Approximating the Minimum Weight Spanning Tree of a Set of Points in the Hausdorff Metric

Victor Alvarez* Raimund Seidel†

March 31, 2009

Abstract

We study the problem of approximating MST(P), the minimum weight spanning tree of a set P of n points in $[0,1]^d$, by a spanning tree of some subset $Q \subset P$. We show that if the *weight* of MST(P) is to be approximated, then in general Q must be large. If the *shape* of MST(P) is to be approximated, then this is always possible with a small Q.

More specifically, for any $0 < \varepsilon < 1$ we prove:

(i) There are sets $P \subset [0,1]^d$ of arbitrarily large size n with the property that any subset $Q' \subset P$ that admits a spanning tree T' with $\big||T'| - |\mathrm{MST}(P)|\big| < \varepsilon \cdot |\mathrm{MST}(P)|$ must have size at least $\Omega(n^{1-1/d})$. (Here $|T|$ denotes the weight, i.e. the sum of the edge lengths of tree T.)

(ii) For any $P \subset [0,1]^d$ of size n there exists a subset $Q \subseteq P$ of size $O(1/\varepsilon^d)$ that admits a spanning tree T that is ε-close to MST(P) in terms of Hausdorff distance (which measures shape dissimilarity).

(iii) This set Q and this spanning tree T can be computed in time $O(\tau_{d,p}(n) + 1/\varepsilon^d \log(1/\varepsilon^d))$ for any fixed dimension d. Here $\tau_{d,p}(n)$ denotes the time necessary to compute the minimum weight spanning tree of n points in $\mathbb{R}^d$ under any fixed metric L_p, $1 \leq p \leq \infty$, where $\tau_{2,p}(n) = O(n \log n)$, see [9], $\tau_{3,2}(n) = O((n \log n)^{4/3})$, and $\tau_{d,2}(n) = O(n^{2-2/(\lceil d/2 \rceil+1)+\phi})$, with $\phi > 0$ arbitrarily small, for $d > 3$, see [1]. Also $\tau_{3,1}(n)$ and $\tau_{3,\infty}(n)$ is known to be $O(n \log n)$, see [6].

1 Introduction

The approximation of geometric problems by means of reducing the size of the input has been the subject of study of many researchers. The idea is the fast identification of the part of the input that matters for the problem at hand and the use of this extracted data to speed up the computations.

In [2], Agarwal *et al.* developed a framework, called Coresets, to approximate extent measures of a given set of points P in any fixed dimension d. Such extent measures include the diameter, the

*Fachrichtung Informatik, Universität des Saarlandes, alvarez@cs.uni-sb.de
†Fachrichtung Informatik, Universität des Saarlandes, rseidel@cs.uni-sb.de

width, the radius of the minimum enclosing cylinder, etc. Their idea is basically the computation of a subset P' of P whose size depends exclusively on ε and d and, whose convex hull approximates the convex hull of P. Then, use this new convex hull for further computations and argue that this produces good approximations for the desired extent measures.

In this paper we are interested in approximating the minimum weight spanning tree of a set $P \subset \mathbb{R}^d$ of points, but not in the sense of, say, Clarkson [5], who wants to quickly find some spanning tree of P whose weight is close to that of $\mathrm{MST}(P)$. We are instead interested in finding a spanning tree of a *small subset* of P that in some sense approximates $\mathrm{MST}(P)$. We will show that the core set approach outlined above cannot work in this context if the approximation measure is the weight of the trees. However, if we want to approximate $\mathrm{MST}(P)$ in a more topological (or shape) sense, then this is indeed possible using a spanning tree of a subset of P whose size depends exclusively on ε, the approximation parameter, and on d.

Throughout the paper let $0 < \varepsilon < 1$ be a fixed constant. The dimension d is meant to be fixed. Also, let L_p, $1 \leq p \leq \infty$, be a chosen but fixed metric.

2 MST(P) admits no constant size subset approximation with respect to weight

The goal of this section is to prove the following result:

Theorem 1 *For each $n = k^d$ with $k \in \mathbb{N}$ there exists a set $P \subset [0,1]^d$ of n points such that any subset Q' of P that admits a spanning tree T' with $|T'| \geq (1-\varepsilon)|MST(P)|$ must have size at least $\Omega(k^{(d-1)})$.*

Note that this theorem clearly implies Claim (i) of the abstract.

