Lower Bounds for Embedding Edit Distance into Normed Spaces

A. Andoni, MIT
M. Deza, ENS, France
A. Gupta, Bell Labs
P. Indyk, MIT
S. Raskhodnikova, MIT

1 Introduction

The edit distance (also called *Levenshtein metric*) between two strings is the minimum number of operations (insertions, deletions and character substitutions) needed to transform one string into another. This distance is of key importance in computational biology, as well as text processing and other areas. Algorithms for problems involving this metric have been extensively investigated. In particular, the quadratic-time dynamic programming algorithm for computing the edit distance between two strings is one of the most investigated and used algorithms in computational biology.

Recently, a new approach to problems involving edit distance has been proposed. Its basic component is construction of a mapping f (called an *embedding*), which maps any string s into a vector $f(s) \in \Re^d$, so that for any pair of strings s, s', the l_p distance $\|f(s)-f(s')\|_p$ is approximately equal to the edit distance between s and s'. The approximation factor is called *distortion* of the embedding f. A low-distortion embedding of edit distance into l_p norm would be very useful, for the following reasons:

- One could reduce a similarity search (e.g., nearest neighbor computation) in large sequence databases to an analogous problem in normed spaces; for the latter problem many efficient solutions are known.
- If computing $f(s)$ took subquadratic time, then the edit distance between two strings could be approximated in subquadratic time as well.

Unfortunately, so far essentially nothing is known about embeddability of the edit distance into a normed space[1]. If we modify the definition of the distance by allowing to move an arbitrarily long contiguous block of characters as a single operation, then the resulting *block-edit* metric can be embedded into l_1 with distortion $O(\log d \cdot \log^* d)$, where d is the length of the embedded strings (see [CPSV00, MS00, CM02] and references therein). This suggests that a similar result could be achievable for the (character) edit distance as well. So far, however, such a result seems quite elusive. This raises the possibility that a low-distortion embedding might be not possible.

In this paper we present the first non-trivial *lower bound* for embeddability of edit distance into l_p norms. In particular, we show that such metric cannot be embedded into l_1 norm (with arbitrary dimension) with distortion better than $3/2$. In fact, we show that the metric cannot be embedded with better distortion into the square of l_2; since any l_1-metric can be embedded isometrically into $(l_2)^2$ [LLR94], this implies the former result. We also show that for our approach, the factor $3/2$ is tight, even for embeddings into the l_1 norm.

Although our lower bound of $3/2$ is quite modest when compared with the current best upper bound of $O(d)$, it should be noted that proving lower bounds for embedding into l_1 norm is a quite difficult task. In fact, the only known technique for obtaining *super-constant* lower bounds, applicable to shortest path metrics over a graph (say G), is to show that G is an expander. We believe that this approach might not be applicable here, since our computational experiments for edit metrics over small length strings suggest that the graph underlying the edit metric is *not* a good expander. Instead our approach is to identify a subgraph of the edit metric for which the lower bound can be shown using direct arguments. In particular, we show that the edit metric contains the shortest path metric over the $K_{2,n}$ graph (we call it a $K_{2,n}$*-metric*) as an induced subgraph. Then we show that the latter metric cannot be embedded into the square of l_2 norm with low distortion.

2 "Hard" subsets of the edit metric

We show that the $K_{2,n}$-metric is an induced subgraph of the edit distance. If we label the vertices of $K_{2,n}$ with $A_1, A_2, B_1, B_2, ..., B_n$, such that there are edges only between A_i and B_j, $1 \le i \le 2, 1 \le j \le n$, then we can make the following correspondence:

- A_1 corresponds to the string $\underbrace{101010...10}_{2n \text{ bits}}$, where the block 10 repeats exactly n times;

[1] Except for generic embeddability results for *arbitrary* metric spaces, which do not seem to provide any interesting bounds.

- A_2 corresponds to the string $\underbrace{1010...10}_{2n-2 \text{ bits}}$, where the block 10 repeats exactly $n-1$ times;
- B_i corresponds to the string $\underbrace{1010...101011010...1010}_{2n-1 \text{ bits}}$, that is the string that corresponds to A_1 with the i^{th} zero bit deleted.

With these correspondences we have the following distances between the strings. The distance between strings corresponding to A_1 and A_2 is precisely 2: A_2 is obtained by performing 2 deletions on A_1 string. The distance between A_1 and any B_i is precisely 1 by definition of the B_i strings. The distance between any B_i and A_2 is also 1 because A_2 can be obtained by deleting the i^{th} 1. Finally, the distance between the strings corresponding to some B_i and some B_j, for $i \neq j$, is equal to 2 because these are two strings of the same lengths that differ in at least 2 positions.

3 Analytic lower bounds

THEOREM 3.1. *For any $\epsilon > 0$, there exists n, such that the distortion of any embedding f of $K_{2,n}$-metric into the square of the l_2 norm (and therefore into the l_1 norm as well) is at least $3/2 - \epsilon$.*

Proof: Take $n = 2k$, where $k = \lceil \frac{1/\epsilon}{2} \rceil$. Let c be the distortion of f. We use the notation $B_i = A_{i+2}$, for $i = -1, 0$. The metric over the set of points $X = f(B_{-1}), \dots f(B_n)$ needs to satisfy the following *negative type inequality* [DL97] for any sequence $b_{-1}, \dots b_n$ of integers which sum up to 0:

$$\sum_{-1 \le i < j \le n} b_i b_j \|f(B_i) - f(B_j)\|_2^2 \le 0$$

Let $b_{-1} = b_0 = -k$, and $b_i = 1$, $i = 1 \dots n$. We obtain

$$k^2 \cdot 2 + \binom{n}{2} \cdot 2 \le 2nk \cdot c$$

Therefore,

$$c \ge \frac{k^2 + (2k-1)k}{(2k)k} \ge 3/2 - \frac{1}{2k}.$$

The following theorem shows that the factor $3/2$ is indeed tight.

