When Hamming Meets Euclid: The Approximability of Geometric TSP and Steiner Tree[*]

Luca Trevisan[†]

Abstract

We prove that the Traveling Salesman Problem (Min TSP) is Max SNP-hard (and thus NP-hard to approximate within some constant $r > 1$) even if all cities lie in a Euclidean space of dimension $\log n$ (n is the number of cities) and distances are computed with respect to any l_p norm. The running time of recent approximation schemes for geometric Min TSP is doubly exponential in the number of dimensions. Our result implies that this dependence is necessary unless NP has sub-exponential algorithms.

We prove a similar, but weaker, inapproximability result for the Steiner Minimal Tree Problem (Min ST). We also prove, as an intermediate step, the hardness of approximating Min TSP in Hamming spaces.

The reduction for Min TSP uses error-correcting codes and random sampling; the reduction for Min ST uses the integrality property of MIN-CUT linear programming relaxations. The only previous inapproximability results for metric Min TSP involved metrics where all distances are 1 or 2.

1 Introduction

Given a metric space and a set U of points into it, the Metric Traveling Salesman Problem (Min TSP) is to find a closed tour of shortest total length visiting each point exactly once, while the Metric Steiner Minimum Tree Problem (Min ST) is to find the minimum length tree connecting all the points of U; the tree can possibly contain points not in U, that are called "Steiner points".

Both problems are among the most classical and most widely studied ones in Combinatorial Optimization, Operations Research and Computer Science during the past few decades, and before. Important special cases arise when the metric space is $\mathcal{R}^k$ and the distance is computed according to the ℓ_1 norm (the *rectilinear* case) or the ℓ_2 norm (the *Euclidean* case).

We establish the first non-approximability results for this class of problems. As an intermediate step, we use the fact that they are also hard to approximate in *Hamming spaces.* The approximability of the Hamming versions of Min TSP seems to have never been considered before. The hardness of approximating this problem is one of the main technical results of this paper. The hardness of approximating Min ST in Hamming spaces has been studied for its application to problems in computational biology (specifically, reconstructing evolutionary trees), but it has not been linked to the hardness of the problem in geometric norms.

We now state and discuss our results for Min TSP and Min ST.

[*] A preliminary version of this paper appeared in the proceedings of the 29th ACM Symposium on Theory of Computing [Tre97].

[†] luca@cs.columbia.edu. Columbia University, Department of Computer Science. Work done while at the University of Geneva, at the University of Rome "La Sapienza" and while visiting the Technical University of Catalonia (UPC).

1.1 The Traveling Salesman Problem

Interest in the Min TSP started during the 1930's. In 1966, the (already) long-standing failure of developing an efficient algorithm for the Min TSP led Edmonds [Edm66] to conjecture that the problem is not in P: this is sometimes referred to as the first statement of the $\mathsf{P} \neq \mathsf{NP}$ conjecture. See the book of Lawler et al. [LLKS85] for an extensive survey of results on Min TSP. Here we will only review the results that are relevant to the present paper. The Min TSP is NP-hard even if the cities are restricted to lie in $\mathcal{R}^2$ and the distances are computed according to the ℓ_2 norm [GGJ76, Pap77]. Due to such a negative result, research concentrated on developing good heuristics. Recall that an r-approximate algorithm ($r > 1$) is a polynomial-time heuristic that is guaranteed to deliver a tour whose cost is at most r times the optimum cost. A 3/2-approximation algorithm that works for any metric space is due to Christofides [Chr76]. For more than twenty years, no improvement of this bound had been found, even in the restricted case of geometric metrics.

In the late 1980's, the discovery of the theory of Max SNP-hardness [PY91] gave a tool for understanding this lack of results. Indeed, Papadimitriou and Yannakakis [PY93] proved that the Min TSP is Max SNP-hard even when restricted to metric spaces (as we shall see later, the result also holds for a particularly restricted class of metric spaces). As later shown by Arora et al. [ALM$^+$98], this implies that there exists a constant $\epsilon > 0$ such that metric Min TSP cannot be approximated within a factor $(1 + \epsilon)$ in polynomial time, unless $\mathsf{P} = \mathsf{NP}$. The complexity of approximating Min TSP in the case of geometric metrics remained a major open question. In his PhD thesis, Arora noted that proving the Max SNP-hardness of Euclidean Min TSP in $\mathcal{R}^2$ should be very difficult, but that this could perhaps be done in $\mathcal{R}^{k(n)}$ for sufficiently large $k(n)$ [Aro94, Chapter 9]. In [GKP95], Grigni, Koutsopias and Papadimitriou proved that the restriction of the Min TSP to shortest paths metrics of planar graphs can be approximated within $(1 + \epsilon)$ in time $n^{O(1/\epsilon)}$. Such an approximation algorithm is called a *Polynomial Time Approximation Scheme (PTAS).* This result led Grigni et al. [GKP95] to conjecture that Euclidean Min TSP has a PTAS in $\mathcal{R}^2$. They again posed the question of determining the approximability of the problem for higher dimensions. In a recent breakthrough, Arora [Aro96] developed a PTAS for the Min TSP in $\mathcal{R}^2$ under any ℓ_p metric. Such an algorithm also works in higher dimensional spaces and, in particular, it runs in time $n^{\tilde{O}((\log^{d-2} n)/\epsilon^{d-1})}$ in $\mathcal{R}^d$. A similar approximation scheme (but only for spaces of dimension 2) was also found later by Mitchell [Mit97]. Arora has subsequently improved the running time of his approximation scheme [Aro98]; his new scheme runs in nearly-linear time for any fixed number of dimensions. Specifically, the algorithm of [Aro98] finds a $(1+\epsilon)$-approximate solution in $\mathcal{R}^d$ in time $n(\log n)^{O\left((\sqrt{d}/\epsilon)^d\right)}$. An additional improvement is due to Rao and Smith [RS98]. Their algorithm runs in time $(\sqrt{d}/\epsilon)^{O(d(\sqrt{d}/s)^{d-1})}n + O(dn \log n)$. The dependence of the running time on the number of dimensions is however still doubly exponential.