Proof. Let $n = k^d$ with $k \in \mathbb{N}$ and let $\mathbb{G}^d$ be the d-dimensional grid over $[0,1]^d$ of cell size $\delta = 1/(k-1)$. Let P be the set consisting of the grid points of $\mathbb{G}^d$. It is clear that $|P| = n$.

Any minimum weight spanning tree of such a set P only contains grid edges. Thus $|\mathrm{MST}(P)| = (n-1) \cdot \delta = (n-1)/(k-1) > n/k$.

Now let T' be a spanning tree of some $Q' \subset P$ such that $|T'| \geq (1-\varepsilon)|\mathrm{MST}(P)|$. Under the chosen L_p metric, every edge inside the unit cube $[0,1]^d$ has length at most $\sqrt[p]{d}$. Hence $|T'| < |Q'| \cdot \sqrt[p]{d}$. Combining this last inequality with the ones above we have

$$|Q'| \cdot \sqrt[p]{d} > |T'| \geq (1-\varepsilon)|\mathrm{MST}(P)| > (1-\varepsilon)\frac{n}{k}\,.$$

Since $n = k^d$ and ε and d are constant, the result follows.

■

3 The Hausdorff Metric

The Hausdorff metric allows to define distances between subsets of a metric space. In our case the metric space is $\mathbb{R}^d$ with the chosen L_p metric.

Definition 1 (Hausdorff distance) *The* Hausdorff distance $H(A, B)$ *between two non-empty subsets* A, B *of* $\mathbb{R}^d$ *is defined to be the radius of the largest open ball centered in one set and not meeting the other set.*

We say that A *and* B *are* ε-close *iff* $H(A, B) \leq \varepsilon$.

It is well known that the Hausdorff distance constitutes a metric on the space of all non-empty compact subsets of $\mathbb{R}^d$. Moreover, in a way it expresses the shape similarity, or rather dissimilarity between sets: $H(A, B) = 0$ means A and B must be the same, i.e. they are not at all dissimilar, and A and B are ε-close means that they are only ε-dissimilar in the sense that for any point in one set within distance ε there must be a point of the other set. Many computational geometry papers have used the Hausdorff distance as a measure of similarity/dissimilarity between subsets of $\mathbb{R}^d$, see e.g. [3]. We will use the Hausdorff distance to measure similarity/dissimilarity between spanning trees of finite sets embedded in $\mathbb{R}^d$, where such a tree is considered a subset of $\mathbb{R}^d$, namely the union of the segments formed by its edges.

It will turn out that if instead of closeness in weight we consider closeness in Hausdorff distance the minimum weight spanning tree of any finite $P \subset \mathbb{R}^d$ admits a good approximation by a spanning tree of a constant sized subset of P.

4 Approximating MST(P) by shape

At first a few graph theoretic preliminaries.

Let G be a complete undirected graph with vertex set P and with weighted edges. For the sake of exposition we assume that all edge weights are distinct, and thus the shortest edge of any cut of G and also the minimum weight spanning tree MST(P) are unique. This assumption can be justified using a standard perturbation argument. Let $\overline{P} = \langle P_1, \ldots, P_k \rangle$ be a partition of P into $k \geq 2$ non-empty "clusters," and let $\overline{G}$ be the graph obtained from G by contracting each cluster in $\overline{P}$ into a single node. $\overline{G}$ has parallel edges and self-loops, still, its minimum weight spanning tree MST($\overline{P}$) is unique. Consider the forest on P formed by the $k-1$ edges of G that induce the edges of MST($\overline{P}$). Let us call this forest the *minimum weight cluster forest of* P *with respect to* $\overline{P}$, for short MCF($P, \overline{P}$).

What is the relationship between the edges in MCF($P, \overline{P}$) and MST(P)?

Lemma 2 *Every edge in MCF*($P, \overline{P}$) *also is an edge of MST*(P).

Proof. Let e be an edge of MCF$(P, \overline{P})$ and let $\overline{e}$ be the corresponding edge of MST$(\overline{P})$. The removal of $\overline{e}$ from MST$(\overline{P})$ results in two subtrees producing a partition of the node set $\overline{P}$ into two sets $\overline{R}$ and $\overline{S}$. The edge $\overline{e}$ must be the shortest edge between nodes (i.e. clusters) in $\overline{R}$ and in $\overline{S}$ and hence e must be the shortest edge between (original) vertices in $R = \bigcup \overline{R}$ and in $S = \bigcup \overline{S}$. Since R and S form a partition of P this means that e must be an edge of MST(P). ■

Let us call an edge of G *long* (with respect to $\overline{P}$) iff it is longer than any edge connecting two vertices in the same cluster of $\overline{P}$.

