Algorithms for Metric Learning via Contrastive Embeddings

Diego Ihara
University of Illinois at Chicago, USA
https://dihara2.people.uic.edu/
dihara2@uic.edu

Neshat Mohammadi
University of Illinois at Chicago, USA
nmoham24@uic.edu

Anastasios Sidiropoulos
University of Illinois at Chicago, USA
http://sidiropoulos.org
sidiropo@uic.edu

Abstract

We study the problem of supervised learning a metric space under *discriminative* constraints. Given a universe X and sets $\mathcal{S}, \mathcal{D} \subset \binom{X}{2}$ of *similar* and *dissimilar* pairs, we seek to find a mapping $f : X \to Y$, into some target metric space $M = (Y, \rho)$, such that similar objects are mapped to points at distance at most u, and dissimilar objects are mapped to points at distance at least ℓ. More generally, the goal is to find a mapping of maximum *accuracy* (that is, fraction of correctly classified pairs). We propose approximation algorithms for various versions of this problem, for the cases of Euclidean and tree metric spaces. For both of these target spaces, we obtain fully polynomial-time approximation schemes (FPTAS) for the case of perfect information. In the presence of imperfect information we present approximation algorithms that run in quasi-polynomial time (QPTAS). We also present an exact algorithm for learning line metric spaces with perfect information in polynomial time. Our algorithms use a combination of tools from metric embeddings and graph partitioning, that could be of independent interest.

2012 ACM Subject Classification Theory of computation → Computational geometry

Keywords and phrases metric learning, contrastive distortion, embeddings, algorithms

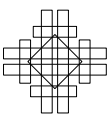
Digital Object Identifier 10.4230/LIPIcs.SoCG.2019.45

Related Version A full version of this paper is available at https://arxiv.org/abs/1807.04881.

Funding Supported by NSF under award CAREER 1453472, and grants CCF 1815145 and CCF 1423230.

1 Introduction

Geometric algorithms have given rise to a plethora of tools for data analysis, such as clustering, dimensionality reduction, nearest-neighbor search, and so on; we refer the reader to [20, 19, 25] for an exposition. A common aspect of these methods is that the underlying data is interpreted as a metric space. That is, each object in the input is treated as a point, and the pairwise dissimilarity between pairs of objects is encoded by a distance function on pairs of points. An important element that can critically determine the success of this data analytic framework, is the choice of the actual metric. Broadly speaking, the area of *metric learning* is concerned with methods for recovering an underlying metric space that agrees with a given set of observations (we refer the reader to [29, 23] for a detailed exposition). The problem of learning the distance function is cast as an optimization problem, where the objective function quantifies the extend to which the solution satisfies the input constraints.



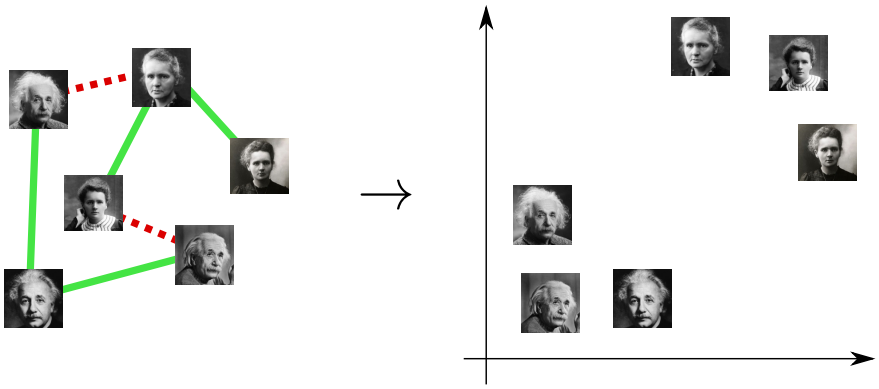



Figure 1 An illustration of the metric learning framework. The input consists of a set of objects (here, a set of photos), with some pairs labeled as "similar" (depicted in green), and some pairs labeled "dissimilar" (depicted in red). The output is an embedding into some metric space (here, the Euclidean plane), such that similar objects are mapped to nearby points, while dissimilar objects are mapped to points that are far from each other.

Supervised vs. unsupervised metric learning. The problems studied in the context of metric learning generally fall within two main categories: *supervised* and *unsupervised* learning. In the case of unsupervised metric learning, the goal is to discover the intrinsic geometry of the data, or to fit the input into some metric space with additional structure. A prototypical example of an unsupervised metric learning problem is dimensionality reduction, where one is given a high-dimensional point set and the goal is to find a mapping into some space of lower dimension, or with a special structure, that approximately preserves the geometry of the input (see e.g. [15, 26, 12]).

In contrast, the input in a supervised metric learning problem includes *label constrains*, which encode some prior knowledge about the ground truth. As an example, consider the case of a data set where experts have labeled some pairs of points as being "similar" and some pairs as being "dissimilar". Then, a metric learning task is to find a transformation of the input such that similar points end up close together, while dissimilar points end up far away from each other. As an illustrative example of the above definition, consider the problem of recognizing a face from a photo. More concretely, a typical input consists of some *universe* X, which is a set of photos of faces, together with some $\mathcal{S}, \mathcal{D} \subseteq \binom{X}{2}$. The set $\mathcal{S}$ consists of pairs of photos that correspond to the same person, while $\mathcal{D}$ consists of pairs of photos from different people. A typical approach for addressing this problem (see e.g. [14]) is to find a mapping $f : X \to \mathbb{R}^d$, for some $d \in \mathbb{N}$, such that for all similar pairs $\{x, y\} \in \mathcal{S}$ we have $\|f(x) - f(y)\|_2 \leq u$, and for all dissimilar pairs $\{x, y\} \in \mathcal{D}$ we have $\|f(x) - f(y)\|_2 \geq \ell$, for some $u, \ell > 0$. More generally, one seeks to find a mapping f that maximizes the fraction of correctly classified pairs of photos (see Figure 1).

1.1 Problem formulation

At the high level, an instance of a metric learning problem consists of some universe of objects X, together with some similarity information on subsets of these objects. Here, we focus on similarity and dissimilarity constraints. Specifically, an instance is a tuple $\phi = (X, \mathcal{S}, \mathcal{D}, u, \ell)$, where X is a finite set, with $|X| = n$, $\mathcal{S}, \mathcal{D} \subset \binom{X}{2}$, which are sets of pairs of objects that are labeled as "similar" and "dissimilar" respectively, and $u, \ell > 0$. We refer to the elements of $\mathcal{S} \cup \mathcal{D}$ as *constraints*. We focus on the case where $\mathcal{S} \cap \mathcal{D} = \emptyset$, and $\mathcal{S} \cup \mathcal{D} = \binom{X}{2}$. Let

$f : X \to Y$ be a mapping into some target metric space (Y, ρ). As it is typical with geometric realization problems, we relax the definition to allow for a small multiplicative error $c \geq 1$ in the embedding. We say that f *satisfies* $\{x, y\} \in \mathcal{S}$, if

$$\rho(f(x), f(y)) \leq u \cdot c, \tag{1}$$

and we say that it satisfies $\{x, y\} \in \mathcal{D}$ if

$$\rho(f(x), f(y)) \geq \ell/c. \tag{2}$$

If f does not satisfy some $\{x, y\} \in \mathcal{S} \cup \mathcal{D}$, then we say that it *violates* it. We refer to the parameter c as the *contrastive distortion* of f. We also refer to f as a *constrastive embedding*, or *c-embedding*.