This is a typical occurrence of the "curse of dimensionality", a phenomenon empirically observed in several cases in computational geometry, that is, the fact that the complexity of a geometric problem grows exponentially or more in the number of dimensions of the space. In some cases (e.g. nearest neighbor search [Kle97, KOR98, IM98]) clever algorithmic solutions can be developed to avoid this exponential growth. In a preliminary version of [Aro96] Arora asked whether an approximation scheme for geometric Min TSP exists for any arbitrary number of dimensions.

Our Results. In this paper we essentially answer negatively to this questions. We prove that Min TSP in $\mathcal{R}^{\log n}$ is Max SNP-hard using any ℓ_p metric. It follows from our result that there cannot be a PTAS for these problems (unless $\mathsf{P}= \mathsf{NP}$) and that there cannot be $(1+\epsilon)$-approximate algorithms in $\mathcal{R}^d$ running in time $n^{O(d/\epsilon)}$ for any $\epsilon > 0$, unless $\mathsf{NP} \subseteq \mathsf{DTIME}(n^{O(\log n)})$.

The Max SNP-hardness is proved by means of a reduction from the version of the metric Min TSP that was shown to be Max SNP-hard in [PY93]. In such metric spaces, any pair of points is either at distance one or two, and an additional technical condition holds. The reduction uses a mapping (see Lemma 9) of the metric spaces of [PY93] into Hamming spaces and the observation (see Proposition 3) that, for elements of $\{0,1\}^n$ a "gap" in the Hamming distance is preserved if distances are computed according to a ℓ_p metric. Our mapping of the metric spaces of [PY91] into Hamming spaces does not preserve distances up to negligible distortion (in fact we suspect that such kind of mapping would be provably impossible). Instead, our mapping introduces a fairly high (yet constant) distortion, but satisfies an additional condition: cities at distance one are mapped into cities at distance $\approx D_1$; cities at distance 2 are mapped into cities at distance $\approx D_2$, and D_2 is larger than D_1 by a multiplicative constant factor. This is sufficient to make the mapping be an *approximation preserving reduction.* Our mapping uses *error-correcting codes* (namely, Hadamard codes) to map cities into an $O(n)$-dimensional Hamming space, and then *random sampling* to reduce the number of dimensions to $O(\log n)$.

The Minimum k-Cities Traveling Salesman Problem (Min k-TSP) and the Minimum Degree-Restricted Steiner Tree Problem (two problems mentioned in Arora's papers [Aro96, Aro98] on approximation schemes for geometric problems) are generalizations of the Min TSP. The hardness results that we prove for Min TSP clearly extend to them.

1.2 The Minimum Steiner Tree Problem

The origins of the Min ST problem are even more remote than the Min TSP's ones: the case when $|U| = 3$ and the metric space is $\mathcal{R}^2$ with the ℓ_2 norm has been studied by Torricelli in the 17th century. A more general case was later considered by Gauss. Recent results about this problem are similar to the ones for Min TSP: exact optimization is NP-hard in $\mathcal{R}^2$ both in the Rectilinear (ℓ_1) case [GJ77] and in the Euclidean (ℓ_2) case [GGJ77]. Constant-factor approximation is achievable in any metric space (the best factor is 1.644 due to Karpinski and Zelikovsky [KZ97]), in general metric spaces the problem is Max SNP-hard [BP89]. Arora [Aro96, Aro98], Mitchell [Mit97], and Rao and Smith [RS98] show how to extend their geometric TSP approximation schemes to geometric Min ST. The running time of these approximation schemes are the same as reported in the previous section for TSP. See also the books by Hwang, Richards and Winters [HRW92] and by Ivanov and Tuzhilin [IT94] and the web page maintained by Ganley [Gan] for extended surveys on the Steiner tree literature. No non-approximability result was known for geometric versions of the Steiner Tree problem. Steiner Tree in Hamming spaces is known to be Max SNP-hard, though this result is usually expressed in terms of an equivalent problem in computational biology (see Wareham's Master Thesis [War93] and also [JW94, War95]).

Our Results. We prove the Max SNP-hardness of Min ST in $\mathcal{R}^n$ under the ℓ_1 norm. We establish this result by means of a reduction from the Min ST problem in Hamming spaces. The reduction is based on the following combinatorial result (Theorem 14): for an instance where all the points are in $\{0,1\}^n \subset \mathcal{R}^n$, there exists an optimum solution where all the Steiner points lie in $\{0,1\}^n$. We prove this fact using the *integrality property* of Min-CUT linear programming relaxations.

