Approximation Algorithms for Low-Distortion Embeddings Into Low-Dimensional Spaces

by

Anastasios Sidiropoulos

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[June 2005]
May 2005

Author ....
Department of Electrical Engineering and Computer Science
May 6, 2005

Certified by ....
Piotr Indyk
Associate Professor
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Approximation Algorithms for Low-Distortion Embeddings Into Low-Dimensional Spaces

by

Anastasios Sidiropoulos

Submitted to the Department of Electrical Engineering and Computer Science
on May 6, 2005, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

We present several approximation algorithms for the problem of embedding metric spaces into a line, and into the two-dimensional plane. We give an $O(\sqrt{n})$-approximation algorithm for the problem of finding a line embedding of a metric induced by a given unweighted graph, that minimizes the (standard) multiplicative distortion. For the same problem, we give an exact algorithm, with running-time exponential in the distortion. We complement these results by showing that the problem is NP-hard to α-approximate, for some constant $\alpha > 1$.

For the two-dimensional case, we show a $O(\sqrt{n})$ upper bound for the distortion required to embed an n-point subset of the two-dimensional sphere, into the plane. We prove that this bound is asymptotically tight, by exhibiting n-point subsets such that any embedding into the plane has distortion $\Omega(\sqrt{n})$. These techniques yield a $O(1)$-approximation algorithm for the problem of embedding an n-point subset of the sphere into the plane.

Thesis Supervisor: Piotr Indyk
Title: Associate Professor

Acknowledgments

I would like to thank my research supervisor Piotr Indyk, for his constant help and guidance.

I would also like to thank Ioannis Caragiannis, Stavros Cosmadakis, Efstratios Gallopoulos, Christos Kaklamanis, and Periklis Papakonstantinou, for their help and influence during my undergraduate studies in the University of Patras.

I also thank my co-authors here are MIT, Noga Alon, Mihai Badoiu, Erik Demaine, Martin Farach-Colton, Vahab Mirrokni, Yuri Rabinovich, Mohammad Taghi-Hajiaghayi, and Vinod Vaikunthanathan.

I am very grateful to the family of Paris Kanelakis, and the Onassis Public Benefit Foundation, for supporting my research with fellowships during the past two years.

Finally, I want to thank my friends Nikos, Theofilos, Thodoris, Giannis^3, Dimitra, Manolis, and Dimitris, and my mother Litsa, my grandmother Katerina, and my sister Maria, for all their love and support.

This Thesis is dedicated to the memory of my father, Nikos Sidiropoulos.

Anastasios Sidiropoulos
Spring 2005

Contents

List of Figures

Chapter 1

Introduction

Embedding distance matrices into geometric spaces (most notably, into low-dimensional spaces) is a fundamental problem occurring in many applications. In the context of data visualization, this approach allows the user to observe the structure of the data set and discover its interesting properties. In computational chemistry, this approach is used to recreate the geometric structure of the data from the distance information. The problem is of interest in many other areas, see [24] for a discussion.

The methods for computing such embeddings have their roots in work going back to the first half of the 20th century, and in the more recent work of Shepard [22, 23], Kruskal [15, 16], and others. The area is usually called *Multi-dimensional Scaling* (MDS) and is a subject of extensive research [24]. However, despite significant practical interest, few theoretical results exist in this area (see Related Work). The most commonly used algorithms are heuristic (e.g., gradient-based method, simulated annealing, etc) and are often not satisfactory in terms of the running time and/or quality of the embeddings.

In this paper we present algorithms for the following fundamental embedding problem: given a graph $G = (V, E)$ inducing a shortest path metric $M = M(G) = (V, D)$, find a mapping f of V into a *line* that is non-contracting (i.e., $|f(u) - f(v)| \geq D(u, v)$ for all $u, v \in V$) and minimizes the distortion $c_{line}(M, f) = \max_{u,v \in V} \frac{|f(u)-f(v)|}{D(u,v)}$. That is, our goal is to find $c_{line}(M) = \min_f c_{line}(M, f)$. For the case when G is an *unweighted* graph, we show the following algorithms for this problem (denote $n = |V|$):

- A polynomial (in fact, $O(n^3c)$-time) c-approximation algorithm for metrics M for which $c_{line}(M) \leq c$. This also implies an $O(\sqrt{n})$-approximation algorithm for any M (Chapter 2).
- An exact algorithm, with running time $n^{O(c_{line}(M))}$ (Chapter 3).

We complement our algorithmic results by showing that a-approximating the value of $c_{line}(M)$ is NP-hard for certain $a > 1$ in Chapter 4. In particular, this justifies the exponential dependence on $c_{line}(M)$ in the running time bound for the exact algorithm.

We also study the problem of embedding metrics into the *plane* in Chapter 5. In particular, we focus on embedding metrics $M = (X, D)$ which are induced by a set of points in a unit sphere S^2. Embedding such metrics is important, e.g., for the purpose of visualizing point-sets representing places on Earth or other planets, on a (planar) computer screen.[1] In general, we show that an n-point spherical metric can be embedded with distortion $O(\sqrt{n})$, and this bound is optimal in the worst case. (The lower bound is shown by resorting to the Borsuk-Ulam theorem [3], which roughly states that any continuous mapping from S^2 into the plane maps two antipodes of S^2 into the same point.) For the algorithmic problem of embedding M into the plane, we give a 3.512-approximation algorithm, when D is Euclidean distance in $\mathbb{R}^3$.

