2.2	2.0	3.0	67.0
tu-en EMD	0.982	0.960	0.946	0.922	0.889	0.839
Time (s)	10.1	10.4	11.2	11.9	15.9	287.0
ch-en EMD	1.014	1.012	0.990	0.962	0.923	0.842
Time (s)	1.9	1.7	2.2	2.4	2.7	51.6

If we run the same number of rounds on the original instance (A, B) by the approach (Cohen and Guibas 1999), the total running time will be $O(\lambda(\Gamma(n, d) + n^2d))$ by Proposition 1. When $k \ll n$, Algorithm 1 achieves a significant reduction on the running time.

4 Experiments

We implement our proposed algorithm and test the performance on both random and real datasets. All of the experimental results are obtained on a Windows workstation with 2.4GHz Intel Xeon CPU and 32GB DDR4 Memory. For each dataset, we run 20 trials and report the average results. We set the iterative approach (Cohen and Guibas 1999) to terminate when the change of the objective value is less than 10^{-3} .

To construct a random dataset, we randomly generate two manifolds in \mathbb{R}^{500} , which are represented by the polynomial functions with low dimension (≤ 50); in the manifolds, we take two sets of randomly sampled points having the sizes of $n_1 = 2 \times 10^4$ and $n_2 = 3 \times 10^4$, respectively; also, for each sampled point, we randomly assign a positive weight; finally, we obtain two weighted point sets as an instance of geometric alignment.

For real datasets, we consider the two applications mentioned in Section 1, unsupervised bilingual lexicon induction and PPI network alignment. For the first application, we have 5 pairs of languages: Chinese-English, Spanish-English, Italian-English, Japanese-Chinese, and Turkish-English. Given by (Zhang et al. 2017), each language has a vocabulary list containing 3000 to 13000 words; we also follow their preprocessing idea to represent all the words by vectors in \mathbb{R}^{50} through the embedding technique (Mikolov, Le, and Sutskever 2013). Actually, each vocabulary list is represented by a distribution in the space where each vector has the weight equal to the corresponding frequency in the language. For the second application, we use the popular benchmark dataset NAPAbench (Sayed Mohammad and

Yoon 2012) of PPI networks. It contains 3 pairs of PPI networks, where each network is a graph of 3000-10000 nodes. As the step of preprocessing, we apply the recent node2vec technique (Grover and Leskovec 2016) to represent each network by a group of vectors in \mathbb{R}^{100} ; following (Liu et al. 2017), we assign a unit weight to each vector.

Results. For each instance, we try different compression levels. According to Algorithm 1, we compress the size of each point set to be k . We set $k = \gamma \times \max\{n_1, n_2\}$ where $\gamma \in \{1/50, 1/40, 1/30, 1/20, 1/10, 1\}$; in particular, $\gamma = 1$ indicates that we directly run the algorithm (Cohen and Guibas 1999) without compression. The purpose of our proposed approach is to design a compression method, such that the resulting qualities on the original and compressed datasets are close (as stated in Section 1.2 and the paragraph after the proof of Theorem 1). So in the experiment, we focus on the comparison with the results on the original data (i.e., the results of $\gamma = 1$).

The results on random dataset are shown in Table 1. The obtained EMDs by compression are only slightly higher than the ones of $\gamma = 1$, while the advantage of the compression on running time is significant. For example, the running time of $\gamma = 1/50$ is less than 5% of the one of $\gamma = 1$. We obtain the similar performances on the real datasets. The results on Linguistic are shown in Table 2 (due to the space limit, we put the results on PPI network dataset in the full version of our paper); the EMD for each compression level is always at most 1.2 times the baseline with $\gamma = 1$, but the corresponding running time is dramatically reduced.

To further show the robustness of our method, we particularly add Gaussian noise to the random dataset and study the change of the objective value by varying the noise level. We set the standard variance of the Gaussian noise to be $\eta \times \Delta$, where Δ is the maximum diameter of the point sets and η is from 0.5×10^{-2} to 2.5×10^{-2} . Figure. 1 shows that the obtained EMD over baseline remains very stable (slightly higher than 1) of each noise level η .

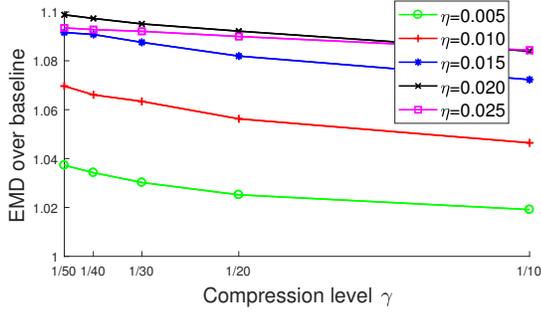


Figure 1: The EMDs over baseline for different noise levels.