1.3 Discussion

For Euclidean Min TSP, there is still a slight slackness between recent approximation schemes and our hardness result. Specifically, a running time $2^{(2^d)/\epsilon}\text{poly}(n)$ would be compatible with our results, but if we believe that NP does not admit sub-exponential algorithms (i.e. $\mathsf{NP} \not\subseteq \mathsf{DTIME}(2^{n^{o(1)}})$), then even a running time $2^{2^{o(d)}/\epsilon}\text{poly}(n)$ is infeasible. For a fixed ϵ, the approximation scheme of Rao and Smith [RS98] runs in time $(\sqrt{d}/\epsilon)^{O(d(\sqrt{d}/s)^{d-1})}n + O(dn \log n)$ which, for fixed ϵ, is roughly $2^{2^{(d \log d)/2 + \log d + \log\log d + \log\log n + O(1)}}$.

There is much more room for improvements for the Min STproblem, however our results at least imply that the number of dimensions *does matter* in the running time of an approximation scheme for this geometric problem.

We feel that one important contribution of this paper is the recognition of Hamming spaces as a class of metric spaces that are somewhat “close” both to arbitrary metrics and to geometric metrics, while also having a nice combinatorial structure. This combination of characteristics makes problems on Hamming metrics an ideal “intermediate” step in reducing a combinatorial problem to a geometric problem. Our hardness result for Hamming Min TSP has been used by Crescenzi et al. [CGP+98] in order to prove the NP-hardness of the protein folding problem.

We also think that it should be worth trying to improve Christofides algorithm in Hamming spaces. While the well-behaved structure of Hamming spaces should not make this task impossible, it is likely that such an improved algorithm could give useful ideas for more general cases.

2 Preliminaries

We denote by $\mathcal{R}$ the set of real numbers. For an integer n we denote by $[n]$ the set $\{1, \ldots, n\}$. For a vector $a \in \mathcal{R}^n$ and an index $i \in [n]$, we denote by $a[i]$ the i-th coordinate of a. The *weight* of a Boolean vector $a \in \{0,1\}^n$ is the number of non-zero entries.

Given an instance x of an optimization problem A, we will denote by $\mathsf{opt}_A(x)$ the cost of an optimum solution for x, we will also typically omit the subscript. For a feasible solution y (usually a tour or a tree) of an instance x of an optimization problem A, we denote its cost by $\mathsf{cost}_A(x, y)$ or, more often, as $\mathsf{cost}(y)$. See e.g. [BC93, Pap94] for formal definitions about optimization problems. In this paper we will use the notions of L-reduction and Max SNP-hardness. Max SNP is a class of constant-factor approximable optimization problems that includes Max 3SAT, we refer the reader to [PY91] for the formal definition.

Definition 1 (L-reduction) *An optimization problem A us said to be L-reducible to an optimization problem B if two constants α and β and two polynomial-time computable functions f and g exist such that*

1. *For an instance x of A, $x' = f(x)$ is an instance of B, and it holds $\mathsf{opt}_B(x') \leq \alpha \mathsf{opt}_A(x)$.*

2. *For an instance x of A, and a solution y' feasible for $x' = f(x)$, $y = g(x, y')$ is a feasible solution for x and it holds $|\mathsf{opt}_A(x) - \mathsf{cost}_A(x, y)| \leq \beta |\mathsf{opt}_B(x') - \mathsf{cost}_B(x', y')|$.*

We say that an optimization problem A is Max SNP-hard if all Max SNP-problems are L-reducible to A. From [ALM+98] it follows that if a problem A is Max SNP-hard, then a constant $\epsilon > 0$ exists such that $(1 + \epsilon)$-approximating A is NP-hard.

A function $d : U \times U \to \mathcal{R}$ is a *metric* if the following properties hold:

1. $d(u, v) \geq 0$ for all $u, v \in U$;

2. $d(u, v) = 0$ if and only if $u = v$;

3. $d(u, v) = d(v, u)$ for any $u, v \in U$ (symmetry);
4. $d(u, v) \leq d(u, z) + d(z, v)$ for any $u, v, z \in U$ (triangle inequality.)

If all properties but (2) hold, then d is said to be a *semi-metric*. Abusing notation, we will usually adopt the term "metric" both for metrics and semi-metrics.

Definition 2 (**$(1, 2) - B$ metrics**) *For a positive integer B, a metric $d : U \times U \to \mathcal{R}$ is a $(1, 2) - B$ metric if it satisfies the following properties:*

1. *For any $u, v \in U$, $u \neq v$, $d(u, v) \in \{1, 2\}$.*
2. *For any $u \in U$, at most B elements of U are at distance 1 from u.*

Papadimitriou and Yannakakis [PY93] have shown that a constant $B_0 > 0$ exists such that the Min TSP is Max SNP-hard even when restricted to $(1, 2) - B_0$ metrics.

For an integer $p \geq 1$, the ℓ_p norm in $\mathcal{R}^n$ is defined as $||(u_1, \ldots, u_n)||_p = (\sum_{i=1}^n |u_i|^p)^{(1/p)}$. The distance induced by the ℓ_p norm is defined as $d_p(u, v) = ||u - v||_p$. For a positive integer n, we denote by d_H^n the Hamming metric in $\{0, 1\}^n$ (we will usually omit the superscripts). We will make use of the following fact.

Proposition 3 *Let $u, v \in \{0, 1\}^n \subseteq \mathcal{R}^n$. Then $d_p(u, v) = d_H(u, v)^{1/p}$.*

Before starting with the presentation of our results, we make the following important caveat.