1.1 Related work

1.1.1 Combinatorial vs Algorithmic Problem.

The problem of finding low-distortion embeddings of metrics into geometric spaces has been long a subject of extensive mathematical studies. During the last few years, such embeddings found multiple and diverse uses in computer science as well; many such applications have been surveyed in [11]. However, the problems addressed in this paper are fundamentally different from those investigated in the aforementioned

[1] Indeed, the whole field of cartography is devoted to low-distortion representations of spherical maps into the plane.

literature. In a nutshell, our problems are *algorithmic*, as opposed to *combinatorial.* More specifically, we are interested in finding the best distortion embedding of a *given* metric (which is an algorithmic problem) as opposed to the best distortion embedding for a *class* of metrics (which is a combinatorial problem). Thus, we define the quality of an embedding algorithm as the worst-case *ratio* of the distortion obtained by the algorithm to the best achievable distortion. In contrast, the combinatorial approach focuses on providing the worst-case upper bounds for the distortion itself. Thus, the problems are fundamentally different, which raises new interesting issues.

Despite the differences, we mention two combinatorial results that are relevant in our context. The first one is the [18] adaptation of Bourgain's construction [4] that enables embedding of an arbitrary metric into $l_2^{O(\log^2 n)}$ with maximum multiplicative distortion $O(\log n)$. It should be noted, however, that for the applications mentioned earlier, the most interesting spaces happen to be low-dimensional. Similarly, any metric can be embedded into d-dimensional Euclidean space with multiplicative distortion $O(\min[n^{\frac{2}{d}} \log^{3/2} n, n])$ and no better than $\Omega(n^{1/\lfloor (d+1)/2 \rfloor})$ [20]. However, the worst-case guarantees are rather large for small d, especially for the case $d = 1$ that we consider here.

1.1.2 Previous Work on the Algorithmic Problem.

To our knowledge there have been few *algorithmic* embedding results. Hastad et al. gave a 2-approximation algorithm for embedding an arbitrary metric into a line $\mathbb{R}$, when the *maximum additive two-sided error* was considered; that is, the goal was to optimize the quantity $\max_{u,v} ||f(u) - f(v)| - D(u,v)|$. They also showed that the same problem cannot be approximated within 4/3 unless $P = NP$ [10, 12]. Bădoiu extended the algorithm to the 2-dimensional plane with maximum two-sided additive error when the distances in the target plane are computed using the l_1 norm [5]. Bădoiu, Indyk and Rabinovich [2] gave a weakly-quasi-polynomial time algorithm for the same problem in the l_2 norm.

Very recently, Kenyon, Rabani and Sinclair [13] gave *exact* algorithms for minimum (multiplicative) distortion embeddings of metrics *onto* simpler metrics (e.g., line metrics). Their algorithms work as long as the minimum distortion is small, e.g., constant. We note that constraining the embeddings to be *onto* (not *into*, as in our case) is crucial for the correctness of their algorithms.

In general, one can choose non-geometric metric spaces to serve as the host space. For example, in computational biology, approximating a matrix of distances between different genetic sequences by an ultrametric or a tree metric allows one to retrace the evolution path that led to formation of the genetic sequences. Motivated by these applications M. Farach-Colton and S. Kannan show how to find an *ultrametric* T with minimum possible maximum additive distortion [7]. There is also a 3-approximation algorithm for the case of embedding arbitrary metrics into weighted tree metrics to minimize the maximum additive two-sided error [1]. [6] recently gave an $O(\log^{1/p} n)$-approximation for embedding arbitrary n-point metrics into the line to minimize the ℓ_p norm of the two-sided error vector $||f(u) - f(v)| - D(u, v)|$.

Distortion vs Bandwidth. In the context of unweighted graphs, the notion of minimum distortion of an embedding into a line is closely related to the notion of a graph *bandwidth.* Specifically, if the non-contraction constraint $|f(u)-f(v)| \geq D(u, v)$ is replaced by a constraint $|f(u) - f(v)| \geq 1$ for $u \neq v$, then $c_1(M(G))$ becomes precisely the same as the bandwidth of the graph G.

There are several algorithms that approximate the bandwidth of a graph [8, 9]. Unfortunately, they do not seem applicable in our setting, since they do not enforce the non-contraction constraint for all node pairs. However, in the case of *exact* algorithms the situation is quite different. In particular, our exact algorithm for computing the distortion is based on the analogous algorithm for the bandwidth problem by Saxe [21].

Chapter 2

A $O(c)$-Approximation Algorithm

In this chapter we present a $O(c)$-approximation algorithm for embedding the shortest-path metric of an unweighted graph into the line, where c denotes that optimal distortion. That is, our algorithm outputs an embedding with distortion $O(c^2)$. By combining this algorithm with the fact that $c = O(n)$ for any input metric, we can obtain a $O(\sqrt{n})$-approximation algorithm.

2.1 Description of the Algorithm

We start by stating an algorithmic version of a fact proved in [19].

Lemma 1. *Any shortest path metric over an unweighted graph $G = (V, E)$ can be embedded into a line with distortion at most $2n - 1$ in time $O(|V| + |E|)$.*

Proof. Let T be a spanning tree of the graph. We replace every (undirected) edge of T with a pair of opposite directed edges. Since the resulting graph is Eulerian, we can consider an Euler tour C in T. Starting from an arbitrary node, we embed the nodes in T according to the order that they appear in C, ignoring multiple appearances of a node, and preserving the distances in C. Clearly, the resulting embedding is non-contracting, and since C has length $2n$, the distortion is at most $2n - 1$. □

Note that the $O(n)$ bound is tight, e.g. when G is a star, or a cycle.

Let $G = (V, E)$ be a graph, such that there exists an embedding of G with distortion c. The algorithm for computing an embedding of distortion at most $O(c^2)$ is the following:

1. Let f_{OPT} be an optimal embedding of G (note that we just assume the existence of such an embedding, without computing it). Guess nodes $t_1, t_2 \in V$, such that $f_{OPT}(t_1) = \min_{v \in V} f_{OPT}(v)$, and $f_{OPT}(t_2) = \max_{v \in V} f_{OPT}(v)$.