1.2 Our contribution

We now briefly discuss our main contributions.

We focus on the problem of computing an embedding with low contrastive distortion, and with maximum *accuracy*, which is defined to be the fraction of satisfied constraints. Since we are assuming *complete information*, that is $\mathcal{S} \cup \mathcal{D} = \binom{X}{2}$, it follows that the accuracy is equal to $k/\binom{n}{2}$, where k is the number of satisfied constraints. We remark that the setting of complete information requires a *dense* set of constraints. This case is important in applications when learning a metric space based on a set of objects with fully-labeled pairs. This is also the setting of other important machine learning primitives, such as Correlation Clustering (see [7]), which we discuss in Section 1.4.

Our results are concerned with two main cases: *perfect* and *imperfect* information. Here, the case of perfect information corresponds to the promise problem where there exists an embedding that satisfies all constraints. On the other hand, in the case of imperfect information, no such promise is given. As we shall see, the latter scenario appears to be significantly more challenging. Our results are concerned with two different families of target spaces: d-dimensional Euclidean space, and trees.

Learning Euclidean metric spaces. We begin our investigation by observing that the problem of computing a contrastive embedding into d-dimensional Euclidean space with perfect information is polynomial-time solvable for $d = 1$, and it becomes NP-hard even for $d = 2$. The result for $d = 1$ is obtained via a simple greedy algorithm, while the NP-hardness proof uses a standard reduction from the problem of recognizing unit disk graphs (see [11]).

▶ **Theorem 1** (Learning the line with perfect information). *Let $\gamma = (X, \mathcal{S}, \mathcal{D}, u, \ell)$ be an instance of the metric learning problem. Then there exists a polynomial-time algorithm which given γ, either computes an 1-embedding $\widehat{f} : X \to \mathbb{R}$, with accuracy 1, or correctly decides that no such embedding exists.*

▶ **Theorem 2** (Hardness of learning the plane with perfect information). *Given an instance $\gamma = (X, \mathcal{S}, \mathcal{D}, u, \ell)$ of the metric learning problem, it is NP-hard to decide whether there exists an 1-embedding $\widehat{f} : X \to \mathbb{R}^2$, with accuracy 1.*

The above two results indicate that, except for what is essentially its simplest possible case, the problem is generally intractable. This motivates the study of approximation algorithms. Our first main result in this direction is a FPTAS for the case of perfect information, summarized in the following.

▶ **Theorem 3** (Learning Euclidean metric spaces with perfect information). *Let $u, \ell > 0$ be fixed constants. Let $\gamma = (X, \mathcal{S}, \mathcal{D}, u, \ell)$ be an instance of the problem of learning a d-dimensional Euclidean metric space with perfect information, with $|X| = n$, for some $d \geq 2$. Suppose that γ admits an 1-embedding $f^* : X \to \mathbb{R}^d$ with accuracy 1. Then for any $\varepsilon, \varepsilon' > 0$, there exists a randomized algorithm which given γ, ε, and ε', computes a $(1+\varepsilon')$-embedding $\widehat{f} : X \to \mathbb{R}^d$, with accuracy at least $1 - \varepsilon$, in time $n^{O(1)} g(\varepsilon, \varepsilon', d) = n^{O(1)} 2^{(\frac{d}{\varepsilon\varepsilon'})^{O(d^2)}}$, with high probability. In particular, for any fixed d, ε, and ε', the running time is polynomial.*

We also obtain the following QPTAS for the case of imperfect information.

▶ **Theorem 4** (Learning Euclidean metric spaces with imperfect information). *Let $d \geq 1$. Let $u, \ell > 0$ be fixed constants. Let $\gamma = (X, \mathcal{S}, \mathcal{D}, u, \ell)$ be an instance of the problem of learning a d-dimensional Euclidean metric space, with $|X| = n$, for some $d \geq 1$. Suppose that γ admits an 1-embedding $f^* : X \to \mathbb{R}^d$ with accuracy $1-\zeta$, for some $\zeta > 0$. Then for any $\varepsilon, \varepsilon' > 0$, there exists an algorithm which given γ, ε, ε', and ζ, computes a $(1+\varepsilon')$-embedding $\widehat{f} : X \to \mathbb{R}^d$, with accuracy at least $1 - O(\zeta^{1/2} \log^{3/4} n (\log\log n)^{1/2}) - \varepsilon$, in time $n^{O(1)} 2^{\varepsilon^{-2}(\frac{d \log n}{\zeta\varepsilon'})^{O(d)}}$. In particular, for any fixed d, ε, ε', and ζ, the running time is quasi-polynomial.*

Learning tree metric spaces. We next consider the case of embedding into tree metric spaces. More precisely, we are given an instance $(X, \mathcal{S}, \mathcal{D}, u, \ell)$ of the metric learning problem, and we wish to find a tree $\widehat{T}$, and an embedding $\widehat{f} : X \to \widehat{T}$, with maximum accuracy. Note that the tree $\widehat{T}$ is not part of the input. Our results closely resemble the ones from the case of learning Euclidean metric spaces. Specifically, for the case of perfect information we obtain a PTAS, summarized in the following.

▶ **Theorem 5** (Learning tree metric spaces with perfect information). *Let $u, \ell > 0$ be fixed constants. Let $\gamma = (X, \mathcal{S}, \mathcal{D}, u, \ell)$ be an instance of the problem of learning a tree metric space with perfect information, with $|X| = n$. Suppose that γ admits an 1-embedding $f^* : X \to V(T^*)$, for some tree T^*, with accuracy 1. Then for any $\varepsilon, \varepsilon' > 0$, there exists a randomized algorithm which given γ, ε, and ε', computes some tree $\widehat{T}$, and a $(1+\varepsilon')$-embedding $\widehat{f} : X \to V(\widehat{T})$, with accuracy at least $1 - \varepsilon$, in time $n^{O(1)} g(\varepsilon, \varepsilon') = n^{O(1)} 2^{1/(\varepsilon\varepsilon')^{O(1/\varepsilon)}}$, with high probability. In particular, for any fixed ε and ε', the running time is polynomial.*

As in the case of learning Euclidean metric spaces, we also obtain a QPTAS for learning tree metric spaces in the presence of imperfect information, summarized in the following.

▶ **Theorem 6** (Learning tree metric spaces with imperfect information). *Let $u, \ell > 0$ be fixed constants. Let $\gamma = (X, \mathcal{S}, \mathcal{D}, u, \ell)$ be an instance of the problem of learning a tree metric space with imperfect information, with $|X| = n$. Suppose that γ admits an 1-embedding $f^* : X \to V(T^*)$, for some tree T^*, with accuracy $1-\zeta$, for some $\zeta > 0$. Then for any $\varepsilon, \varepsilon' > 0$, there exists a randomized algorithm which given γ, ε, ε', and ζ, computes some tree $\widehat{T}$, and a $(1+\varepsilon')$-embedding $\widehat{f} : X \to V(\widehat{T})$, with accuracy at least $1 - O(\zeta^{1/2} \log^{3/4} n (\log\log n)^{1/2}) - \varepsilon$, in time $2^{((\log n)/(\zeta^{1/2}\varepsilon\varepsilon'))^{O(1/\varepsilon)}}$. In particular, for any fixed ε, ε', and ζ, the running time is quasi-polynomial.*

1.3 Overview of our techniques

We now give a brief overview of the techniques used in obtaining our approximation algorithms for the metric learning problem. Interestingly, our algorithms for the Euclidean case closely resemble our algorithms for tree metric spaces.