Remark 4 *In some of the proofs of this paper we implicitly make the (unrealistic) assumption that arbitrary real numbers can appear in an instance and that arithmetic operations (including squared roots) can be computed over them in constant time. However, our results still hold if we instead assume that numbers are rounded and stored in a floating point notation using $O(\log n)$ bits. This fact follows from a modification of the argument used in [Aro96] to reduce a general instance of Euclidean TSP or Steiner Tree into an instance where coordinates are positive integers whose value is $O(n^2)$.*

3 The Min TSP

Our hardness result is based on a "distance preserving" embeddings of $(1, 2) - B$ metric spaces into Hamming spaces. We first define the kind of embedding we are looking for.

Definition 5 *For an integer B, a $(1, 2) - B$ metric space (U, d), an integer k and positive reals $D_1, D_2 > 0$ and $0 < \epsilon < 1/2$, we say that a mapping $f : U \to \{0, 1\}^k$ is (k, D_1, D_2, ϵ)-good if for any $u, v \in U$:*

1. *If $d(u, v) = 1$, then $D_1(1 - \epsilon) \leq d_H(f(u), f(v)) \leq D_1(1 + \epsilon)$.*
2. *If $d(u, v) = 2$, then $D_2(1 - \epsilon) \leq d_H(f(u), f(v)) \leq D_2(1 + \epsilon)$.*

In particular, if f is a $(k, D_2, D_1, 0)$-good embedding, then pairs at distance 2 are mapped into pairs at distance D_2 nd pairs at distance 1 are mapped into pairs at distance D_1.

Recall that, for any $n = 2^h$ that is a power of 2, the *first-order Reed-Muller Code* (which is also an *Hadamard Code*) $H_n \subset \{0, 1\}^n$ is a set of n binary strings of length n whose pairwise Hamming distance is $n/2$. The elements of H_n can be seen as the set of liner functions $l : \{0, 1\}^h \to \{0, 1\}$. See e.g. [vLW92, Chapter 18] and the references therein for more details. The set H_n is computable in time polynomial in n.

Lemma 6 *There exists a polynomial time algorithm that on input a $(1,2)-B$ metric with n points, where n is a power of two, finds a $((B+1)n, Bn/2, (B+1)n/2, 0)$-good embedding.*

PROOF: Let $U = \{u_1, \ldots, u_n\}$. Recall that a $(1,2) - B$ metric (U, d) can be represented as an undirected graph $G = (U, E)$, where $\{u, v\} \in E$ iff $d(u, v) = 1$ (see [PY93]). Note that G has maximum degree B.

We claim that we can find in polynomial time a partition of E into $B+1$ matchings $E_1, \ldots, E_{B+1}$. To prove this claim, it suffices to observe that the problem of partitioning the set of edges of a graph into disjoint matchings is a restatement of the *edge coloring problem.* In a graph of maximum degree B, an edge coloring with $B+1$ colors can be found in polynomial time [Hol81].

We now describe the embedding. Each node $u \in U$ is mapped into a string $f(u)$ that is the concatenation of $B+1$ strings $a_u^1, \ldots, a_u^{B+1} \in H_n$:

$$f(u) = a_u^1 \circ \ldots \circ a_u^{B+1} \ .$$

For a fixed $i \in \{1, \ldots, B+1\}$, the strings $\{a_u^i\}_{u \in U}$ are chosen arbitrarily in H_n such that $a_u^i = a_v^i$ if and only if $\{u, v\} \in E_i$. Since H_n can be generated in polynomial time in n, the overall construction can be carried out in poly(n) time.

Let us now compute the distance between two strings $f(u)$ and $f(v)$. There are two cases to be considered.

1. If $\{u, v\} \notin E$, then $a_u^i \neq a_v^i$ for all $i = 1, \ldots, B+1$, and so $d_H(f(u), f(v)) = (B+1) \cdot n/2$.
2. If $\{u, v\} \in E$, then $\{u, v\} \in E_j$ for some j, and we have $a_u^j = a_v^j$ and $a_u^i \neq a_v^i$ for $i \neq j$. It follows that $d_H(f(u), f(v)) = B \cdot n/2$.

□

We also observe the following simpler result

Lemma 7 *There exists a polynomial time algorithm that on input a $(1,2)-B$ metric with n points finds a $((B+1)n, 2B, 2(B+1), 0)$-good embedding. Furthermore, any vector in the embedding has weight precisely $B+1$.*

PROOF: Use the same construction of the proof of Lemma 6, but using the code $I_n \subset \{0,1\}^n$ composed of all the n vectors of length n having exactly one non-zero entry. Any two elements of such a code have distance precisely 2. □

Our next goal is to map the instances of Hamming TSP produced by Lemma 6 into Hamming spaces of logarithmic dimension. Observe that the distance between any two points is a constant fraction of the number of dimensions. We will show how to exploit this property using *random sampling.* The main fact about random sampling is the following: let $b_1, \ldots, b_n \in \{0,1\}$ be unknown values. If we pick a random sub(multi-)set $b_{i_1}, \ldots, b_{i_m}$ of m elements, where $m = O((\log 1/\delta)/\epsilon^2)$, then with probability $1 - \delta$ it holds that

$$\left| \sum_{i=j}^{n} b_j - \frac{n}{m} \sum_{j=1}^{m} b_{i_j} \right| \leq \epsilon n \ .$$

Now, if we pick $m = O((\log n)/\epsilon^2)$ coordinates from the target Hamming space of the previous reduction, and we project the mapping on this (multi)subset of coordinates, the distance between two fixed cities will deviate from the expected one by a factor at most ϵm with probability $(1 - 1/\text{poly}(n))$. In particular, there is a constant probability that all the pairwise distances are simultaneously "distorted" by at most ϵm. Using the oblivious sampler of Bellare and Rompel [BR94] (or alternatively, the Chernoff bound for random walks on expander graphs [Gil98]) we can find such a set of $O((\log n)/\epsilon^2)$ coordinates *deterministically* in polynomial time. Details follow. We first state formally the result of Bellare and Rompel.