2. Compute the shortest path $p = v_1, v_2, \ldots, v_L$ from t_1 to t_2.

3. Partition V into disjoint sets $V_1, V_2, \ldots V_L$, such that for each $u \in V_i$, $D(u, v_i) = \min_{1 \leq j \leq L} D(u, v_j)$. Break ties so that each V_i is connected.

4. For $i = 1 \ldots L$, compute a spanning tree T_i of the subgraph induced by V_i, rooted at v_i. Embed the nodes of V_i as in the proof of Lemma 1, leaving a space of length $|V_i| + |V_{i+1}| + 1$ between the nodes of V_i and V_{i+1}.

2.2 Analysis

We will analyze the algorithm presented in the previous section. The first step is to show that every set V_i has small diameter.

Lemma 2. *For every $i, 1 \leq i \leq L$, and for every $x \in V_i$, we have $D(v_i, x) \leq c/2$.*

Proof. Assume that the assertion is not true. That is, there exists v_i, and $x \in V_i$, such that $D(x, v_i) > c/2$. Consider the optimal embedding f_{OPT}. By the fact that v_1 and v_L are the left-most and right-most embedded nodes in the embedding f_{OPT}, it follows that there exists $j, 1 \leq j < L$, such that $f_{OPT}(x)$ lies between $f_{OPT}(v_j)$, and $f_{OPT}(v_{j+1})$. W.l.o.g., assume that

$$f_{OPT}(v_j) < f_{OPT}(x) < f_{OPT}(v_{j+1}).$$

Since $x \in V_i$, we have

$$\begin{aligned} |f_{OPT}(v_{j+1}) - f_{OPT}(v_j)| &= f_{OPT}(v_{j+1}) - f_{OPT}(x) + f_{OPT}(x) - f_{OPT}(v_j) \\ &\geq D(v_{j+1}, x) + D(x, v_j) \\ &\geq 2D(x, v_i) \\ &> c. \end{aligned}$$

This is a contradiction, since the expansion of f_{OPT} is at most c. □

We also need to bound the total size of c consecutive sets V_i.

Lemma 3. *For every i, $1 \leq i \leq L - c + 1$, we have $\sum_{j=i}^{i+c-1} |V_j| \leq 2c^2$.*

Proof. Assume that there exists i such that $\sum_{j=i}^{i+c-1} |V_j| > 2c^2$. Note that

$$\max_{i \leq j_1 < j_2 \leq i+c-1} |f_{OPT}(v_{j_1}) - f_{OPT}(v_{j_2})| \leq c(c-1).$$

Moreover, since $\sum_{j=i}^{i+c-1} |V_j| > 2c^2$, we have $\max_{u,w \in \bigcup_{j=i}^{i+c-1} V_j} |f_{OPT}(u) - f_{OPT}(w)| \geq 2c^2$. It follows that there exists $u \in V_l$, for some l, with $i \leq l \leq i + c - 1$, such that $|f_{OPT}(v_l) - f_{OPT}(u)| \geq \frac{2c^2 - c(c-1)}{2} > c^2/2$. Since the expansion is at most c, we have $D(v_l, u) > c/2$, contradicting Lemma 2. □

Lemma 4. *The embedding computed by the algorithm is non-contracting.*

Proof. Let $x, y \in V$. If x and y are in the same set V_i, for some i, then clearly $|f(x) - f(y)| \geq D(x, y)$, since the distance between x and y produced by a traversal of the spanning tree of the graph induced by V_i is at least the distance of x and y on T_i, which is at least $D(x, y)$.

Assume now that $x \in V_i$ and $y \in V_j$, for some $i < j$. We have

$$\begin{aligned} |f(y) - f(x)| &\geq \sum_{t=i}^{j-1}(|V_t| + |V_{t+1}| + 1) \\ &\geq |V_i| + (j - i) + |V_j| \\ &\geq D(x, v_i) + D(v_i, v_j) + D(v_j, y) \\ &\geq D(x, y) \end{aligned}$$

□

The next Lemma bounds the contraction of the embedding.

Lemma 5. *The expansion of the embedding computed by the algorithm is at most $4c^2$.*

Proof. It suffices to show that for each $\{x, y\} \in E$, $|f(x) - f(y)| \leq 4c^2$. Let $x \in V_i$, and $y \in V_j$. If $|i - j| \leq 2c$, then by Lemma 3 we obtain that $|f(x) - f(y)| \leq 4c^2$.

Assume now that there exist nodes $x \in V_i$ and $y \in V_j$, with $\{x, y\} \in E$, and $|i - j| > 2c$. By Lemma 2, we obtain that $D(v_i, x) \leq c/2$, and $D(y, v_j) \leq c/2$, and thus $|i - j| = D(v_i, v_j) \leq c + 1$, a contradiction. □

Theorem 1. *The described algorithm computes a non-contracting embedding of maximum distortion $O(c^2)$, in time $O(n^3 c)$.*

Proof. By Lemmata 4 and 5, it follows that the computed embedding is non-contracting and has distortion at most $O(c^2)$. In the beginning of the algorithm, we compute all-pairs shortest paths for the graph. Next, for each possible pair of nodes t_1 and t_2, the described embedding can be computed in linear time. Thus, the total running time is $O(n^2|E|) = O(n^3 c)$. □

Theorem 2. *There exists a $O(\sqrt{n})$-approximation algorithm for the minimum distortion embedding problem.*