Learning Euclidean metric spaces. At the high level, our algorithms for learning Euclidean metric spaces consist of the following steps:

Step 1: Partitioning. We partition the instance into sub-instances, each admitting an embedding into a ball of small diameter. For the case of perfect information, it is easy to show that such a partition exists, using a Lipschitz partition of $\mathbb{R}^d$ (see [13]). Such a partition can be chosen so that only a small fraction of similarity constraints are "cut" (that is, their endpoints fall in different clusters of the partition). However, computing such a partition is a difficult task, since we do not have an optimal embedding (which is what we seek to compute). We overcome this obstacle by using a result of Krauthgamer and Roughgarden [22], which allows us to compute such a partition, without access to an optimal embedding.

For the case of imperfect information, the situation is somewhat harder since there might be no embedding that satisfies all the constraints. This implies that the partition that we use in the case of perfect information, is not guaranteed to exist anymore. Therefore, instead of using a geometric partitioning procedure, we resort to graph-theoretic methods. We consider the graph $G_{\mathcal{S}}$, with $V(G) = X$, and with $E(G) = \mathcal{S}$. Roughly speaking, we partition $G_{\mathcal{S}}$ into expanding subgraphs. This can be done by deleting only a relatively small fraction of edges. For each expanding subgraph, we can show that the corresponding sub-instance admits an embedding into a subspace of small diameter.

Step 2: Embedding each sub-instance. Once we have partitioned the instance into sub-instances, with each one admitting an embedding into a ball of small diameter, it remains to find such an embedding. This is done by first discretizing the target ball, and then using the theory of pseudoregular partitions (see e.g. [17]) to exhaustively find a good embedding. The discretization step introduces contrastive distortion $1 + \varepsilon'$, for some $\varepsilon' > 0$ that can be made arbitrarily small in expense of the running time.

Step 3: Combining the embeddings. Finally, we need to combine the embeddings of the sub-instances to an embedding of the original instance. This can easily be done by ensuring that the images of different clusters are sufficiently far from each other. Since only a small fraction of similarity constraints are cut by the original partition, it follows that the resulting accuracy is high.

Learning tree metric spaces. Surprisingly, the above template is also used, almost verbatim, for the case of learning tree metric spaces. The only difference is that the partitioning step now resembles the random partitioning scheme of Klein, Plotkin and Rao [21] for minor-free graphs. The embedding step requires some modification, since the space of trees of bounded diameter is infinite, and thus exhaustive enumeration is not directly possible. We resolve this issue by first showing that if there exists an embedding into a tree of small diameter, then there also exists an embedding into a *single* tree, which we refer to as *canonical*, and with only slightly worse accuracy.

1.4 Related work

Metric embeddings. The theory of metric embeddings is the source of several successful methods for processing metrical data sets, which have found applications in metric learning (see, e.g. [29, 23]). Despite the common aspects of the use of metric embeddings in metric learning and in algorithm design, there are some key differences. Perhaps the most important difference is that the metric embedding methods being used in algorithm design often correspond to unsupervised metric learning tasks. Consequently, problems and methods that are studied in the context of metric learning, have not received much attention from the

theoretical computer science community. Many of the existing methods used in supervised metric learning tasks are based on convex optimization, and are thus limited to specific kinds of objective functions and constraints (see e.g. [30, 16]).

The problems considered in this paper are closely related to questions studied in the context of computing low-distortion embeddings. For example, the problem of learning a tree metric space is closely related to the problem of embedding into a tree, which has been studied under various classical notions of distortion (see e.g. [12, 9, 28, 2, 1]). Similarly, several algorithms have been proposed for the problem of computing low-distortion embeddings into Euclidean space (see e.g. [24, 3, 26, 27, 5]). Various extensions and generalizations have also been considered for maps that approximately preserve the relative ordering of distances (see e.g. [10, 2, 6]). However, we remark that all of the above results correspond to the unsupervised version of the metric learning problem, and are thus not directly applicable in the supervised setting considered here.

Other types of supervision and embeddings. We note that other notions of supervision have also been considered in the metric learning literature. For example, instead of pairs of objects that are labeled either "similar" or "dissimilar", another possibility is to have triple constraints of the form $(x, y, z) \in \binom{X}{3}$, encoding the fact that "x is more similar to y than z". In this case, one seeks to find a mapping $f : X \to Y$, such that $\rho(f(x), f(y)) < \rho(f(x), f(z)) - m$, for some *margin* parameter $m > 0$ (see e.g. [30]). However, the form of supervision that we consider is one of the most popular ones in practice. Furthermore, it is often desirable that the mapping f has a special form, such as linearity (when X is a subset of some linear space), or being specified implicitly via a set of parameters, as is the case of mappings computed by neural networks. We believe that our work can lead to further theoretical understanding of the metric learning problem under different types of supervision, and for specific classes of embeddings.

Correlation Clustering. The supervised metric learning problem considered here can be thought of as a generalization of the classical Correlation Clustering problem [7]. More specifically, the Correlation Clustering problem is precisely the metric learning problem with finite contrastive distortion, for the special case when the host is the uniform metric space, with $u = 0$ and $\ell = 1$.

1.5 Organization

The rest of the paper is organized as follows. Section 2 introduces some notation and definitions. Section 3 shows how pseudoregular partitions can be used to compute near-optimal embeddings into spaces of bounded cardinality. Section 4 presents the algorithm for learning Euclidean spaces with perfect information. The algorithm for learning Euclidean spaces with imperfect information is given in Section 5.

The algorithms for learning tree metric spaces under perfect and imperfect information, the exact algorithm for learning line metric spaces with perfect information, and the NP-hardness proof of learning the plane with perfect information can be found in the full version of this paper.

2 Preliminaries

For any non-negative integer n, we use the notation $[n] = \{1, \ldots, n\}$, with the convention $[0] = \emptyset$. Let $M = (X, \rho)$ be some metric space. We say that M is a *line* metric space if it can be realized as a submetric of $\mathbb{R}$, endowed with the standard distance. For a graph G and some $v \in V(G)$, we write

$$N_G(v) = \{u \in V(G) : \{v, u\} \in E(G)\}.$$

For some $U \subset V(G)$, we denote by $E(U)$ the set of edges in G that have both endpoints in U; that is $E(U) = \{\{u, v\} \in E(G) : \{u, v\} \subseteq U\}$. For some $U, U' \subset V(G)$, we also write $E(U, U') = \{\{u, v\} \in E(G) : u \in U, v \in U'\}$.

3 Pseudoregular partitions and spaces of bounded cardinality

In this Section we present an algorithm for computing a nearly-optimal embedding, provided that there exists a solution contained inside a ball of small cardinality.