Lemma 8 (Randomness-efficient sampling [BR94]) *There exists a randomized algorithm (called* sampler*) that on input parameters $\epsilon > 0, \delta > 0$ and n, computes a sequence of (not necessarily distinct) indices $i_1, \ldots, i_k$ where $k = O((1/\epsilon^2)\log 1/\delta)$ such that for any fixed sequence of boolean values $b_1, \ldots, b_n \in \{0,1\}$, the following holds:*

$$\mathbf{Pr}\left[\left|\frac{1}{n}\sum_j a_j - \frac{1}{k}\sum_j a_{i_j}\right| > \epsilon\right] \leq \delta$$

where the probability is taken over the random choices of the sampler. Furthermore, the sampler only uses $O(1/\epsilon^2 + \log 1/\delta)$ random bits.

We can now state and prove our result about embeddings in Hamming spaces with a logarithmic number of dimensions.

Lemma 9 *There exists a polynomial time algorithm that, for any B and $\gamma > 0$, on input a $(1,2)-B$ metric (U,d) with n points finds a $(k, DB/(B+1), D, \gamma)$-good embedding of U, where $k = O((\log Bn)/\gamma^2)$ and $D = k/2$.*

Proof: The algorithm first computes an embedding of (U,d) into $\{0,1\}^m$ according to Lemma 6, where $m \leq 2(B+1)|U|$ (recall that Lemma 6 can be applied only to metric spaces whose number of points is power of two, but this can be achieved by adding an appropriate number of dummy points; this at most doubles the number of dimensions of the final embedding.) Then Lemma 8 is applied, with parameters $\epsilon = \gamma/3$ and $\delta = 1/n^2$. Let $i_1, \ldots, i_k$ be the sequence of indices given by the sampler, where $k = O(\log Bn/\gamma^2)$. Consider the embedding f' defined as $f'(u)[j] = f(u)[i_j]$, that is f' maps a point u into a substring of $f(u)$ defined by the indices $i_1, \ldots, i_k$. For any pair of vertices u and v, consider the string $a_1, \ldots, a_m$ where $a_j = |f(u)[j] - f(v)[j]|$ Clearly the Hamming distance between $f(u)$ and $f(v)$ is equal to $\sum_{j=1}^m a_j$. On the other hand, the Hamming distance between $f'(u)$ and $f'(v)$ is $\sum_j a_{i_j}$ We observe that with probability at least $1 - \delta = 1 - 1/n^2$ it holds that

$$\left|\frac{1}{m}d_H(f(u), f(v)) - \frac{1}{k}d_H(f'(u), f'(v))\right| \leq \epsilon$$

and such a relation holds for all the pairs u, v simultaneously with positive probability. Thus, one of the possible outputs of the sampler is a sequence $i_1, \ldots, i_k$ such that for all u and v

$$\left|\frac{k}{m}d_H(f(u), f(v)) - d_H(f'(u), f'(v))\right| \leq k\epsilon$$

such a sequence can be found in polynomial time by considering all the polynomially many possible random choices of the sampler. It is left to the reader to verify that the resulting embedding satisfies the requirement of the Lemma. □

Theorem 10 *For any p there exists a constant $\epsilon_p > 0$ such that* Min TSP *is* NP*-hard to approximate within a factor $(1+\epsilon_p)$ even when restricted to ℓ_p spaces of logarithmic dimension.*

Proof: From [PY93] and [ALM+98] we have the following result: constants $B_0 > 0$ and $r_0 > 1$ exist such that, given an instance (U, d) of Min TSP with a $(1,2) - B_0$ metric and n cities, and given the promise that either $\mathsf{opt}(U, d) = n$ or $\mathsf{opt}(U, d) \geq r_0 n$, it is NP-hard to distinguish which of the two cases holds.

Fix a constant γ such that

$$\frac{1-\gamma}{1+\gamma}\left(1+\frac{(r_0-1)}{B_0}\right) > 1 + \frac{(r_0-1)}{2B_0} \stackrel{\text{def}}{=} 1+\epsilon_p \ .$$

Such a constant γ must exists since for $\gamma \to 0$ the left-hand side tends to a value strictly greater than the right-hand side.

Given an instance (U, d) of $(1,2) - B_0$ Min TSP with n cities, we use Lemma 9 to map it into a Hamming space of dimension $k = O(\log n)$ using a $(k, DB_0/(B_0+1), D, \gamma)$-good embedding with $D = k/2$. Let $f : U \to \{0,1\}^k$ denote such embedding. We consider two cases.