Proof. If the optimal distortion c is at most $\sqrt{n}$, then the described algorithm computes an embedding of distortion at most $O(c\sqrt{n})$. Otherwise, the algorithm de-

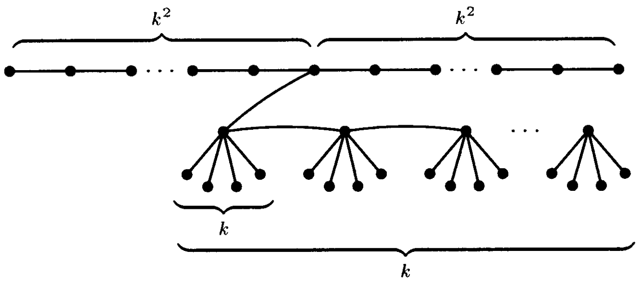

Figure 2-1: A tight instance for the algorithm.

scribed in Lemma 1, computes an embedding of distortion $O(n)$. Thus, by taking the best of the above two embeddings, we obtain an $O(\sqrt{n})$-approximation. □

2.3 Tightness of the Analysis

Figure 2-1 depicts an instance which demonstrates that the above analysis is actually tight, even for the case of trees. More specifically, the depicted tree can be optimally embedded with distortion $O(k)$. To see that, observe that there exist nodes of degree $O(k)$, thus this is indeed a lower bound for the optimal distortion. Furthermore, we can achieve this bound by first embedding the left half of the long path, and then interleaving the stars of size $O(k)$ with the nodes of the right half of the long path, thus achieving $O(k)$ distortion. It is easy to see that the described algorithm will embed (in left-to-right order) first the left half of the long path, then all the k stars of size k, and finally the right half of the long path. This embedding yields distortion $O(k^2)$.

Chapter 3

A Dynamic Programming Algorithm for Graphs of Small Distortion

Given a connected simple graph $G = (V, E)$ and an integer c, we consider the problem of deciding whether there exists a non-contracting embedding of G into the integer line with maximum distortion at most c.

Note that the maximum distance between any two points in an optimal embedding can be at most $c(n-1)$, and there always exists an optimal embedding with all the nodes embedded into integer coordinates. W.l.o.g., in the rest of this section, we will only consider embeddings of the form $f : V \to \{0, 1, \ldots, c(n-1)\}$. Furthermore, if G admits an embedding of distortion c, then the maximum degree of G is at most $2c$. Thus, we may also assume that G has maximum degree $2c$.

3.1 Definitions

The algorithm that we will describe is based on dynamic programming. More specifically, the algorithm will compute a solution by augmenting carefully chosen partial solutions. We will first define formally the notion of a partial solution.

Definition 1 (Partial Embedding). *Let $V' \subseteq V$. A* partial embedding *on V' is a function $g : V' \to \{0, 1, \dots, c(n-1)\}$.*

Definition 2 (Feasible Partial Embedding). *Let f be a partial embedding on V'. f is called* feasible *if there exists an embedding g of distortion at most c, such that for each $v \in V'$, we have $g(v) = f(v)$, and for each $u \notin V'$, it is $g(u) > \max_{w \in V'} f(w)$.*

Definition 3 (Plausible Partial Embedding). *Let f be a partial embedding on V'. f is called* plausible *if*

- *For each $u, v \in V'$, we have $|f(u) - f(v)| \geq D(u, v)$.*
- *For each $u, v \in V'$, if $\{u, v\} \in E$, then $|f(u) - f(v)| \leq c$.*
- *Let $L = \max_{v \in V'} f(v)$. For each $u \in V'$, if $f(u) \leq L - c$, then for each $w \in V$ such that $\{u, w\} \in E$, we have $w \in V'$.*

3.2 Some Technical Properties

We will now give some useful properties of the feasible and plausible embeddings defined above.

Lemma 6. *If a partial embedding is feasible, then it is also plausible.*

Proof. Let f be a partial embedding over V', such that f is feasible, but not plausible, and let $L = \max_{v \in V'} f(v)$. It follows that there exists $\{u, w\} \in E$, with $u \in V'$, such that $f(u) \leq L - c$, and $w \notin V'$. Since f is feasible, there exists an embedding g of distortion at most c, satisfying $g(u) = f(u) \leq L - c$, and $g(w) > L$. Thus, $|g(u) - g(w)| > c$, a contradiction. □

Definition 4 (Active Region). *Let f be a partial embedding over V'. The* active region *of f is a couple (X, Y), where $X = \{(u_1, f(u_1)), \dots, (u_{|X|}, f(u_{|X|}))\}$ is a set of $\min\{2c+1, |V'|\}$ couples, where $\{u_1, \dots, u_{|X|}\}$ is a subset of V', such that $f(u_i) = \max_{u \in V' \setminus \{u_{i+1}, \dots, u_{|X|}\}} f(u)$, and Y is the set of all edges in E having exactly one endpoint in V'.*

Lemma 7. *Let f_1 be a plausible partial embedding over V_1, and f_2 be a plausible partial embedding over V_2. If f_1 and f_2 have the same active region, then*

- $V_1 = V_2$.
- *f_1 is feasible if and only if f_2 is feasible.*

Proof. Let $L = \max_{v \in V'} f(v)$. To prove that $V_1 \subseteq V_2$, assume that there exists $v \in V_1 \setminus V_2$. Let p be a path starting at v, and terminating at some node in $V_1 \cap V_2$, and let v'' be the first node in $V_1 \cap V_2$ visited by p, and v' be the node visited exactly before v''. Clearly, $v' \in V_1 \setminus V_2$, and v' is not in the active region, thus $f_1(v') < L - 2c$. Furthermore, by the definition of a plausible partial embedding, since the edge $\{v'', v'\}$ has exactly one endpoint in V_2, it follows that $f_2(v'') > L - c$. Thus, $|f_1(v') - f_1(v'')| = |f_1(v') - f_2(v'')| > c$, contradicting the fact that f_1 is plausible. Similarly we can show that $V_2 \subseteq V_1$, and thus $V_1 = V_2$.