Lemma 3 *Every long edge of MST(P) is also an edge of MCF$(P, \overline{P})$.*

Proof. Let e be a long edge in MST(P). Similar to the previous proof the edge e induces a partition of P into R and S, and e is the shortest edge connecting vertices in R with vertices in S. No cluster of $\overline{P}$ can have a vertex both in R and in S, since such two vertices would be connected by an edge shorter than the long edge e, a contradiction to e being the shortest edge between R and S. Thus R and S induce a partition of the cluster set $\overline{P}$ into $\overline{R}$ and $\overline{S}$, and $\overline{e}$ (induced by e) is the shortest edge connecting a cluster in $\overline{R}$ with a cluster in $\overline{S}$. Thus $\overline{e}$ is an edge of MST$(\overline{P})$ and therefore e is an edge of MCF$(P, \overline{P})$. ■

In the following P will a set of points in $\mathbb{R}^d$. Recall that the distance between two points $x, y \in \mathbb{R}^d$ in the L_p metric is given by $L_p(x, y) = (\sum_{i=1}^{d} |x_i - y_i|^p)^{1/p}$ with $L_\infty(x, y) = \max(|x_i - y_i|),\ 1 \leq i \leq d$. Note that given a cube of dimension d and of side length δ, the largest distance between two points inside the cube under the L_p metric is $\leq \delta \sqrt[p]{d}$.

We are now able to present the main result of this section which will prove Claim (ii) of the abstract.

Theorem 4 *Let P be a set of points in $[0, 1]^d$ and let $0 < \varepsilon < 1$ be a given parameter. It is possible to find a spanning tree T of some subset Q of P such that MST(P) and T are ε-close and $|Q| = O(1/\varepsilon^d)$.*

Proof. We will start by imposing a d-dimensional grid $\mathbb{G}^d$ of cell size $\delta = \frac{2\varepsilon}{3\sqrt[p]{d}}$ over P, where p is the chosen metric. The grid $\mathbb{G}^d$ induces a partition $\overline{P}$ of P into $k = O(1/\varepsilon^d)$ clusters, with each cluster begin composed of the set of points contained in a cell of $\mathbb{G}^d$. See Figure 1. Note that by the observation right above, two points in the same cluster are at most $2\varepsilon/3$ apart.

The claimed set Q will be the points in P incident to the edges of the minimum weight cluster forest MCF$(P, \overline{P})$. Since there are $k - 1$ edges in MCF$(P, \overline{P})$ it follows that $|Q| = O(1/\varepsilon^d)$.

The claimed spanning tree T of Q will contain all edges in MCF$(P, \overline{P})$ and in addition for each cluster C in $\overline{P}$ an arbitrary spanning tree of the points of Q in C. See Figure 1.

We claim that T and MST(P) are ε-close.

We need to prove that for every point on T there is a point on MST(P) within distance at most ε, and vice versa.

Let e be an edge of T. If e is an edge of $\mathrm{MCF}(P,\overline{P})$, then by Lemma 2 it is also an edge of $\mathrm{MST}(P)$ and thus every point x on e is within distance $0 < \varepsilon$ of some point of $\mathrm{MST}(P)$. If e is an edge connecting two points of the same cluster, then its length is at most $2\varepsilon/3$. Thus any point x on e is at most at distance $\varepsilon/3 < \varepsilon$ from one of e's endpoints, which are both in $\mathrm{MST}(P)$.

Now let e be an edge of $\mathrm{MST}(P)$. If it has length bigger than $2\varepsilon/3$, then it is long in the sense of Lemma 3, and therefore it is contained in $\mathrm{MCF}(P,\overline{P})$ and hence also in T. Thus every point x on e is within distance $0 < \varepsilon$ of some point of T. If e has length less than $2\varepsilon/3$, then every point x is within distance $\varepsilon/3$ of an endpoint v of e. Let q some point of Q in the cluster containing v. The distance between v and q is at most $2\varepsilon/3$, and thus by the triangle inequality the distance between x and q (which lies on T) is at most ε. ■

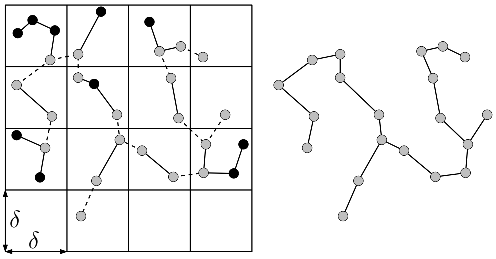

Figure 1: The points chosen to form Q are highlighted in light gray. The dashed edges connect points in different clusters of P induced by $\mathbb{G}^2$

This result says that it is possible to find a constant-size subset Q of P along with a spanning tree T of Q such that shape-wise T and $\mathrm{MST}(P)$ look essentially the same. This gives a method to sort of "compress" $\mathrm{MST}(P)$ to a tree that is close in shape but has constant size. Note, however, that one cannot conclude anything from T about the total weight $|\mathrm{MST}(P)|$.