- If $\mathsf{opt}(U, d) = n$, then an optimum solution for U will have cost at most $n \cdot D(B_0/(B_0+1)) \cdot (1+\gamma)$ for U'.
- If $\mathsf{opt}(U, d) \geq nr_0$, then there can be no solution for U' of cost less than $nD(B_0/(B_0+1))(1-\gamma) + (r_0-1)n(1/(B_0+1))D(1-\gamma)$. Otherwise, the same solution would have cost less than nr_0 for U.

Distinguishing between the two cases is NP-hard, therefore it is NP-hard to approximate the target instance to within a factor

$$\frac{1-\gamma}{1+\gamma} \cdot \frac{(B_0+r_0-1)/(B_0+1)}{B_0/(B_0+1)} \geq 1+\epsilon_p$$

□

Remark 11 *The claim of the Theorem asks for the cities to be in $\mathcal{R}^{\log n}$, rather than in $\mathcal{R}^{c\log n}$ as in the previous construction. However, we can add $(n^c - n)$ new cities, all at distance $1/n^{c+1}$ from a given one. This perturbs the optimum in a negligible way, and gives an instance with $N = n^c$ cities in $\mathcal{R}^{\log N}$.*

Using techniques of Khanna et al. [KMSV99], the non-approximability result of Theorem 10 implies that geometric Min TSP in $\mathcal{R}^{\log n}$ under any ℓ_p norm is APX PB-hard (in particular, Max SNP-hard) under E-reductions and APX-complete under AP-reductions [CKST95].

3.1 Additional Remarks

Using the embedding of Lemma 7 in the proof of Theorem 10 one can prove the Max SNP-hardness of Min TSP when restricted to Hamming instances with a constant bound on the weight of the points. Reducing from TSP(1,2) in graphs with maximum degree 3 (an NP-hard problem) it is possible to prove the NP-hardness of the Hamming TSP problem in instances where all points have weight precisely 4. These problems were not known to be NP-hard before. It is reasonable to conjecture that Hamming TSP is solvable in polynomial time for instances where all points have weight at most 2 and is NP-hard for instances where all points have weight 3. We do not know how to prove this conjecture.

The use of Hadamard codes in the proof of Lemma 6 is motivated by our desire to produce Hamming instances where any two points have a linear (in the number of dimensions) distance. This property is essential to project the embedding down to a logarithmic number of dimensions.

4 The Min ST Problem

The hardness of approximating Min ST will be established by means of a reduction from the Hamming version of the problem.

The following result appears in [War93, Theorem 45, Part 3]. It is based on the observation that the NP-hardness proof of Hamming Min TSP appeared in [DJS86] yields an L-reduction from Vertex Cover in bounded degree graphs (a problem proved to be Max SNP-hard in [PY91]) to Hamming Min ST.

Theorem 12 ([War93]) Min ST *is* Max SNP*-hard even when restricted to Hamming spaces.*

Remark 13 *Unaware of these previous results, we presented a (independently found) proof of Theorem 12 in a preliminary version of this paper [Tre97].*

Our goal is to reduce the Steiner Tree problem in Hamming spaces to the Steiner Tree problem under the ℓ_1 distance. We note that for points in $\{0,1\}^n$ the ℓ_1 distance equals the Hamming distance. However, the reduction is non-trivial since $\mathcal{R}^n$ contains many points that are not in $\{0,1\}^n$ and we have to argue that having much more choice for the Steiner nodes does not make the problem easier. The Rectilinear Min ST problem looks very much like a *relaxation* of the Hamming Min ST problem; our reduction makes use of a *rounding scheme* proving that the relaxation does not change the optimum.

Theorem 14 *Let $U \subseteq \{0,1\}^n \subset \mathcal{R}^n$ be an instance of Rectilinear* Min ST *all whose points are in the Boolean cube. Let T be a feasible solution for U. Then it is possible to find in polynomial time (in the size of T) another solution T' such that $\mathsf{cost}(T') \leq \mathsf{cost}(T)$ and all the Steiner nodes of T' are in $\{0,1\}^n$.*

Before proving the theorem, we note the following relevant consequence.

Corollary 15 *For any instance $U \subseteq \{0,1\}^n$ of Rectilinear* Min ST*, an optimum solution exists all whose Steiner points are in $\{0,1\}^n$.*

We now prove Theorem 14.

Proof:[Of Theorem 14] Let $S = \{s_1, \ldots, s_m\}$ be the set of Steiner points of T, and let E be the set of edges of T. For any $s_j \in S$ we will find a new point $s'_j \in \{0,1\}^n$, so that if we let T' be the tree obtained from T by substituting the s points with the corresponding s' points, the cost of T' is not greater than the cost of T. The latter statement is equivalent to

$$\begin{aligned} & \sum_{(s_j,u)\in E, u\in U} ||s_j - u||_1 + \sum_{(s_j,s_h)\in E} ||s_j - s_h||_1 \\ \geq & \sum_{(s'_j,u)\in E, u\in U} ||s'_j - u||_1 + \sum_{(s'_j,s'_h)\in E} ||s'_j - s'_h||_1 \end{aligned}$$

We will indeed prove something stronger, namely, that for any $i \in [n]$ it holds

$$\begin{aligned}&\sum_{(s_j,u)\in E, u\in U} |s_j[i]-u[i]| + \sum_{(s_j,s_h)\in E} |s_j[i]-s_h[i]| \\ \geq &\sum_{(s'_j,u)\in E, u\in U} |s'_j[i]-u[i]| + \sum_{(s'_j,s'_h)\in E} |s'_j[i]-s'_h[i]|\end{aligned} \tag{1}$$