Assume now that f_1 is feasible, thus there exists an embedding g_1 of distortion at most c, such that for each $v \in V_1$, we have $f_1(v) = g_1(v)$, and for each $v \notin V_1$, we have $g_1(v) > L$. Consider the embedding g_2, where $g_2(u) = f_2(u)$, if $u \in V_2$, and $g_2(u) = g_1(u)$ otherwise. It suffices to show that g_2 is non-contracting and has distortion at most c.

If g_2 has distortion more than c, then since f_2 is a plausible partial embedding, and g_1 has distortion at most c, it follows that there exists an edge $\{u, w\}$, with $u \in V_2$ and $w \notin V_2$, such that $|g_2(u) - g_2(w)| > c$. Since the edge $\{u, w\}$ has exactly one endpoint in V_2, it follows that $f_2(u) > L - c$, and thus u is in the active region, and $f_2(u) = f_1(u)$. Thus, we obtain that $|g_1(u) - g_1(w)| = |g_2(u) - g_2(w)| > c$, a contradiction. Thus, g_2 has distortion at most c, and f_2 is feasible. □

Lemma 8. *For fixed values of c, the number of all possible active regions of all the plausible partial embeddings is at most $O(n^{4c+2})$.*

Proof. Let f be a plausible partial embedding, with active region (X, Y), such that $|X| = i$. It is easy to see that every edge in Y has exactly one endpoint in X. Since the degree of every node is at most $2c$, after fixing X, the number of possible values for Y is at most 2^{2ic}. Also, the number of possible different values for X is at most $\binom{n}{i}(nc)^i$. Thus, the number of possible active regions for all plausible partial embeddings is at most $\sum_{i=1}^{2c+1}\binom{n}{i}(nc)^i 2^{2ic} = O(n^{4c+2})$. □

3.3 The Algorithm

Definition 5 (Successor of a Partial Embedding). *Let f_1 and f_2 be plausible partial embeddings on V_1 and V_2 respectively. f_2 is a* successor *of f_1 if and only if*

- $V_2 = V_1 \cup \{u\}$, *for some* $u \notin V_1$.
- *For each* $u \in V_1 \cap V_2$, *we have* $f_1(u) = f_2(u)$.
- *If* $u \in V_2$ *and* $u \notin V_1$, *then* $f_2(u) = \max_{v \in V_2} f_2(v)$.

Let P be the set of all plausible partial embeddings, and let $\hat{P}$ be the set of all active regions of the embeddings in P. Consider a directed graph H with $V(H) = \hat{P}$. For each $\hat{x}, \hat{y} \in V(H)$, $(\hat{x}, \hat{y}) \in E(H)$ if and only if there exist plausible embeddings x, y, such that $\hat{x}$ and $\hat{y}$ are the active regions of x and y respectively, and y is a successor of x.

Lemma 9. *Let x_0 be the active region of the empty partial embedding. G admits a non-contracting embedding of distortion at most c, if and only if there exists a directed path from x_0 to some node x in H, such that $x = (X, Y)$, with $X \neq \emptyset$ and $Y = \emptyset$.*

Proof. If there exists a path from x_0 to some node $x = (X, Y)$, with $X \neq \emptyset$ and $Y = \emptyset$, then since $X \neq \emptyset$, it follows that x is not the active region of the empty partial embedding. Furthermore, since G is connected and $Y = \emptyset$, it follows that x is the active region of a plausible embedding f of all the nodes of G. By the definition of a plausible embedding, it follows that f is a non-contracting embedding of G with distortion at most c.

If there exists a non-contracting embedding f of G, with distortion at most c, then we can construct a path in H, visiting nodes $y_0, y_1, \ldots, y_{|V|}$, as follows: For each i let f_i be the partial embedding obtained from f by considering only the i leftmost embedded nodes, and let y_i be the active region of f_i. Clearly, each f_i is a feasible embedding, and thus by Lemma 6, it is also plausible. Moreover, $y_0 = x_0$, and for each $0 < i \leq |V|$, it is easy to see that f_i is a successor of f_{i-1}, and thus $(y_{i-1}, y_i) \in E(H)$. Since, $f_{|V|}$ is an embedding of all the nodes of G, the active region $y_{|V|} = (X_{|V|}, Y_{|V|})$ satisfies $X_{|V|} \neq \emptyset$, and $Y_{|V|} = \emptyset$. □

Using Lemma 9, we can decide whether there exists an embedding of G as follows: We begin at node x_0, and we repeatedly traverse edges of H, without repeating nodes. Note that we do not compute the whole H from the beginning, but we instead compute only the neighbors of the current node. This is done as follows: At each step i, we maintain a plausible partial embedding g_i, such that each partial embedding induced by the j leftmost embedded nodes in g_i, has active region equal to the jth node in the path from x_0 to the current node. We consider all the plausible embeddings obtained by adding a rightmost node in g_i. The key property is that by Lemma 7, the active regions of these embeddings are exactly the neighbors of the current node. This is because an active region completely determines the subset of embedded nodes, as well as the feasibility of such a plausible embedding. By Lemma 8, the above procedure runs in polynomial time when c is fixed.

We have thus obtained the following Theorem.

