Embeddings and Non-approximability of Geometric Problems

Venkatesan Guruswami* Piotr Indyk†

1 Introduction

In this paper we present two results. Our first theorem shows that "sparse" (see definition later) metric spaces, with distances 1 and 2, are isometrically embeddable into low-dimensional l_∞ norm. A similar in "spirit" result was earlier proved for l_p norms with *finite* p by Trevisan [Tre97]. Our result is obtained using a different technique (Bourgain's sampling approach) and could not have been achieved using the techniques of [Tre97].

In the second part of the paper we apply the aforementioned embeddings to show non-approximability results for several geometric problems in l_p norms with logarithmic dimension, namely TSP, k-median and min-sum k-clustering. In particular, we provide the first known constant factor hardness for the min-sum k-clustering problem.

2 The embedding

A metric $M = (X, D)$ is a $(1, 2) - B$ metric, if all distances are either 1 or 2, and for any node $u \in X$ there are at most B nodes within distance 1 from u. Below we show that such metrics can embedded into l_∞ with no distortion. In fact, we prove the following lemma, which is a bit more general.

Lemma 1. *Let $M = (X, D)$ be a metric with all distances in $\{1, 2\}$, $|X| = n$. Then for any $B \geq 1$ there exists a mapping $f : X \to [0, 2]^{O(B \log n)}$ such that $|f(p) - f(q)|_\infty \leq D(p, q)$, for any $p, q \in X$, and $|f(p) - f(q)|_\infty = D(p, q)$, for any $p, q \in X$ such that $|\{u : D(u, q) < 2\}| \leq B$.*

Proof: By probabilistic method (which can be derandomized in polynomial time via the method of conditional probabilities). For $i = 1 \ldots l = O(B \log n)$, choose a set $S_i \subset X$ such that each element of X is included in S_i independently with probability $1/B$. Then we define $f(p) = [D(p, S_1), \ldots, D(p, S_l)]$, where $D(p, S) = \min_{s \in S} D(p, s)$. Observe that for each i and each q such that $|\{u : D(u, q) < 2\}| \leq B$ we have $\Pr[|D(p, S_i) - D(q, S_i)| = D(p, q)] = \Omega(1/B)$. Indeed, consider first the case $D(p, q) = 2$. Then

- with probability $1/B$ we have $p \in S$
- with constant probability $\{u : D(u, q) < 2\} \cap S_i = \emptyset$

and the probabilities are independent. In such a case we have $|D(p, S_i) - D(q, S_i)| = |0 - 2| = D(p, q)$. The case $D(p, q) = 1$ is similar. By repeating the process $O(B \log n)$ times we make sure the embedding is correct with high probability. The non-expansion of the mapping follows from (a standard use of) triangle inequality. □

Comments: In [Tre97], Trevisan showed that for any finite $p > 1, \epsilon > 0$, $B \geq 1$, there exists $\alpha > 0$ so that for any $(1, 2) - B$ metric $M = (X, D)$, $|X| = n$, one can construct a mapping $f : X \to l_p^k$, $k = O(\log n/\epsilon^2)$, such that for any $u, v \in X$:

- if $D(u, v) = 1$, then $|f(u) - f(v)|_p = (1 \pm \epsilon)$
- if $D(u, v) = 2$, then $|f(u) - f(v)|_p = (1 + \alpha)(1 \pm \epsilon)$

The value of α tends to 0 as p tends to ∞, and therefore the construction cannot be used when $p = \infty$. Our theorem extends this result to the case $p = \infty$ (in fact, with optimal $\alpha = 1$ and $\epsilon = 0$).

3 Hardness results

For many "metric" problems, a polynomial time approximation scheme (PTAS) for $(1, 2) - O(1)$ metrics does not exist (unless $P = NP$ or similar collapse of complexity classes). The results from the previous section imply that the same hardness results hold even when the distances are induced by n points in the $l_p^{O(\log n)}$ norm. This implies that the doubly exponential dependence on dimension, present in the polynomial time approximation schemes for some of those problems (e.g., [Ar98, RS98, ARR98, KR99]), is indeed necessary.

*Computer Science Division, UC Berkeley.
†Laboratory for Computer Science, MIT.

In the following we focus on the following problems.

TSP. Considered earlier in [Tre97]; our contribution is to extend his results to the case $p = \infty$.

k**-median.** The problem is defined as follows: given a metric space $M = (X, D)$ and $k > 0$, find a set C of k points in X to minimize $\sum_{p \in X} \min_{c \in C} D(p, c)$. It was shown in [GK98] that a closely related problem of *uncapacitated facility location with unit demands* is MaxSNP-hard for $(1, 2) - O(1)$ metrics. It is easy to see that this problem (and its approximate version) is polynomial-time reducible to (the approximate version of) k-median over the same metrics. Thus, it follows that the existence of a PTAS for k-median for $(1, 2) - B$ metrics is unlikely. By using the embeddings, we obtain the same conclusion for the metrics induced by the $l_p^{O(\log n)}$ norm.

Min-sum k**-clustering.** The problem is defined as follows: given a metric space $M = (X, D)$ and $k > 0$, partition X into k clusters $C_1 \ldots C_k$ to minimize $\sum_i \sum_{p,q \in C_i} D(p, q)$. Below we show that this problem is hard to approximate within a certain constant factor. In addition to the implications mentioned earlier, this result justifies the exponential in k dependence of the running times of the approximation schemes of [dlVKKR02] (since their running times are only polynomial in the dimension). Moreover, it complements the $\log^{O(1)} n$-approximation algorithm of [BCR01].