Let $i \in [n]$ be fixed, we now see how to find values of $s'_1[i], \ldots, s'_m[i] \in \{0,1\}$ such that (1) holds. We express as a linear program the problem of finding values of $s'_1[i], \ldots, s'_m[i]$ that minimize the right-hand side of (1). For any $j \in [m]$ we have a variable x_j (representing the value to be given to $s'_j[i]$) and for any edge $e=(a,b)$ such that at least one endpoint is in S we have a variable y_e, representing the length $|a[i]-b[i]|$. The linear program is as follows

$$\begin{array}{lll} \min & \sum_e y_e & \\ \text{s.t.} & & \\ & y_e \geq x_j - x_h & \forall e=(s_j,s_h)\in E \\ & y_e \geq x_h - x_j & \forall e=(s_j,s_h)\in E \\ & y_e \geq x_j & \forall e=(s_j,u_h)\in E. u_h[i]=0 \\ & y_e \geq 1-x_j & \forall e=(s_j,u_h)\in E. u_h[i]=1 \\ & x_j \geq 0 & \\ & y_e \geq 0 & \end{array}$$

(LP).

Setting $x_j = s_j[i]$ and setting $y_{(a,b)} = |a[i]-b[i]|$ yields a feasible solution, and its cost is the left-hand side of (1). Let (x^*, y^*) be an optimum solution for (LP). From the previous observation we have that setting $s'_j[i] = x^*_j$ we satisfy (1). It remains to be seen that (LP) has an optimum solution where all variables take value from $\{0,1\}$. This follows from the fact that (LP) is the linear programming relaxation of an undirected Min-CUT problem, where all the u such that $u[i]=0$ (respectively, $u[i]=1$) are identified with the source (respectively, the sink), each s_j is a node, and the edges are like in T. It is well known (see e.g. [PS82]) that a Min-CUT linear programming relaxation has optimum 0/1 solutions, and that such a solution can be found in polynomial time. □

Remark 16 *There seems to be no natural analog of Theorem 14 in other norms. Even in $\mathcal{R}^2$, using the Euclidean metric, we have that the optimum solution of the instance $\{(0,0),(1,0),(0,1)\}$ must use a Steiner point not in $\{0,1\}^2$.*

Theorem 17 *Rectilinear* MIN ST *is* Max SNP*-hard.*

PROOF: We reduce from Hamming MIN ST. The reduction leaves the instance unchanged. For an instance $U \subseteq \{0,1\}^n$, we let $\mathsf{opt}_H(U)$ (respectively, $\mathsf{opt}_R(U)$) be the cost of an optimum solution for U, when seen as an instance of Hamming MIN ST (respectively, of Rectilinear MIN ST). Clearly, we have that $\mathsf{opt}_R(U) \leq \mathsf{opt}_H(U)$. Given a solution T for U, we find a solution T' as in Theorem 14. Since in $\{0,1\}^n$ the distance induced by the ℓ_1 norm equals the Hamming distance, we have that $\mathsf{cost}_H(T') = \mathsf{cost}_R(T') \leq \mathsf{cost}_R(T)$. We have an L-reduction with $\alpha = \beta = 1$. □

5 Conclusions and Open questions

We do not know how to extend our non-approximability result for Min ST to the Euclidean case. Arora [Aro96] notes that, by inspecting the way his algorithm works, it is possible to claim that, for any instance of Euclidean Min ST, there exists a near-optimal solution where the Steiner points lie in some well-specified positions (either at "portals" or in positions chosen at the bottom of the recursion). This observation could perhaps be a starting point.

We do not have explicit estimations of the constants to within which it is hard to approximate geometric Min TSP and rectilinear Min ST. The constant for Min TSP should be only slightly smaller than the corresponding constant for the $(1,2)-B$ case. An explicit non-approximability factor has been estimated by Engebresten [Eng98] for the latter problem, and it is very close to 1. The constant for Min ST should be slightly better. Finding much stronger estimations (comparable to the 3/2 bound of Christofides and the 1.644 bound of Karpinski and Zelikovsky) is an open and challenging question.

Acknowledgments. I am grateful to Sanjeev Arora, Madhu Sudan, László Babai and Janos Körner for very interesting and helpful discussions. I thank Ray Greenlaw for having read a preliminary version of this paper and given several useful remarks. I wish to thank Jens Lagergren, Tao Jiang and Todd Wareham for providing me references and reprints about their work in computational biology, and for explaining to me the relation with the Hamming Steiner Tree Problem.

References

[ALM+98] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, 1998. Preliminary version in *Proc. of FOCS'92.*

[Aro94] S. Arora. *Probabilistic Checking of Proofs and Hardness of Approximation Problems.* PhD thesis, University of California at Berkeley, 1994.

[Aro96] S. Arora. Polynomial time approximation schemes for Euclidean TSP and other geometric problems. In *Proceedings of the 37th IEEE Symposium on Foundations of Computer Science*, pages 2–11, 1996.

[Aro98] S. Arora. Polynomial time approximation schemes for Euclidean Traveling Salesman and other geometric problems. *Journal of the ACM*, 45(5), 1998.

[BC93] D.P. Bovet and P. Crescenzi. *Introduction to the Theory of Complexity.* Prentice Hall, 1993.

[BP89] M. Bern and P. Plassmann. The Steiner tree problem with edge lengths 1 and 2. *Information Processing Letters*, 32:171–176, 1989.