Let G be a n-vertex graph. For any disjoint $A, B \subseteq V(G)$, let $e(A, B) = |E(A, B)|$, where $E(A, B)$ denotes the set of edged with one endpoint in A and one in B, and let $\mathbf{d}(A, B) = \frac{e(A,B)}{|A| \cdot |B|}$. Let $\mathcal{V} = V_1, \ldots, V_k$ be a partition of $V(G)$. For any $i, j \in [k]$, let $\mathbf{d}_{i,j} = \mathbf{d}(V_i, V_j)$. For any $U \subseteq V(G)$, let $U_i = U \cap V_i$. The partition $\mathcal{V}$ is called *ε-pseudoregular* if for all disjoint $S, T \subseteq V(G)$, we have $\left| e(S, T) - \sum_{i \in [k]} \sum_{j \in [k]} \mathbf{d}_{i,j} |S \cap V_i||T \cap V_j| \right| \leq \varepsilon n^2$. It is also called *equitable* if for all $i, j \in [k]$, we have $||V_i| - |V_j|| \leq 1$. We recall the following result due to Frieze and Kannan [18] on computing pseudoregular partitions (see also [17, 8]).

▶ **Theorem 7** ([18]). *There exists a randomized algorithm which given an n-vertex graph G, and $\varepsilon, \delta > 0$, computes an equitable ε-pseudoregular partition of G with at most $k = 2^{O(\varepsilon^{-2})}$ parts, in time $2^{O(\varepsilon^{-2})} n^2 / (\varepsilon^2 \delta^3)$, with probability at least $1 - \delta$.*

The following is the main result of this Section.

▶ **Lemma 8** (Pseudoregular partitions and embeddings into spaces of small cardinality). *Let $\varepsilon, \varepsilon' > 0$. Let $\gamma_C = (C, \mathcal{S}, \mathcal{D}, u, \ell)$ be an instance of the metric learning problem, with $|C| = n$. Let $M = (N, \rho)$ be some metric space. Suppose that γ_C admits a $(1 + \varepsilon')$-embedding $g^* : C \to N$ with accuracy r^*. Then, there exists an algorithm which given γ_C and M, computes a $(1 + \varepsilon')$-embedding $\widehat{g} : C \to N$, with accuracy at least $r^* - \varepsilon$. The running time is $n^{O(1)} 2^{O(|N|^5/\varepsilon^2)}$.*

Due to lack of space, the proof of Lemma 8 is available in the full version of this paper.

4 Learning Euclidean metric spaces with perfect information

It this section we describe our algorithm for learning d-dimensional Euclidean metric spaces with perfect information. First, we present some tools from the theory of random metric partitions, and show how they can be used to partition an instance to smaller ones, each corresponding to a subspace of bounded diameter. The final algorithm combines this decomposition with the algorithm from Section 3 for learning metric spaces of bounded diameter.

4.1 Euclidean spaces and Lipschitz partitions

We introduce the necessary tools for randomly partitioning a metric space, and use them to obtain a decomposition of the input into pieces, each admitting an embedding into a ball in $\mathbb{R}^d$ of small diameter.

Let $M = (X, \rho)$ be a metric space. A partition P of X is called Δ-*bounded*, for some $\Delta > 0$, if every cluster in P has diameter at most Δ. Let $\mathcal{P}$ be a probability distribution over partitions of X. We say that $\mathcal{P}$ is (β, Δ)-*Lipschitz*, for some $\beta, \Delta > 0$, if the following conditions are satisfied:

- Any partition in the support of $\mathcal{P}$, is Δ-bounded.
- For all $p, q \in X$, $\Pr_{P\sim\mathcal{P}}[P(x) \neq P(y)] \leq \beta \cdot \rho(x, y)$.

▶ **Lemma 9** ([13]). *For any $d \geq 1$, and any $\Delta > 0$, we have that d-dimensional Euclidean space admits a $(O(\sqrt{d}/\Delta), \Delta)$-Lipschitz distribution.*

Let $G_{\mathcal{S}}$ be the graph with vertex set $V(G_{\mathcal{S}}) = X$, and edge set $E(G_{\mathcal{S}}) = \mathcal{S}$. We use $\rho_{\mathcal{S}}$ to denote the shortest-path distance of $G_{\mathcal{S}}$, where the length of each edge is set to u.

▶ **Lemma 10.** *Let $X' \subset X$, such that the subgraph of $G_{\mathcal{S}}$ induced on X' (i.e. $G_{\mathcal{S}}[X']$) is connected. Suppose further that $\mathsf{diam}(f^*(X')) \leq \Delta$, for some $\Delta > 0$. Then, for any $p, q \in X'$, we have $\rho_{\mathcal{S}}(p, q) \leq 2u(1 + 4\Delta/u)^d$.*

Proof. Let $G' = G_{\mathcal{S}}[X']$. Let $Q = x_1, \ldots, x_L$ be the shortest path between p and q in G', with $x_1 = p$, $x_L = q$. We first claim that for all $i, j \in \{1, \ldots, L\}$, with $i < j - 1$, we have that

$$\|f^*(x_i), f^*(x_j)\|_2 > u. \tag{3}$$

Indeed, note that if $\|f^*(x_i), f^*(x_j)\|_2 \leq u$, then since γ has full information, i.e. $\mathcal{S} \cup \mathcal{D} = \binom{X}{2}$, it must be that $\{x_i, x_j\} \in \mathcal{S}$, and thus the edge $\{x_i, x_j\}$ is present in the graph G'. This implies that $\rho_{G'}(x_i, x_j) = u$, which contradicts the fact that Q is a shortest path. We have thus established (3).

For each $i \in \{1, \ldots, L\}$, let B_i be the ball in $\mathbb{R}^d$ centered at $f^*(x_i)$ and of radius $u/2$. By (3) we get that the balls $B_2, B_4, \ldots, B_{2\lfloor L/2 \rfloor}$ are pairwise disjoint. Since $\mathsf{diam}(f^*(X')) \leq \Delta$, it follows that there exists some $x^* \in \mathbb{R}^d$, such that $f^*(X') \subset \mathsf{ball}(x^*, 2\Delta)$. Therefore

$$\bigcup_{i=1}^{\lfloor L/2 \rfloor} B_{2i} \subseteq \bigcup_{i=1}^{\lfloor L/2 \rfloor} \mathsf{ball}(x_{2i}, u/2) \subseteq \mathsf{ball}(x^*, 2\Delta + u/2). \tag{4}$$

Recall that the volume of the ball of radius r in $\mathbb{R}^d$ is equal to $V_d(r) = \frac{\pi^{d/2}}{\Gamma(1+d/2)} r^d$. From (4) we get $\lfloor L/2 \rfloor \cdot V_d(u/2) \leq V_d(2\Delta + u/2)$, and thus $\rho_{\mathcal{S}}(p, q) \leq \rho_{G'}(p, q) = u \cdot (L - 1) \leq 2u \frac{V_d(2\Delta+u/2)}{V_d(u/2)} = 2u(1 + 4\Delta/u)^d$, which concludes the proof. ◀

We now show that the shortest-path metric of the similarity constraint graph admits a Lipschitz distribution.

▶ **Lemma 11.** *For any $\Delta > 0$, the metric space $(X, \rho_{\mathcal{S}})$ admits a $(O(\sqrt{d}/\Delta), u(1+4\Delta/u)^d)$-Lipschitz distribution.*

Proof. Let $f^* : X \to \mathbb{R}^d$ be some mapping with accuracy 1. By Lemma 9 there exist some $(O(\sqrt{d}/\Delta), \Delta)$-Lipschitz distribution, $\mathcal{P}$, of $(\mathbb{R}^d, \|\cdot\|_2)$. Let P be a random partition sampled from $\mathcal{P}$. We define a partition P' of X as follows. Let $G'_{\mathcal{S}}$ be the graph obtained from $G_{\mathcal{S}}$ by deleting all edges whose endpoints under f^* are in different clusters of P. That is,

$V(G'_{\mathcal{S}}) = X$, and $E(G'_{\mathcal{S}}) = \{\{p,q\} \in \mathcal{S} : P(f^*(p)) = P(f^*(q))\}$, where $P(p)$ denotes the cluster of P containing p. We define P' to be the set of connected components of $G'_{\mathcal{S}}$. Let $\mathcal{P}'$ be the induced distribution over partitions of $(X, \rho_{\mathcal{S}})$. It remains to show that $\mathcal{P}'$ is $(O(\sqrt{d}/\Delta), u(1+4\Delta/u)^d)$-Lipschitz.