5 Conclusion

In this paper, we propose a novel framework for compressing point sets in high dimension, so as to approximately preserve the quality of alignment. This work is motivated by several emerging applications in the fields of machine learning, bioinformatics, and wireless network. Our method utilizes the property of low doubling dimension, and yields a significant speedup on alignment. In the experiments on random and real datasets, we show that the proposed compression approach can efficiently reduce the running time to a great extent.

6 Acknowledgments

The research of this work was supported in part by NSF through grant CCF-1656905 and a start-up fund from Michigan State University. The authors also want to thank the anonymous reviewers for their helpful comments and suggestions for improving the paper.

7 Appendix

Lemma 3. *The multiplication $M_A \times M_B^T$ can be computed in $O(n_1 n_2 d + \min\{n_1, n_2\} \cdot d^2)$ time.*

Proof. With a slight abuse of notations, we also use F to denote the $n_1 \times n_2$ matrix of the EMD flow where each entry is f_{ij} ; also, $F_{i,:}$ represents the i -th row of the matrix F . Given a vector t , we use \sqrt{t} to denote the new vector with each entry being the square root of the corresponding one in t . Also, we use $\text{diag}(t)$ to denote the diagonal matrix where the i -th diagonal entry is the i -th entry of t . Following the constructions of M_A and M_B , we have

$$\begin{aligned}
 M_A &= [\sqrt{f_{11}}a_1^1, \dots, \sqrt{f_{1n_2}}a_1^{n_2}; \\
 &\quad \dots \dots \dots \\
 &\quad \sqrt{f_{n_11}}a_{n_1}^1, \dots, \sqrt{f_{n_1n_2}}a_{n_1}^{n_2}] \\
 &= [a_1\sqrt{F_{1,:}}, a_2\sqrt{F_{2,:}}, \dots, a_{n_1}\sqrt{F_{n_1,:}}]; \\
 M_B &= [\sqrt{f_{11}}b_1^1, \dots, \sqrt{f_{1n_2}}b_1^{n_2}; \\
 &\quad \dots \dots \dots \\
 &\quad \sqrt{f_{n_11}}b_{n_1}^1, \dots, \sqrt{f_{n_1n_2}}b_{n_1}^{n_2}] \\
 &= B[\text{diag}(\sqrt{F_{1,:}}), \text{diag}(\sqrt{F_{2,:}}), \\
 &\quad \dots, \text{diag}(\sqrt{F_{n_1,:}})],
 \end{aligned}$$

by some simple calculation. Then,

$$\begin{aligned}
 M_A \times M_B^T &= \sum_{i=1}^{n_1} (a_i \sqrt{F_{i,:}}) \times (\text{diag}(\sqrt{F_{i,:}}) B^T) \\
 &= \sum_{i=1}^{n_1} a_i F_{i,:} B^T = AFB^T.
 \end{aligned}$$

It is easy to know that computing AFB^T takes $O(n_1 n_2 d + \min\{n_1, n_2\} \cdot d^2)$ time. \square

References

- Agarwal, P. K.; Fox, K.; Panigrahi, D.; Varadarajan, K. R.; and Xiao, A. 2017. Faster algorithms for the geometric transportation problem. In *33rd International Symposium on Computational Geometry, SoCG 2017, July 4-7, 2017, Brisbane, Australia*, 7:1–7:16.
- Ahuja, R. K.; Magnanti, T. L.; and Orlin, J. B. 1993. *Network flows: theory, algorithms, and applications*. Prentice Hall.
- Andoni, A.; Do Ba, K.; Indyk, P.; and Woodruff, D. 2009. Efficient sketches for earth-mover distance, with applications. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, 324–330. IEEE.
- Belkin, M. 2003. *Problems of learning on manifolds*. The University of Chicago.
- Benamou, J.; Carlier, G.; Cuturi, M.; Nenna, L.; and Peyré, G. 2015. Iterative bregman projections for regularized transportation problems. *SIAM J. Scientific Computing* 37(2).
- Besl, P., and McKay, N. D. 1992. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14(2):239–256.
- Cabello, S.; Giannopoulos, P.; Knauer, C.; and Rote, G. 2008. Matching point sets with respect to the earth mover’s distance. *Computational Geometry* 39(2):118–133.
- Cao, X.; Wei, Y.; Wen, F.; and Sun, J. 2014. Face alignment by explicit shape regression. *International Journal of Computer Vision* 107(2):177–190.
- Cohen, S., and Guibas, L. 1999. The earth mover’s distance under transformation sets. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, 1.
- Cornea, N. D.; Demirci, M. F.; Silver, D.; Dickinson, S.; and Kantor, P. 2005. 3d object retrieval using many-to-many matching of curve skeletons. In *Shape Modeling and Applications, 2005 International Conference*, 366–371. IEEE.
- Courty, N.; Flamary, R.; Habrard, A.; and Rakotomamonjy, A. 2017. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, 3733–3742.
- Dasgupta, S., and Sinha, K. 2013. Randomized partition trees for exact nearest neighbor search. In *Conference on Learning Theory*, 317–337.
- Ding, H., and Xu, J. 2017. FPTAS for minimizing the earth mover’s distance under rigid transformations and related problems. *Algorithmica* 78(3):741–770.