[BR94] M. Bellare and J. Rompel. Randomness-efficient oblivious sampling. In *Proceedings of the 35th IEEE Symposium on Foundations of Computer Science*, pages 276–287, 1994.

[CGP+98] P. Crescenzi, D. Goldman, C.H. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. In *Proceedings of the 30th ACM Symposium on Theory of Computing*, pages 597–603, 1998.

[Chr76] N. Christofides. Worst-case analysis of a new heuristic for the travelling salesman problem. Technical report, Carnegie-Mellon University, 1976.

[CKST95] P. Crescenzi, V. Kann, R. Silvestri, and L. Trevisan. Structure in approximation classes. In *Proceedings of the 1st Combinatorics and Computing Conference*, pages 539–548. LNCS 959, Springer-Verlag, 1995.

[DJS86] W.H.E. Day, D.S. Johnson, and D. Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81:33–42, 1986.

[Edm66] J. Edmonds. Optimum branchings. *Journal of Research National Bureau of Standards, Part B*, 17B(4):233–240, 1966.

[Eng98] L. Engebrester. An explicit lower bound for TSP with distances one and two. Technical Report TR98-046, Electronic Colloquium on Computational Complexity, 1998.

[Gan] J. Ganley. The Steiner tree page. Web page, URL http://www.cs.virginia.edu/~jlg8k/steiner.

[GGJ76] M.R. Garey, R.L. Graham, and D.S. Johnson. Some NP-complete geometric problems. In *Proceedings of the 8th ACM Symposium on Theory of Computing*, pages 10–22, 1976.

[GGJ77] M.R. Garey, R.L. Graham, and D.S. Johnson. The complexity of computing Steiner minimal trees. *SIAM Journal of Applied Mathematics*, 34:477–495, 1977.

[Gil98] D. Gillman. A Chernoff bound for random walks on expander graphs. *SIAM Journal on Computing*, 27(4):1203–1220, 1998.

[GJ77] M.R. Garey and D.S. Johnson. The rectilinear Steiner tree problem is NP-complete. *SIAM Journal on Applied Mathematics*, 32(4):826–834, 1977.

[GKP95] M. Grigni, E. Koutsoupias, and C.H. Papadimitriou. An approximation scheme for planar graph TSP. In *Proceedings of the 36th IEEE Symposium on Foundations of Computer Science*, pages 640–645, 1995.

[Hol81] I. Holyer. The NP-completeness of edge-coloring. *SIAM Journal on Computing*, 10:718–720, 1981.

[HRW92] F.K. Hwang, D.S. Richards, and P. Winter. *The Steiner Tree Problem.* North-Holland, 1992.

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th ACM Symposium on Theory of Computing*, pages 604–613, 1998.

[IT94] A.O. Ivanov and A.A. Tuzhilin. *Minimal Networks: The Steiner Problem and its Generalizations.* CRC Press, 1994.

[JW94] T. Jiang and L. Wang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348, 1994.

[Kle97] Jon M. Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In *Proceedings of the 29th ACM Symposium on Theory of Computing*, pages 599–608, 1997.

[KMSV99] S. Khanna, R. Motwani, M. Sudan, and U. Vazirani. On syntactic versus computational views of approximability. *SIAM Journal on Computing*, 28(1):164–191, 1999. Preliminary version in *Proc. of FOCS'94*.

[KOR98] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. In *Proceedings of the 30th ACM Symposium on Theory of Computing*, pages 614–623, 1998.

[KZ97] M. Karpinski and A. Zelikovsky. New approximation algorithms for the Steiner tree problems. *Journal of Combinatorial Optimization*, 1:1–19, 1997.

[LLKS85] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys. *The Traveling Salesman Problem*. John Wiley, 1985.

[Mit97] J.S.B. Mitchell. Guillotine subdivisions approximate polygonal subdivisions: A simple polynomial-time approximation scheme for geometric TSP, K-MST, and related problems. *SIAM Journal on Computing*, 1997. To appear.

[Pap77] C.H. Papadimitriou. Euclidean TSP is NP-complete. *Theoretical Computer Science*, 4:237–244, 1977.

[Pap94] C.H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[PS82] C.H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, 1982.

[PY91] C. H. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences*, 43:425–440, 1991. Preliminary version in *Proc. of STOC'88*.

[PY93] C.H. Papadimitriou and M. Yannakakis. The travelling salesman problem with distances one and two. *Mathematics of Operations Research*, 18:1–11, 1993.

[RS98] S.B. Rao and W.D. Smith. Approximating geometrical graphs via "spanners" and "banyans". In *Proceedings of the 30th ACM Symposium on Theory of Computing*, pages 540–550, 1998.

[Tre97] L. Trevisan. When Hamming meets Euclid: The approximability of geometric TSP and MST. In *Proceedings of the 29th ACM Symposium on Theory of Computing*, pages 21–29, 1997.

[vLW92] J.H. van Lint and R.M. Wilson. *A Course in Combinatorics*. Cambridge University Press, 1992.

[War93] H.T. Wareham. On the computational complexity of inferring evolutionary trees. Master's thesis, Memorial University of Newfoundland, Canada, 1993. Available at `http://www.csr.uvic.ca/~harold/`.

[War95] H.T. Wareham. A simplified proof of the NP- and MAX SNP-hardness of multiple sequence tree alignment. *Journal of Computational Biology*, 2(4):509–514, 1995.