Let us first bound the probability of separation. Let $p, q \in X$. Let $x_1, \ldots, x_t$, with $x_1 = p$, $x_t = q$, be a shortest-path in $G_{\mathcal{S}}$ between p and q. Since $\mathcal{P}$ is $(O(\sqrt{d}/\Delta), \Delta)$-Lipschitz, it follows that $\Pr[f^*(p) \neq f^*(q)] \leq \frac{O(\sqrt{d})}{\Delta} \|f^*(p) - f^*(q)\|_2 \leq \frac{O(\sqrt{d})}{\Delta} \sum_{i=1}^{t-1} \|f^*(x_i) - f^*(x_{i+1})\|_2 \leq \frac{O(\sqrt{d})}{\Delta} \sum_{i=1}^{t-1} \rho_S(x_i, x_{i+1}) = \frac{O(\sqrt{d})}{\Delta} \rho_{\mathcal{S}}(p,q)$.

Finally, let us bound the diameter of the clusters in P'. Let C be a cluster in P'. By construction, $\mathsf{diam}(f^*(C)) \leq \Delta$, and $G_{\mathcal{S}}[C]$ is a connected subgraph, thus by Lemma 10 we obtain that the diameter of C in the metric space (X, ρ) is at most $u(1+4\Delta/u)^d$. Thus $\mathcal{P}'$ is $(O(\sqrt{d}/\Delta), u(1+4\Delta/u)^d)$-Lipschitz, concluding the proof. ◀

The following result allows us to sample from a Lipschitz distribution, without additional information about the geometry of the underlying space.

▶ **Lemma 12** ([22]). *Let $M = (X, \rho)$ be a metric space, and let $\Delta > 0$. Suppose that M admits a (β, Δ)-Lipschitz distribution, for some $\beta > 0$. Then there exists a randomized polynomial-time algorithm, which given M and Δ, outputs a random partition P of X, such that P is sampled from a $(2\beta, \Delta)$-Lipschitz distribution $\mathcal{P}$.*

4.2 The algorithm: Combining Lipschitz and pseudoregular partitions

We now present our final algorithm for learning Euclidean metric spaces with perfect information. First, we show that, using the algorithm from Section 3, we can learn d-dimensional Euclidean metric spaces of bounded diameter. The final algorithm combines this result with an efficient procedure for decomposing the instance using a random Lipschitz partition.

▶ **Lemma 13** (Computing Euclidean embeddings of small diameter). *Let $\varepsilon, \varepsilon' > 0$. Let $\gamma_C = (C, \mathcal{S}, \mathcal{D}, u, \ell)$, with $|C| = n$, be an instance of the metric learning problem. Suppose that γ_C admits an embedding $g^* : C \to \mathbb{R}^d$ with accuracy r^*, such that $\mathsf{diam}(g^*(C)) \leq \Delta$, for some $\Delta > 0$. Then, there exists an algorithm which given γ_C and Δ, computes a $(1+\varepsilon')$-embedding $\widehat{g} : C \to \mathbb{R}^d$, with accuracy at least $r^* - \varepsilon$. Furthermore for any fixed u and ℓ, the running time is $T(n, d, \Delta, \varepsilon, \varepsilon') = 2^{\varepsilon^{-2}(\Delta\sqrt{d}/\varepsilon')^{O(d)}}$.*

Proof. Let $c_1 = \frac{\varepsilon'}{2\sqrt{d}} \min\{\ell, u\}$. Since $\mathsf{diam}(g^*(C)) \leq \Delta$, it follows that there exists some ball $B \subset \mathbb{R}^d$ of radius 2Δ such that $g^*(C) \subset B$. Let $c_1\mathbb{Z}^d$ denote the integer lattice scaled by a factor of c_1. Let $N = B \cap (c_1\mathbb{Z}^d)$. Note that $|N| \leq (2\Delta/c_1)^d$. Let $g' : C \to \mathbb{R}^d$ be defined such that for each $p \in C$, we have that $g'(p)$ is a nearest neighbor of $g^*(p)$ in N, breaking ties arbitrarily. For any $\{p,q\} \in \mathcal{S} \cap \binom{C}{2}$, we have $\|g'(p) - g'(q)\|_2 \leq \|g'(p) - g^*(p)\|_2 + \|g^*(p) - g^*(q)\|_2 + \|g^*(q) - g'(q)\|_2$
$\leq u + \sqrt{d}c_1 \leq u(1+\varepsilon')$. Similarly, for any $\{p,q\} \in \mathcal{D} \cap \binom{C}{2}$, we have $\|g'(p) - g'(q)\|_2 \geq -\|g'(p) - g^*(p)\|_2 + \|g^*(p) - g^*(q)\|_2 - \|g^*(q) - g'(q)\|_2$
$\geq \ell - \sqrt{d}c_1 \leq \ell/(1+\varepsilon')$, and thus g' is a $(1+\varepsilon')$-embedding, with accuracy r^*.

Let $M = (N, \|\cdot\|_2)$; that is, the metric space on N endowed with the Euclidean distance. By Lemma 8 we can therefore compute a $(1+\varepsilon')$-embedding $\widehat{g} : C \to N$, with accuracy $r^* - \varepsilon$, in time $n^{O(1)} 2^{O(|N|^5/\varepsilon^2)} = 2^{\varepsilon^{-2}(\Delta\sqrt{d}/\varepsilon')^{O(d)}}$. ◀

We are now ready to prove the main result in this Section.

Proof of Theorem 3. We first construct the graph $G_{\mathcal{S}}$ as above. Let $\Delta = c\sqrt{d}u/\varepsilon$, for some universal constant $c > 0$. By Lemma 11 the metric space $(X, \rho_{\mathcal{S}})$ admits some $(O(\sqrt{d}/\Delta), \Delta')$-Lipschitz distribution, for some $\Delta' = u(1 + 4\Delta/u)^d$. By Lemma 12 we can compute, in polynomial time, some random partition P of X, such that P is distributed according to some $(O(\sqrt{d}/\Delta), \Delta')$-Lipschitz distribution $\mathcal{P}$.

Let $\mathcal{S}' = \{\{p, q\} \in \mathcal{S} : P(p) \neq P(q)\}$. Since $\mathcal{P}$ is $(O(\sqrt{d}/\Delta), \Delta')$-Lipschitz, it follows by the linearity of expectation that

$$\mathbb{E}[|\mathcal{S}'|] = \sum_{\{p,q\}\in\mathcal{S}} \Pr[P(p) \neq P(q)] \leq \sum_{\{p,q\}\in\mathcal{S}} O(\sqrt{d})\frac{\rho_{\mathcal{S}}(p,q)}{\Delta} = O(\sqrt{d})\frac{u}{\Delta}|\mathcal{S}| \leq \varepsilon|\mathcal{S}|/4, \quad (5)$$

where the last inequality holds for some large enough constant $c > 0$.