Theorem 3. *For any fixed integer c, we can compute in polynomial time a non-contracting embedding of G, with distortion at most c, if one exists.*

Chapter 4

Hardness of Approximation

In this section we show that the problem of computing minimum distortion embedding of unweighted graphs is NP-hard to a-approximate for certain $a > 1$. This is done by a reduction from TSP over $(1, 2)$-metrics. Recall that the latter problem is NP-hard to approximate up to some constant $a > 1$.

4.1 The Reduction

Recall that a metric $M = (V, D)$ is a $(1, 2)$-metric, if for all $u, v \in V$, $u \neq v$, we have $D(u, v) \in \{1, 2\}$. Let $G(M)$ be a graph (V, E) where E contains all edges $\{u, v\}$ such that $D(u, v) = 1$.

The reduction F from the instances of TSP to the instances of the embedding problem is as follows. For a $(1, 2)$-metric M, we first compute $G = (V, E) = G(M)$. Then we construct a copy $G' = (V', E')$ of G, where V' is disjoint from V. Finally, we add a vertex o with an edge to all vertices in $V \cup V'$. In this way we obtain the graph $F(M)$.

The properties of the reduction are as follows.

Lemma 10. *If there is a tour in M of length t, then $F(M)$ can be embedded into a line with distortion at most t.*

Proof. The embedding $f : F(M) \to \mathbb{R}$ is constructed as follows. Let $v_1, \ldots, v_n, v_1$ be the sequence of vertices visited by a tour T of length t. The embedding f is obtained by placing the vertices V in the order induced by T, followed by the vertex o and then the vertices V'. Formally:

- $f(v_1) = 0$, $f(v_i) = f(v_{i-1}) + D(v_{i-1}, v_i)$ for $i > 1$
- $f(o) = f(v_n) + 1$
- $f(v'_1) = f(o) + 1$, $f(v'_i) = f(v'_{i-1}) + D(v'_{i-1}, v'_i)$ for $i > 1$

It is immediate that f is non-contracting. In addition, the maximum distortion (of at most t) is achieved by the edges $\{o, v_1\}$ and $\{o, v'_n\}$. □

Lemma 11. *If there is an embedding f of $F(M)$ into a line that has distortion c, then there is a tour in M of length at most $c + 1$.*

Proof. Let $H = F(M)$. Let $U = u_1 \dots u_{2n}$ be the sequence of the vertices of $V \cup V'$ in the order induced by f. Partition the range $\{1 \dots 2n\}$ into maximal intervals $\{i_0 \dots i_1 - 1\}, \{i_1 \dots i_2 - 1\}, \dots, \{i_{k-1} \dots i_k - 1\}$, such that for each interval I, the set $\{u_i : i \in I\}$ is either entirely contained in V, or entirely contained in V'. Recall that H has diameter 2. Since f has distortion c, it follows that $|f(u_1) - f(u_{2n})| \leq 2c$. Moreover, from non-contraction of f it follows that $|f(u_{i_j - 1}) - f(u_{i_j})| = 2$ for all j. It follows that if we swap any two subsequences of U corresponding to different intervals I and I', then the resulting mapping of $V \cup V'$ into $\mathbb{R}$ is still non-contracting (with respect to the metric induced by H). Therefore, there exists a mapping f' of $V \cup V'$ into $\mathbb{R}$ which is non-contracting, in which all vertices of V precede all vertices of V', and such that the diameter of the set $f'(V \cup V')$ is at most $2c$. Without loss of generality, assume that the diameter δ of $f'(V)$ is not greater than the diameter of $f'(V')$. This implies that $\delta \leq (2c-2)/2 = c-1$. Therefore, the ordering of the vertices in V induced by f' corresponds to a tour in M of length at most $\delta + 2 \leq c + 1$. □

By combining Lemmata 10 and 11 we obtain the following result.

Corollary 1. *There exists a constant $a > 1$ such that a-approximating the minimum distortion embedding of an unweighted graph is NP-hard.*

Chapter 5

Embedding Spheres Into the Plane

In this chapter we consider the following embedding problem. We are given a set of n points X on a unit sphere in $\mathbb{R}^3$, and we want to embed X into the two-dimensional Euclidean plane. There are two type of questions that we will study in the context of this problem. First, we will show that there exists an embedding of the metric induced by X, with distortion $O(\sqrt{n})$. We will also show that this bound is tight, by giving sets X such that any embedding has distortion $\Omega(\sqrt{n})$.

Next, we will show that using the same techniques we can obtain a $O(1)$-approximation algorithm for the corresponding optimization problem.

5.1 Worst-case Upper Bound

Let $M = (X, D)$ be a metric induced by a set X of n points on a unit sphere S^2, under the Euclidean distance in $\mathbb{R}^3$. Let c denote the minimum distortion of any embedding of M into the two-dimensional Euclidean plane.

Theorem 4. *If $M = (X, D)$ is the metric induced by a set X of n points on a unit sphere S^2, under the Euclidean distance in $\mathbb{R}^3$, then $c = O(\sqrt{n})$.*

Proof. Since the size of the surface of S^2 is constant, it follows that there exists a cap K in S^2, of size $\Omega(1/n)$, such that $X \cap K = \emptyset$. Let p_0 be the center of K on S^2, and p_0' be its antipode. By rotating S^2, we may assume that $p_0 = (0, 0, 1)$, and thus

$p'_0 = (0,0,-1)$.

For points $p, p' \in S^2$, let $\rho_S(p,p')$ be the geodesic distance between p and p' in S^2. Consider the mapping $f : X \to \mathbb{R}^2$, such that for every point $p \in X$, with $p = (x,y,z)$, we have

$$f(p) \quad = \quad \begin{cases} \left(\rho_S(p,p'_0)\frac{x}{\sqrt{x^2+y^2}}\rho_S(p,p'_0)\frac{y}{\sqrt{x^2+y^2}}\right) & \text{if } p \neq p', \\ (0,0) & \text{if } p = p' \end{cases}$$

It is straightforward to verify that f is non-contracting.