5 Computing T

The only computationally non-trivial step in computing Q and T is the determination of the edges of the cluster forest $\mathrm{MCF}(P,\overline{P})$. The straigtforward way of computing these edges, forming the cluster graph $\overline{G}$ and computing its minimum weight spanning tree $\mathrm{MST}(\overline{P})$, leads to an $\Theta(n^2)$ time algorithm in the worst case, since $\overline{G}$ can have $\Theta(n^2)$ non-loop edges. (We assume here ε and d to be fixed.)

Lemma 2 implies that for computing $\mathrm{MST}(\overline{P})$ it suffices to consider only those edges that are induced by edges of $\mathrm{MST}(P)$. This suggests the following algorithm: Compute $B = \mathrm{MST}(P)$, for each grid imposed cluster contract the edges of B within the cluster to produce a contracted graph $\overline{B}$. Compute the minimum weight spanning tree of $\overline{B}$, which by Lemma 2 is the same as $\mathrm{MST}(\overline{P})$.

If $\tau_{d,p}(n)$ denotes the time necessary to compute the minimum weight spanning tree of n points in $\mathbb{R}^d$ and under the chosen L_p metric, then the time necessary for the outlined method is $\tau_{d,p}(n)$ for computing $\mathrm{MST}(P)$, plus $O(n)$ for computing $\overline{B}$ and $O(n + N\log N)$ for computing the minimum weight spanning tree of $\overline{B}$, where $N = O(1/\varepsilon^d)$ is the number of occupied grid cells, which is constant if ε and d are considered to be constant. The total time for the whole method is then dominated by $\tau_{d,p}(n)$. The following values for $\tau_{d,p}(n)$ are known:

- $\tau_{2,p}(n) = O(n \log n)$, see [9].
- $\tau_{3,2}(n) = O((n \log n)^{4/3})$, see [1].
- $\tau_{d,2}(n) = O(n^{2-2/(\lceil d/2 \rceil+1)+\phi})$, with $\phi > 0$ arbitrarily small, for $d > 3$, see [1].
- $\tau_{3,1}(n)$ and $\tau_{3,\infty}(n)$ are known to be $O(n \log n)$, see [6].

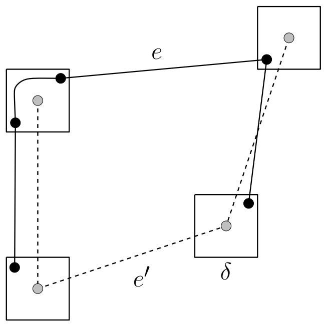

Figure 2: P is the set of black points. Q is the set of gray points. Any point on the edge e of MST(P) (solid edges) is farther than ε to any point on the edge e' of T (dashed edges). Therefore both trees are not ε-close.

Other methods suggest themselves, but they are either incorrect or do not seem to lead to better time bounds. For instance, we could choose a small sample set of points from each occupied cluster and compute the minimum weight spanning tree of the union of these sample sets. However, the tree produced this way may be very different in shape from MST($\overline{P}$) and will not lead to a tree that is ε-close to MST(P), see Figure 2. Or, we could run a minimum weight spanning tree algorithm on the clusters (without forming $\overline{G}$ or some subgraph explicitly) by repeatedly solving so-called bi-chromatic closest pair problems. However, this is unlikely to produce a better running time, since the complexity of solving a bi-chromatic closest pair problem on n points in $\mathbb{R}^d$ under *any* L_p metric is known to be $\Theta(\tau_{d,p}(n))$, see [6].

Finding a faster algorithm for computing a constant sized tree that is ε-close to MST(P) looks like a challenging problem.