Fix some optimal embedding $f^* : X \to \mathbb{R}^d$, that satisfies all the constraints in γ. Let C be a cluster in P. Since P is Δ'-bounded, it follows that the diameter of C in $(X, \rho_{\mathcal{S}})$ is at most Δ'. Thus, for any $p, q \in C$, there exists in $G_{\mathcal{S}}$ some path $Q = x_0, \ldots, x_L$ between p and q, with $L \leq \Delta'/u$ edges. Thus $\|f^*(p) - f^*(q)\|_2 \leq \sum_{i=0}^{L-1} \|f^*(x_i) - f^*(x_{i+1})\|_2 \leq L \cdot u \leq \Delta'$, which implies that $\mathsf{diam}(f^*(C)) \leq \Delta'$. Furthermore f^* has accuracy 1.

Let $\mathcal{S}_C = \mathcal{S} \cap \binom{C}{2}$, $\mathcal{D}_C = \mathcal{D} \cap \binom{C}{2}$. Then $\gamma_C = (C, \mathcal{S}_C, \mathcal{D}_C, u, u)$ is an instance of non-linear metric learning in d-dimensional Euclidean space that admits a solution $f^* : C \to \mathbb{R}^d$ with accuracy 1, with $\mathsf{diam}(f^*(C)) \leq \Delta'$. Therefore, for each $C \in P$, using Lemma 13 we can compute some $(1 + \varepsilon')$-embedding $f_C : C \to \mathbb{R}^d$, with accuracy at least $1 - \varepsilon/2$, in time $T(n, d, \Delta', \varepsilon/2, \varepsilon') = n^{O(1)}2^{(\frac{d}{\varepsilon\varepsilon'})^{O(d^2)}}$. We can now combine all these maps f_C into a single map $\widehat{f} : X \to \mathbb{R}^d$, by translating each image $f_C(C)$ such that for any distinct $C, C' \in P$, for any $p \in C$, $p' \in C'$, we have $\|f_C(p) - f_{C'}(p')\|_2 \geq u$. For any $p \in C$, we set $f(p) = f_C(p)$, where C is the unique cluster in P containing p. It is immediate that any constraint $\{p, q\} \in \mathcal{S}$, with p and q in different clusters is violated by f. The total number of such violations is $|\mathcal{S}'|$. By Markov's inequality and (5), these violations are at most $\varepsilon|\mathcal{S}|/2$, with probability at least $1/2$. Similarly, any $\{p, q\} \in \mathcal{D}$, with p and q in different clusters is satisfied by f. All remaining violations are on constraints with both endpoints in the same cluster of P. Since the accuracy of each f_C is at least $1 - \varepsilon$, it follows that the total number of these violations is at most $\varepsilon(|\mathcal{S}| + |\mathcal{D}|)/2$. Therefore, the total number of violations among all constraints is at most $\varepsilon(|\mathcal{S}| + |\mathcal{D}|)/2 + \varepsilon|\mathcal{S}|/2 \leq \varepsilon(|\mathcal{S}| + |\mathcal{D}|)$, with probability at least $1/2$. In other words, the accuracy of f is at least $1 - \varepsilon$, with probability at least $1/2$. The success probability can be increased to $1 - 1/n^c$, for any constant $c > 0$, by repeating the algorithm $O(\log n)$ times and returning the best embedding found. Finally, the running time is dominated by the computation of the maps f_C, for all clusters C in P. Since there are at most n such clusters, the total running time is $n^{O(1)}2^{(\frac{d}{\varepsilon\varepsilon'})^{O(d^2)}}$, concluding the proof. ◀

5 Learning Euclidean metric spaces with imperfect information

In this Section we describe our algorithm for learning d-dimensional Euclidean metric spaces with imperfect information. We first obtain some preliminary results on graph partitioning via sparse cuts. We next show that for any instance whose similarity constraints induce an expander graph, the optimal solution must be contained inside a ball of small diameter. Our final algorithm combines these results with the algorithm from Section 3 for embedding into host metric spaces of bounded cardinality.

Fix some instance $\gamma = (X, \mathcal{S}, \mathcal{D}, u, \ell)$. For the remainder of this Section we define the graphs $G_{\mathcal{S}} = (X, E_{\mathcal{S}})$, $G_{\mathcal{D}} = (X, E_{\mathcal{D}})$, and $G_{\mathcal{S}\cup\mathcal{D}} = (X, E_{\mathcal{S}\cup\mathcal{D}})$, where $E_{\mathcal{S}} = \mathcal{S}$, $E_{\mathcal{D}} = \mathcal{D}$, and $E_{\mathcal{S}\cup\mathcal{D}} = \mathcal{S} \cup \mathcal{D}$.

5.1 Well-linked decompositions

We recall some results on partitioning graphs into well-connected components. An instance to the Sparsest-Cut problem consists of a graph G, with each edge $\{x,y\} \in E(G)$ having capacity $\mathsf{cap}(x,y) \geq 0$, and each pair $x,y \in V(G)$ having *demand* $\mathsf{dem}(x,y) \geq 0$. The goal is to find a cut $(U,\overline{U})$ of minimum *sparsity*, denoted by $\phi(U)$, which is defined as $\phi(U) = \frac{\mathsf{cap}(U,\overline{U})}{\mathsf{dem}(U,\overline{U})}$, where $\mathsf{cap}(U,\overline{U}) = \sum_{\{x,y\}\in E(U,\overline{U})} \mathsf{cap}(x,y)$, $\mathsf{dem}(U,\overline{U}) = \sum_{\{x,y\}\in E(U,\overline{U})} \mathsf{dem}(x,y)$, and $E(U,\overline{U})$ denotes the set of edges between U and $\overline{U}$. We recall the following result of Arora, Lee and Naor [3] on approximating the Sparsest-Cut problem (see also [4]).

▶ **Theorem 14** ([3]). *There exists a polynomial-time $O(\sqrt{\log n}\log\log n)$-approximation for the Sparsest-Cut problem on n-vertex graphs.*

▶ **Lemma 15.** *Let $\alpha > 0$. There exists a polynomial-time algorithm which computes some $E' \subset \mathcal{S}\cup\mathcal{D}$, with $|E'| \leq \alpha|\mathcal{S}\cup\mathcal{D}|$, such that for every connected component C of $G_{\mathcal{S}} \setminus E'$, and any $U \subset C$, we have $|E_{\mathcal{S}}(U, C\setminus U)| = \Omega\left(\frac{\alpha}{\log^{3/2} n \log\log n}\right)|U|\cdot|C\setminus U|$.*