Claim 1. *The expansion of f is maximized for points p, q, on the perimeter of K, which are antipodals with respect to K.*

Proof. Let $p, q \in S^2$. W.l.o.g., we assume that $p = (0, \sin\phi_p, 1+\cos\phi_p)$, and $q = (\sin\phi_q \sin\theta_q, \sin\phi_q\cos\theta_q, 1+\cos\phi_q)$, for some $0 \leq \phi_p, \phi_q \leq \phi$, and $0 \leq \theta_q \leq \pi$. The images of p and q are $f(p) = (0, \phi_p)$, and $f(q) = (\phi_q \sin\theta_q, \phi_q\cos\theta_q)$, respectively. Let $h = \frac{\|f(p)-f(q)\|}{\|p-q\|}$, be the expansion of f in the pair p, q. We obtain:

$$h^2 \quad = \quad \frac{\phi_q^2 + \phi_p^2 - 2\phi_q\phi_p\cos\theta_q}{2 - 2\cos\phi_p\cos\phi_q - 2\sin\phi_p\sin\phi_q\cos\theta_q}$$

Observe that since $\sin\phi_p \leq \phi_p$, and $\sin\phi_q \leq \phi_q$, it follows that h^2 is maximized when $\cos\theta_q$ is minimized. That is, the expansion is maximized for $\theta_q = \pi$.

Thus, we can assume that the expansion of f is maximized for points $p, q \in S^2$, with $p = (0, \sin\phi_p, 1+\cos\phi_p)$, and $q = (0, -\sin\phi_q, 1+\cos\phi_q)$. For such points, the expansion is $\frac{\phi_p+\phi_q}{2\sin\frac{\phi_p+\phi_q}{2}}$. It follows that the expansion is maximized when $\phi_p+\phi_q$ is maximized, which happens when p and q are on the perimeter of K. □

We pick p and q on the perimeter of K, such that p is the antipode of q w.r.to K. Let ϕ_K be the angle of K, and set $r_K = \phi_K/2$. We have $r_K = \Omega(1/\sqrt{n})$, and $\|f(p)-f(q)\| = 2\pi - 2r_K$, while $\|p-q\| = 2\sin r_K$. Thus, the expansion is at most $\frac{\pi - r_K}{\sin r_K}$. W.l.o.g., we can assume that $r_K \leq \pi/2$, since otherwise we can simply consider a smaller cap K. Thus, $\frac{\pi-r_K}{\sin r_K} \leq 2\frac{\pi - r_K}{\pi r_K} < \frac{2}{r_K} = O(\sqrt{n})$. Since the embedding is non-contracting, it follows that the expansion is $O(\sqrt{n})$. □

5.2 Worst-Case Lower Bound

We will now show that the upper bound given above is optimal within a constant factor.

Theorem 5. *There exists a metric $M = (X, D)$, induced by a set X of n points on a unit sphere S^2, under the Euclidean distance in $\mathbb{R}^3$, such that any mapping $f : X \to \mathbb{R}^2$ has distortion $\Omega(\sqrt{n})$.*

Proof. Let $X \subset S^2$ be a set of n points, such that X is a $O(1/\sqrt{n})$-net of S^2, and let $f : X \to \mathbb{R}^2$ be a non-expanding embedding. Since $S^2 \subset \mathbb{R}^3$, by Kirszbraun's Theorem ([14], see also [17]), we obtain that f can be extended to a non-expanding mapping $f' : S^2 \to \mathbb{R}^2$. Also, by the Borsuk-Ulam Theorem, it follows that there exist antipodals $p, q \in S^2$, such that $f'(p) = f'(q)$. Since X is an $O(1/\sqrt{n})$-net, there exist points $p', q' \in X$, such that $\|p - p'\| = O(1/\sqrt{n})$, and $\|q - q'\| = O(1/\sqrt{n})$. Since f is non-expanding, it follows that $\|f(p') - f(q')\| = O(1/\sqrt{n})$. On the other hand, we have $\|p - q\| = 2$, and thus $\|p' - q'\| = \Omega(1)$. Thus, f has distortion $\Omega(\sqrt{n})$. □

5.3 A $O(1)$-Approximation Algorithm

We are now ready to combine the techniques of the upper and lower bounds, to obtain an approximation algorithm for the optimization version of the problem.

Theorem 6. *There exists a polynomial-time,* 3.512*-approximation algorithm, for the problem of embedding a finite sub-metric of S^2 into $\mathbb{R}^2$.*

Proof. Let S^2 be a unit sphere in $\mathbb{R}^3$. Let $X \subset S^2$ be a set of n points, and let $M = (X, D)$ be the corresponding metric, under the Euclidean distance in $\mathbb{R}^3$. Initially, we compute the largest cap K of S^2, such that $K \cap X = \emptyset$. Let ϕ_K be the angle of K, and $r_K = \phi_K/2$. Let also p_0 be the center of K, and p'_0 be its antipode, w.r.to S^2. By rotating S^2, we may assume that $p_0 = (0, 0, 1)$, and thus $p'_0 = (0, 0, -1)$. Similarly to the proof of Theorem 4, we compute the mapping $f : X \longrightarrow \mathbb{R}^2$, such that for every point $p \in X$, with $p = (x, y, z)$, it is

$$f(p) \quad = \quad \begin{cases} \left(\rho_S(p, p_0') \frac{x}{\sqrt{x^2+y^2}}, \rho_S(p, p_0') \frac{y}{\sqrt{x^2+y^2}}\right) & \text{if } p \neq p' \\ (0, 0) & \text{if } p = p' \end{cases}$$

By Claim 1, we have that the expansion of f is maximized for points which are on the perimeter of K. We pick $p, p' \in S^2$, which are on the perimeter of K, and are antipodals w.r.to K. In this case, it is $\|f(p) - f(p')\| = 2\pi - 2r_K$, and $\|p - p'\| = 2 \sin r_K$. Since f is non-contracting, it follows that the expansion is at most $\frac{\pi - r_K}{\sin r_K}$.