Theorem 1. *There exists an integer B such that Min-sum k-clustering is NP-hard to approximate within some constant factor when the number of clusters k is part of the input, even on $(1, 2) - B$ metrics.*

Proof: The reduction is from the Maximum Triangle Packing problem. Spefically, we use the following hardness result: For some fixed $\delta > 0$ and integer B, given a graph G on n vertices with maximum degree B, it is NP-hard to distinguish between the cases when G has $n/3$ vertex-disjoint triangles and when G has at most $(1 - \delta)n/3$ pairwise vertex-disjoint triangles. This result was shown for 3-dimensional matching by Petrank [Pet94] and a simple gadget gives the result for triangle packing. We assume that $\delta < 1/55$.

Define a metric $M = (X, D)$ on the vertex set of G with $D(x, y) = 1$ if (x, y) is an edge in G, and 2 otherwise. Since G has maximum degree B, this defines a $(1, 2) - B$ metric. Consider the Min-sum k-clustering problem on M with $k = n/3$. If G has a triangle packing of $n/3$ vertex-disjoint triangles, using these triangles as the clusters gives a k-clustering of total cost n.

We now claim that if G does not contain more than $(1 - \delta)n/3$ vertex-disjoint triangles, then every k-clustering of M has Min-sum cost greater than $(1 + \delta^2)n$. Clearly this gives us the claimed hardness result.

Indeed, suppose there is a k-clustering C with cost at most $(1 + \delta^2)n$. Remove all triangles used by the clustering, i.e. clusters of size 3 whose nodes are all connected in G, and suppose a set S of m nodes remain. Since we remove at most $(1 - \delta)n/3$ triangles, we have $m \geq \delta n$. We now have a $m/3$-clustering, say C', of the nodes in S with cost at most $(1 + \delta^2)n - (n - m) = m + \delta^2 n \leq (1 + \delta)m$.

We claim that this cannot happen. The intuition is that the average size of a cluster is 3, and any cluster of size 3 is C' has cost at least 4 (since C' uses no triangles), so the cost per element cannot be less than $(1 + \delta)$. The formal proof has been relegated to the Appendix. □

References

[Ar98] S. Arora. Polynomial-time Approximation Schemes for Euclidean TSP and other Geometric Problems. *Journal of the ACM* 45(5) 753-782, 1998.

[ARR98] S. Arora and P. Raghavan and S. Rao. Polynomial Time Approximation Schemes for Euclidean k-medians and related problems. *Symposium on Theory of Computing*, 1998.

[BCR01] Y. Bartal and M. Charikar and D. Raz. Approximating Min-Sum k-Clustering in Metric Spaces. *Symposium on Theory of Computing*, 2001.

[dlVKKR02] W. Fernandez de la Vega and M. Karpinski and C. Kenyon and Y. Rabani . Polynomial Time Approximation Schemes for Metric Min-Sum Clustering. *ECCC Report TR02-025*, 2002.

[GK98] S. Guha and S. Khuller. Greedy strikes back: Improved Facility Location Algorithms. *ACM-SIAM Symposium on Discrete Algorithms*, 1998

[KR99] S. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the Euclidean k-median problem. *European Symposium on Algorithms*, 1999.

[Pet94] Erez Petrank. The hardness of approximation: gap location. *Computational Complexity*, Vol. 4, 1994. pp. 133-157.

[RS98] S. Rao and W. Smith. Approximating geometrical graphs via spanners and banyans. *Foundations of Computer Science*, 1998.

[Tre97] L. Trevisan. When Hamming Meets Euclid: the Approximability of Geometric TSP and MST. *Symposium on Theory of Computing*, 1997.

A Formal proof

Let γ_i be the fraction of nodes in S which are in clusters of size i, for $i = 1, 2, \ldots$, in the clustering C'. Clearly $\sum_{i\geq 1} \gamma_i = 1$. Also

$$\sum_{i\geq 1} \frac{\gamma_i}{i} = 1/3 \tag{1}$$

since the total number of clusters is $m/3$. The cost of the clustering divided by m is at least $\gamma_2/2+4\gamma_3/3+\sum_{i\geq 4}(i-1)\gamma_i/2$, and thus we have

$$\frac{\gamma_2}{2}+\frac{4\gamma_3}{3}+\sum_{i\geq 4}\frac{i-1}{2}\gamma_i \ \leq\ 1+\delta \ =\sum_{i\geq 1}\gamma_i+\delta\,. \tag{2}$$

By adding (1) and (2) and rearranging the terms we get

$$\frac{2\gamma_3}{3}+\sum_{i\geq 4}\Big(\frac{i-3}{2}+\frac{1}{i}\Big)\gamma_i \ \leq\ \frac{1}{3}+\delta$$

which implies $2\gamma_3/3+3\gamma_4/4+\sum_{i\geq 5}\gamma_i \leq 1/3+\delta$, or $\gamma_1+\gamma_2+\gamma_3/3+\gamma_4/4 \geq 2/3-\delta$ (using $\sum_{i\geq 1}\gamma_i = 1$). Using (1) again, this implies that $\gamma_2/2 \geq (1/3-\delta)$, or in other words the number of 2-element clusters is at least $(1/3-\delta)m/3$. Therefore, there must be at most $\delta m/3$ clusters that cover the at least $m/3$ elements which are not in the 2-element clusters.

This means there must be at least $(1-7\delta)m/3$ elements which are covered by clusters of size 8 or more. The total cost incurred by the clustering of these elements is at least $(1-7\delta)m/3 \cdot 7/2 = \frac{7(1-7\delta)}{6}m$ and this quantity is more than $(1+\delta)m$ for $\delta < 1/55$. So the overall cost of C' is greater than $(1+\delta)m$, a contradiction. $\square$