6 Conclusion and Open Problems

We have shown that in general it is not possible to approximate well the weight of the minimum weight spanning tree of a set of points P in $\mathbb{R}^d$ with a subset of size independent of the size of P. However, changing the notion of approximation, we have shown, that it is possible to compute a spanning tree T of some small subset $Q \subseteq P$ such that the Hausdorff distance between T and the minimum weight spanning tree of P is small, which means that the two trees are very similar in shape. This provides a potential way of compressing $\mathrm{MST}(P)$ in a meaningful and interesting way.

Finally we briefly want to discuss the question whether the results of this paper are restricted to minimum weight spanning trees, or whether they allow generalizations to other structures on point sets, for instance on the Euclidean Travelling Salesperson Tour.

So the question at hand is the following, let us call it the *TSP approximation problem*: Given a set P of n points in $[0,1]^d$, some $\varepsilon > 0$, and some function $f : \mathbb{R}^2 \to \mathbb{R}$, compute a set $Q \subset P$ with $|Q| \leq f(\varepsilon, d)$ along with a cycle C through Q that is "ε-close" to $\mathrm{TSP}(P)$, the shortest Hamiltonian circuit through P. For the case that "ε-close" is to be interpreted as closeness in Euclidean length it is easy to see that in general this TSP approximation problem does in general not admit a solution: the proof of Theorem 1 carries over in an almost verbatim fashion. Similarly, for the case the "ε-close" is to be interpreted in the Hausdorff sense there is a positive answer to this TSP approximation problem, as long as $f(\varepsilon, d) \in \Omega(1/\varepsilon^d)$: The methods used to prove Theorem 4, appropriately adjusted, suffice. However, this positive answer is algorithmically quite unsatisfying, since in the methods of the proof of Theorem 4 computing the approximation requires the knowledge of $\mathrm{TSP}(P)$, and of course computing $\mathrm{TSP}(P)$ is NP-hard. (Replacing $\mathrm{TSP}(P)$ by a length approximation, which by the results of Arora [4] and Mitchell [7] can be computed in polynomial time, seems like a cheat.) So the question remains whether the TSP approximation problem (in the Hausdorff sense) admits polynomial time solution (in n, ε, d). We have not succeeded in finding such a solution and we also have been unable to prove that the problem is hard. We suspect that it is actually hard and conjecture the decision version of the TSP approximation problem (in the Hausdorff sense) to be NP-hard.

Acknowledgements

The first author was supported by IMPRS-CS, the International Max-Planck Reseach School for Computer Science, and also by the Research Training Group 623 "Performance Guarantees for Computer Systems" funded by the Deutsche Forschungsgemeinschaft (DFG). He would also like to acknowledge support by CONACYT-DAAD of México.

References

[1] Pankaj K. Agarwal, Herbert Edelsbrunner, Otfried Schwarzkopf, and Emo Welzl. *Euclidean Minimum Spanning Trees and Bichromatic Closest Pairs.* Discrete Comput. Geom. 6(5):407–422, 1991.

[2] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. *Approximating Extent Measures of Points.* J. ACM, 51(4):606–635, 2004.

[3] Helmut Alt and Leonidas J. Guibas. *Discrete Geometric Shapes: Matching, Interpolation, and Approximation* In J.R. Sack and J. Urrutia, editors, **Handbook of Computational Geometry.** Elsevier Science Publishers B.V. North-Holland, Amsterdam, pp. 121–153, 2000.

[4] Sanjeev Arora. *Polynomial Time Approximation Schemes for Euclidean Traveling Salesman and other Geometric Problems* J. ACM, 45(5):753–782, 1998.

[5] Kenneth L. Clarkson. *Fast Expected-time and Approximation Algorithms for Geometric Minimum Spanning Trees.* In STOC '84: Proceedings of the 16th Annual ACM Symposium on Theory of Computing, 342-3-48, 1984.

[6] Drago Krznaric, Christos Levcopoulos, and Bengt J. Nilsson. *Minimum Spanning Trees in d Dimensions.* Nordic J. of Computing, 6(4):446–461, 1999.

[7] Joseph S. B. Mitchell. *Guillotine Subdivisions Approximate Polygonal Subdivisions: A Simple Polynomial-Time Approximation Scheme for Geometric TSP, k-MST, and Related Problems* SIAM J. Comput., 28(4):1298–1309, 1999.

[8] Andrew C. Yao. *On Constructing Minimum Spanning Trees in k-dimensional Spaces and Related Problems.* SIAM Journal on Computing, 11(4):721–736, 1982.

[9] D. T. Lee. *Two-Dimensional Voronoi Diagrams in the Lp-Metric.* Journal of the ACM, 27(4):604–618, 1980.