Proof. We compute E' inductively starting with $E' = \emptyset$. We construct an instance of the Sparsest-Cut problem on graph $G_{\mathcal{S}}$, where for any $\{x,y\} \in \mathcal{S}$ we have $\mathsf{cap}(x,y) = 1$, and for any $x,y \in X$, we have $\mathsf{dem}(x,y) = 1$. Let $\chi = \frac{\alpha}{c\log^{3/2} n \log\log n}$, for some universal constant $c > 0$ to be specified. If there exists a cut of sparsity at most χ, then by Theorem 14 we can compute a cut $(U,\overline{U})$ of sparsity $O(\chi\sqrt{\log n}\log\log n)$. We add to E' all the edges in $E_{\mathcal{S}}(U,\overline{U})$, and we recurse on the subgraphs $G_{\mathcal{S}}[U]$ and $G_{\mathcal{S}}[\overline{U}]$. If no cut with the desired sparsity exists, then we terminate the recursion. For each cut $(U,\overline{U})$ found, the edges in $E_{\mathcal{S}}(U,\overline{U})$ that were added to E' are charged to the edges in $E_{\mathcal{S}\cup\mathcal{D}}(U,\overline{U})$. In total, we charge only $O(\chi\sqrt{\log n}\log\log n)$-fraction of all edges, and thus $|E'| = O(\chi n \log^{3/2} n \log\log n) \leq \alpha|\mathcal{S}\cup\mathcal{D}|$, where the last inequality follows for some sufficiently large constant $c > 0$, concluding the proof. ◀

5.2 Isoperimetry and the diameter of embeddings

Here we show that if the graph $G_{\mathcal{S}}$ of similarity constraints is "expanding" relative to $G_{\mathcal{S}\cup\mathcal{D}}$, then there exists an embedding with near-optimal accuracy, that has an image of small diameter. Technically, we require that in any cut in $G_{\mathcal{S}\cup\mathcal{D}}$, a significant fraction of its edges are in $\mathcal{S}$. Intuitively, this condition forces almost all points to be mapped within a small ball. The next Lemma formalizes this statement.

▶ **Lemma 16.** *Let $F \subset \mathcal{S}$, with $|F| \leq \zeta\binom{|X|}{2}$, for some $\zeta > 0$. Suppose that for all $U \subset X$, we have*

$$|E_{\mathcal{S}}(U, X\setminus U)| \geq \alpha'\cdot|U|\cdot|X\setminus U|, \tag{6}$$

for some $\alpha' > 0$. Then there exists $J \subseteq X$, satisfying the following conditions:
(1) *No edge in F has both endpoints in J; that is, $F \cap E_{\mathcal{S}}(J) = \emptyset$.*
(2) *$|E_{\mathcal{S}\cup\mathcal{D}}(J)| \geq (1-\zeta/\alpha')\binom{|X|}{2}$.*
(3) *For all $U \subset J$, we have*

$$|E_{\mathcal{S}}(U, J\setminus U)| = \Omega(\alpha')|U|\cdot|J\setminus U|. \tag{7}$$

Proof. Let C be the largest connected component of $G_{\mathcal{S}} \setminus F$, and let $C' = X\setminus C$. By (6) we get

$$|E_{\mathcal{S}\cup\mathcal{D}}(C,C')| = |C|\cdot|C'| = O\left(1/\alpha'\right)|E_{\mathcal{S}}(C,C')| = O\left(1/\alpha'\right)|F| = O\left(\zeta/\alpha'\right)|X|^2. \tag{8}$$

Since $G_{\mathcal{S}\cup\mathcal{D}}$ is a complete graph, it follows from (8) that the number of edges not in $G_{\mathcal{S}\cup\mathcal{D}}[C]$ is at most $O(\zeta/\alpha')|X|^2$, and therefore

$$|E_{\mathcal{S}\cup\mathcal{D}}(C)| \geq (1 - O(\zeta/\alpha'))\binom{|X|}{2}. \tag{9}$$

Next, we compute some $J \subseteq C$ such that for all $U \subset C$,

$$E_{\mathcal{S}}(U, J \setminus U) \geq c'\alpha' \cdot |U| \cdot |J \setminus U|, \tag{10}$$

for some universal constant c' to be determined. This is done as follows. If there is no $U \subset C$ violating (10), then we set $J = C$. Otherwise, pick some $U \subset C$, such that

$$E_{\mathcal{S}}(U, C \setminus U) < c'\alpha' \cdot |U| \cdot |C \setminus U|. \tag{11}$$

Assume w.l.o.g. that $|U| \leq |C \setminus U|$. We delete U and we recurse on $C \setminus U$. All deleted edges are charged to the edges in F that are incident to U. It follows by (11) and (6) that, for some sufficiently small constant $c' > 0$, at least a 2/3-fraction of the edges incident to U are in F. Thus $|F \cap E_{\mathcal{S}}(U, X \setminus U)| = \Omega(\alpha')|U| \cdot |X \setminus U|$. We charge the deleted edges (that is, the edges inside the deleted component U and in $E_{\mathcal{S}}(U, X \setminus U)$) to the edges in $F \cap E_{\mathcal{S}}(U, X \setminus U)$, with each edge thus receiving $O(1/\alpha')$ units of charge. When we recurse on $C \setminus U$, the part of F that was incident to U is replaced by $E_{\mathcal{S}}(U, C \setminus U)$, and thus it decreases in size by at least a factor of 2. Therefore, the total charge that we pay throughout the construction is at most $O(1/\alpha')|F| = O(\zeta/\alpha')|X|^2$, which is an upper bound on the number of all deleted edges. This completes the construction of J. Combining with (9) we get $|E_{\mathcal{S}\cup\mathcal{D}}(J)| \geq |E_{\mathcal{S}\cup\mathcal{D}}(C)| - O(\zeta/\alpha')|X|^2 \geq (1 - O(\zeta/\alpha'))\binom{|X|}{2}$, which concludes the proof. ◀

▶ **Lemma 17.** *Let $M = (Y, \rho)$ be any metric space. Suppose that γ admits an embedding $f^* : X \to Y$ with accuracy $1 - \zeta$, for some $\zeta > 0$. Suppose further that for all $U \subset X$, we have*

$$E_{\mathcal{S}}(U, X \setminus U) \geq \alpha' \cdot |U| \cdot |X \setminus U|, \tag{12}$$

for some $\alpha' > 0$. Then there exists an embedding $f' : X \to Y$, with accuracy at least $1 - O(\zeta/\alpha')$, such that $\mathsf{diam}_M(f'(X)) = O\left(\frac{u \log n}{\alpha'}\right)$.

Proof. Let $J \subseteq X$ be given by Lemma 16, where we set F to be the set of constraints in $\mathcal{S}$ that are violated by f^*. We define $f' : X \to Y$ by mapping each $x \in J$ to $f^*(x)$, and each $x \in X \setminus J$ to some arbitrary point in $f^*(J)$. All constraints in $E_{\mathcal{S}}(J)$ are satisfied by f^*, and thus also by f'. It follows that f' can only violate constraints that are violated by f^*, and constraints that are not in $E_{\mathcal{S}\cup\mathcal{D}}(J)$. Therefore the accuracy of f' is at least $1 - \zeta - O(\zeta/\alpha') = 1 - O(\zeta/\alpha')$.

By Lemma 16 we have that for all $U \subset J$, $|E_{\mathcal{S}}(U, J \setminus U)| = \Omega(\alpha')|U| \cdot |J \setminus U|$. Therefore, the combinatorial diameter (that is, the maximum number of edges in any shortest path) of $G_{\mathcal{S}}[J]$ is at most $O((\log n)/\alpha')$. For all $\{x, y\} \in E_{\mathcal{S}}(J)$ we have $\rho(f'(x), f'(y)) = \rho(f^*(x), f^*(y)) \leq u$, it follows that $\mathsf{diam}_M(f'(X)) = \mathsf{diam}_M(f'(J)) = O(\frac{u \log n}{\alpha'})$. ◀

5.3 The algorithm

We are now ready to prove the main result of this Section.