It remains to show that this embedding is optimal within a constant factor. Let g be a non-expanding embedding $X \longrightarrow \mathbb{R}^2$. Since $S^2 \subset \mathbb{R}^3$, by Kirszbraun's Theorem, we obtain that g can be extended to a non-expanding mapping $g' : S^2 \longrightarrow \mathbb{R}^2$. Also, by the Borsuk-Ulam Theorem, it follows that there exist antipodals $p, q \in S^2$, such that $g'(p) = g'(q)$. Since K is the largest cap with $K \cap X = \emptyset$, it follows that there exist points $p', q' \in X$, such that $\rho_S(p, p') \leq r_K$, and $\rho_S(q, q') \leq r_K$. Since g is non-expanding, we have

$$\begin{aligned} \|g(p') - g(q')\| &= \|g'(p') - g'(q')\| \\ &\leq \|g'(p') - g'(p)\| + \|g'(q') - g'(q)\| \\ &\leq \|p - p'\| + \|q - q'\| \\ &\leq 4 \sin \frac{r_K}{2}. \end{aligned}$$

On the other hand, it is $\|p - q\| = 2$, and thus $\|p' - q'\| \geq 2 \cos r_K$. Thus, g has distortion at least $\max\{\frac{\cos r_K}{2 \sin \frac{r_K}{2}}, 1\}$. By combining the above, we obtain that the approximation ratio of the algorithm is at most $\left(\frac{\pi - r_K}{\sin r_K}\right) / \max\{\frac{\cos r_K}{2 \sin \frac{r_K}{2}}, 1\}$. This value is maximized for $r_K = 2 \tan^{-1} \frac{(\sqrt{3}-1)3^{3/4}\sqrt{2}}{6} \approx 0.749$, for which we obtain that the approximation ratio is less than 3.512. □

Bibliography

[1] R. Agarwala, V. Bafna, M. Farach-Colton, B. Narayanan, M. Paterson, and M. Thorup. On the approximability of numerical taxonomy: (fitting distances by tree metrics). *7th Symposium on Discrete Algorithms*, 1996.

[2] M. Badoiu, P. Indyk, and Y. Rabinovich. Approximate algorithms for embedding metrics into low-dimensional spaces. *Manuscript*, 2003.

[3] K. Borsuk. Drei Sätze über die n-dimensionale euklidische Sphäre. *Fund. Math.*, 20:177–190, 1933.

[4] J. Bourgain. On lipschitz embedding of finite metric spaces into hilbert space. *Isreal Journal of Mathematics*, 52:46–52, 1985.

[5] M Bădoiu. Approximation algorithm for embedding metrics into a two-dimensional space. *14th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2003.

[6] Kedar Dhamdhere. Approximating additive distortion of embeddings into line metrics. *7th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX)*, 2004.

[7] M. Farach-Colton, S. Kannan, and T. Warnow. A robust model for finding optimal evolutionary tree. *Proceedings of the Symposium on Theory of Computing*, 1993.

[8] U. Feige. Approximating the bandwidth via volume respecting embeddings. *Journal of Computer and System Sciences*, 60(3):510–539, 2000.

[9] Anupam Gupta. Improved bandwidth approximation for trees. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2000.

[10] J. Hastad, L. Ivansson, and J. Lagergren. Fitting points on the real line and its application to rh mapping. *Lecture Notes in Computer Science*, 1461:465–467, 1998.

[11] P. Indyk. Tutorial: Algorithmic applications of low-distortion geometric embeddings. *Proceedings of the Symposium on Foundations of Computer Science*, 2001.

[12] L. Ivansson. Computational aspects of radiation hybrid. *Doctoral Dissertation, Department of Numerical Analysis and Computer Science, Royal Institute of Technology*, 2000.

[13] C. Kenyon, Y. Rabani, and A. Sinclair. Low distortion maps between point sets. *Proceedings of the Symposium on Theory of Computing*, 2004. to appear.

[14] M. D. Kirszbraun. Über die zusammenziehenden und lipschitzschen Transformationen. *Fund. Math.*, 22:77–108, 1934.

[15] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[16] J.B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, 1964.

[17] J. R. Lee and A. Naor. Absolute lipschitz extendability. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 338:859–862, 2004.

[18] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Proceedings of 35th Annual IEEE Symposium on Foundations of Computer Science*, pages 577–591, 1994.

[19] J. Matoušek. Bi-lipschitz embeddings into low-dimensional euclidean spaces. *Comment. Math. Univ. Carolinae*, 31:589–600, 1990.

[20] J. Matoušek. On the distortion required for embedding finite metric spaces into normed spaces. *Israel Journal of Mathematics*, 93:333–344, 1996.

[21] J. B. Saxe. Dynamic-programming algorithms for recognizing small-bandwidth graphs in polynomial time. *SIAM J. Algebraic Discrete Methods*, 1:363–369, 1980.

[22] R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function 1. *Psychometrika*, 27:125–140, 1962.

[23] R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function 2. *Psychometrika*, 27:216–246, 1962.

[24] Working Group on Algorithms for Multidimensional Scaling. Algorithms for multidimensional scaling. http://dimacs.rutgers.edu/Workshops/Algorithms/.