- Gonzalez, T. F. 1985. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science* 38:293–306.
- Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. ACM.
- Ham, J.; Lee, D. D.; and Saul, L. K. 2005. Semisupervised alignment of manifolds. In *AISTATS*, 120–127.
- Har-Peled, S., and Mendel, M. 2006. Fast construction of nets in low-dimensional metrics and their applications. *SIAM Journal on Computing* 35(5):1148–1184.
- Indyk, P., and Thaper, N. 2003. Fast color image retrieval via embeddings. In *Workshop on Statistical and Computational Theories of Vision (at ICCV)*.
- Indyk, P. 2007. A near linear time constant factor approximation for euclidean bichromatic matching (cost). In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 39–42. Society for Industrial and Applied Mathematics.
- Karger, D. R., and Ruhl, M. 2002. Finding nearest neighbors in growth-restricted metrics. In *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, 741–750. ACM.
- Klein, O., and Veltkamp, R. C. 2005. Approximation algorithms for computing the earth mover’s distance under transformations. In *International Symposium on Algorithms and Computation*, 1019–1028. Springer.
- Krauthgamer, R., and Lee, J. R. 2004. Navigating nets: simple algorithms for proximity search. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, 798–807. Society for Industrial and Applied Mathematics.
- Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, 957–966.
- Laakso, T. J. 2002. Plane with a_∞ -weighted metric not bilipschitz embeddable to r^n . *Bulletin of the London Mathematical Society* 34(6):667–676.
- Ling, H., and Okada, K. 2007. An efficient earth mover’s distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence* 29(5):840–853.
- Liu, Y.; Ding, H.; Chen, D.; and Xu, J. 2017. Novel geometric approach for global alignment of PPI networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 31–37.
- Malod-Dognin, N.; Ban, K.; and Pržulj, N. 2017. Unified alignment of protein-protein interaction networks. *Scientific Reports* 7(1):953.
- Maltoni, D.; Maio, D.; Jain, A. K.; and Prabhakar, S. 2009. *Handbook of fingerprint recognition*. Springer Science & Business Media.
- Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- Nasser, S.; Jubran, I.; and Feldman, D. 2015. Low-cost and faster tracking systems using core-sets for pose-estimation. *CoRR* abs/1511.09120.
- Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10):1345–1359.
- Pele, O., and Werman, M. 2009. Fast and robust earth mover’s distances. In *Computer vision, 2009 IEEE 12th international conference on*, 460–467. IEEE.
- Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover’s distance as a metric for image retrieval. *International journal of computer vision* 40(2):99–121.
- Sayed Mohammad, E. S., and Yoon, B.-J. 2012. A network synthesis model for generating protein interaction network families. *PLoS one* 7.
- Schönemann, P. H. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31(1):1–10.
- Talwar, K. 2004. Bypassing the embedding: algorithms for low dimensional metrics. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, 281–290.
- Tenenbaum, J. B.; De Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290(5500):2319–2323.
- Todorovic, S., and Ahuja, N. 2008. Region-based hierarchical image matching. *International Journal of Computer Vision* 78(1):47–66.
- Wang, C.; Krafft, P.; and Mahadevan, S. 2011. Manifold alignment.
- Yang, Z.; Wu, C.; and Liu, Y. 2012. Locating in fingerprint space: wireless indoor localization with little human intervention. In *Proceedings of the 18th annual international conference on Mobile computing and networking*, 269–280. ACM.
- Yin, Z.; Lan, H.; Tan, G.; Lu, M.; Vasilakos, A. V.; and Liu, W. 2017. Computing platforms for big biological data analytics: perspectives and challenges. *Computational and structural biotechnology journal* 15:403–411.
- Youn, H.; Sutton, L.; Smith, E.; Moore, C.; Wilkins, J. F.; Maddieson, I.; Croft, W.; and Bhattacharya, T. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences* 113(7):1766–1771.
- Zhang, M.; Liu, Y.; Luan, H.; and Sun, M. 2017. Earth mover’s distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 1934–1945.