Proof of Theorem 4. Fix some optimal embedding $f^* : X \to \mathbb{R}^d$, with accuracy $1-\zeta$. Let $E' \subset \mathcal{S} \cup \mathcal{D}$ be the set of constraints computed by the algorithm in Lemma 15. We have $|E'| \leq \alpha|\mathcal{S} \cup \mathcal{D}|$, for some $\alpha > 0$ to be determined.

For each connected component C of $G_{\mathcal{S}} \setminus E'$, let $\gamma_C = (C, \mathcal{S}_C, \mathcal{D}_C, u, \ell)$ be the restriction of γ on C; that is, $\mathcal{S}_C = \mathcal{S} \cap \binom{C}{2}$, and $\mathcal{S}_D = \mathcal{D} \cap \binom{C}{2}$. Let f^*_C be the restriction of f^* on C, and let the accuracy of f^*_C be $1-\zeta_C$, for some $\zeta_C \in [0,1]$. Let F_C be the set of constraints in $E_{\mathcal{S}}(C)$ that are violated by f^*_C. By Lemma 17 it follows that there exists an embedding $f'_C : X \to \mathbb{R}^d$, with accuracy $1-O(\zeta_C/\alpha')$, and such that $\mathsf{diam}(f'_C(C)) = O(\frac{u \log n}{\alpha'})$, for some $\alpha' = \Omega(\frac{\alpha}{\log^{3/2} n \log\log n})$.

Using Lemma 13 we can compute a $(1+\varepsilon')$-embedding $\widehat{f}_C : C \to \mathbb{R}^d$ with accuracy at least $1-O(\zeta_C/\alpha)-\varepsilon$, in time $n^{O(1)} 2^{\varepsilon^{-2}(\frac{\Delta\sqrt{d}}{\varepsilon'})^{O(d)}}$. We can combine all the embeddings $\widehat{f}_C$ into a single embedding $\widehat{f} : X \to \mathbb{R}^d$ by translating the images of $\widehat{f}_C(C)$ and $\widehat{f}_{C'}(C')$ in $\mathbb{R}^d$ to that their distance is at least ℓ, for all distinct components C, C'. It is immediate that the resulting embedding $\widehat{f}$ can only violate the constraints in E', and the constraints violated by all the $\widehat{f}_C$. Thus, the accuracy of $\widehat{f}$ is at least $1-\alpha-O(\zeta/\alpha')-\varepsilon$. Setting $\alpha = \zeta^{1/2} \log^{3/4} n (\log\log n)^{1/2}$, we get that the accuracy of $\widehat{f}$ is at least $1-O(\zeta^{1/2} \log^{3/4} n (\log\log n)^{1/2})-\varepsilon$. The running time is dominated by the at most n executions of the algorithm from Lemma 13, and thus it is at most $n^{O(1)} 2^{\varepsilon^{-2}(\frac{d \log n}{\zeta\varepsilon'})^{O(d)}}$, for any fixed $u > 0$, concluding the proof. ◀

References

1. Nir Ailon and Moses Charikar. Fitting tree metrics: Hierarchical clustering and phylogeny. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 73–82. IEEE, 2005.
2. Noga Alon, Mihai Bădoiu, Erik D Demaine, Martin Farach-Colton, MohammadTaghi Hajiaghayi, and Anastasios Sidiropoulos. Ordinal embeddings of minimum relaxation: general properties, trees, and ultrametrics. *ACM Transactions on Algorithms (TALG)*, 4(4):46, 2008.
3. Sanjeev Arora, James Lee, and Assaf Naor. Euclidean distortion and the sparsest cut. *Journal of the American Mathematical Society*, 21(1):1–21, 2008.
4. Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):5, 2009.
5. Mihai Badoiu. Approximation algorithm for embedding metrics into a two-dimensional space. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 434–443. Society for Industrial and Applied Mathematics, 2003.
6. Mihai Bădoiu, Erik D Demaine, MohammadTaghi Hajiaghayi, Anastasios Sidiropoulos, and Morteza Zadimoghaddam. Ordinal embedding: Approximation algorithms and dimensionality reduction. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 21–34. Springer, 2008.
7. Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine learning*, 56(1-3):89–113, 2004.
8. Nikhil Bansal and Ryan Williams. Regularity lemmas and combinatorial algorithms. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, pages 745–754. IEEE, 2009.
9. Yair Bartal. Probabilistic Approximations of Metric Spaces and Its Algorithmic Applications. In *37th Annual Symposium on Foundations of Computer Science, FOCS '96, Burlington, Vermont, USA, 14-16 October, 1996*, pages 184–193. IEEE Computer Society, 1996. doi:10.1109/SFCS.1996.548477.

10 Yonatan Bilu and Nati Linial. Monotone maps, sphericity and bounded second eigenvalue. *Journal of Combinatorial Theory, Series B*, 95(2):283–299, 2005.

11 Heinz Breu and David G Kirkpatrick. Unit disk graph recognition is NP-hard. *Computational Geometry*, 9(1-2):3–24, 1998.

12 Mihai Bădoiu, Piotr Indyk, and Anastasios Sidiropoulos. Approximation algorithms for embedding general metrics into trees. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 512–521. Society for Industrial and Applied Mathematics, 2007.

13 Moses Charikar, Chandra Chekuri, Ashish Goel, Sudipto Guha, and Serge Plotkin. Approximating a finite metric by a small number of tree metrics. In *Foundations of Computer Science, 1998. Proceedings. 39th Annual Symposium on*, pages 379–388. IEEE, 1998.

14 Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

15 Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

16 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.

17 Alan Frieze and Ravi Kannan. The regularity lemma and approximation schemes for dense problems. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 12–20. IEEE, 1996.

18 Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

19 Sariel Har-Peled. *Geometric approximation algorithms*, volume 173. American mathematical society Boston, 2011.

20 Piotr Indyk, Jiří Matoušek, and Anastasios Sidiropoulos. Low-Distortion Embeddings of Finite Metric Spaces. In Jacob E. Goodman, Joseph O'Rourke, and Csaba D. Toth, editors, *Handbook of Discrete and Computational Geometry, Second Edition.* Chapman and Hall/CRC, 2017.

21 Philip Klein, Serge A Plotkin, and Satish Rao. Excluded minors, network decomposition, and multicommodity flow. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 682–690. ACM, 1993.

22 Robert Krauthgamer and Tim Roughgarden. Metric Clustering via Consistent Labeling. *Theory OF Computing*, 7:49–74, 2011.

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

24 Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

25 Jiří Matoušek. *Lectures on discrete geometry*, volume 212. Springer Science & Business Media, 2002.

26 Jiří Matoušek and Anastasios Sidiropoulos. Inapproximability for metric embeddings into $\mathbb{R}^d$. *Transactions of the American Mathematical Society*, 362(12):6341–6365, 2010.

27 Amir Nayyeri and Benjamin Raichel. Reality distortion: Exact and approximate algorithms for embedding into the line. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 729–747. IEEE, 2015.

28 Amir Nayyeri and Benjamin Raichel. A Treehouse with Custom Windows: Minimum Distortion Embeddings into Bounded Treewidth Graphs. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 724–736. SIAM, 2017. doi:10.1137/1.9781611974782.46.

29 Gregory Shakhnarovich. *Learning task-specific similarity.* PhD thesis, Massachusetts Institute of Technology, 2005.

30 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.