THEOREM 3.2. *There exists an embedding f of $K_{2,n}$-metric into l_1 with distortion $3/2$.*

Proof: The embedding f is obtained by using a combination of two different embeddings f_1 and f_2. The first embedding (into $l_1^{2^n}$) is defined as follows:

- $f_1(A_1) = (0, \dots, 0)$
- $f_1(A_2) = (1, \dots, 1)/2^n$
- $f_1(B_i) = (\text{bin}(1)_i, \dots, \text{bin}(2^n - 1)_i)/2^n$, where $\text{bin}(j)_i$ denotes the i-th bit of the binary representation of j.

CLAIM 1. *The embedding f_1 satisfies the following properties:*

1. $\|f_1(A_1) - f_1(A_2)\|_1 = 1$
2. $\|f_1(A_i) - f_1(B_j)\|_1 = 1/2$, *for* $i = 1, 2$ *and* $j = 1, \dots n$
3. $\|f_1(B_i) - f_1(B_j)\|_1 = 1/2$, *for* $1 \le i < j \le n$

The second embedding f_2 (into l_1^n) is defined as:

- $f_2(A_1) = f_2(A_2) = (0, \dots, 0)$
- $f_2(B_i) = e_i/2$, where e_i is a vector with 1 at the i-th position and zeros elsewhere

CLAIM 2. *The embedding f_2 satisfies the following properties:*

1. $\|f_2(A_1) - f_2(A_2)\|_1 = 0$
2. $\|f_2(A_i) - f_2(B_j)\|_1 = 1/2$, *for* $i = 1, 2$ *and* $j = 1, \dots n$
3. $\|f_2(B_i) - f_2(B_j)\|_1 = 1$, *for* $1 \le i < j \le n$

Let D_1 and D_2 be the metrics induced by f_1 and f_2. The metric $2D_1 + D_2$ provides the desired distortion.

We mention that by increasing the distortion to $3/2 + \epsilon$ one can reduce the dimension of the host space to only $O(\log n)$. The details are left to the full version of this paper.

The lower bound for the square of the l_2 norm can be also proved directly, without using the negative type inequality. We attach the alternative proof in the appendix.

4 Summary of computational experiments

As mentioned in the introduction, we have performed several computational experiments aimed at improving the lower bound of $3/2$. Specifically, our goal was to estimate the distortion for embedding of the edit metric over binary strings of length up to d into l_1 or square of l_2, for small values of d. In particular, we considered the following three approaches:

1. Optimal distortion embeddings into l_1 using the cut-metric formulation of the l_1 norm

2. Optimal distortion embeddings into square of l_2 using semi-definite programming

3. Lower bounds for distortion via expansion properties of the metric

Embedding into l_1. It is known (e.g., see Proposition 4.2.2 in [DL76]) that a metric $M = (X, D)$ can be embedded into l_1 iff it can be represented as a conic combination of cut semi-metrics, i.e., semi-metrics D_S, $S \subset X$, such that $D_S(p,q) = |S \cap \{p,q\}| \; mod \; 2$. Finding best distortion embedding of M into l_1 can be formulated as a linear program with $2^{|X|-1}$ variables and $O(|X|^2)$ constraints. Unfortunately, in our case $|X| = 2^{d+1} - 1$, which made this approach infeasible for $d > 3$. Thus, we experimented only with $d = 3$. The resulting distortion was $4/3$, which is less than our earlier guarantee.

Embedding into square of l_2. The optimal distortion of an embedding of a metric $M = (X, D)$ into the square of l_2 can be computed in polynomial time using semidefinite programming (e.g., see [Mat]). Thus, we computed best distortion embeddings of the edit metrics for strings of lengths up to $d = 5$. For this purpose, we used Matlab-based package, called SDPpack. The optimal distortion was roughly 1.30, which is less than the analytic bound of $3/2$ proved earlier in this paper. Applying the embedding procedure for $d = 6$ turned out to be infeasible, since (by our estimations) it would require about $2GB$ of memory.

Lower bounds via expansion. Our final attempt to obtain computational lower bounds for embeddings of edit distance was to show that the "edit graph" underlying the edit metric is a "good" expander and then use the bounds as in, e.g., [Mat]. In particular, we considered $G = G_{d,d-1} = (V, E)$, which is the edit graph induced by strings of length d or $d-1$, with a few additional "self-loop" edges to make it regular (with degree $\Delta = 3d - 1$). It is not difficult to see that the shortest path metric over G is an induced subgraph of the edit metric. Our goal was to show that there exists a (large) constant C such that for any set $A \subset V$ we have $|e(A, V-A)| \geq C|A||V-A|/n$, where $e(A, B)$ is the set of edges between A and B. Then it would follow [Mat] that the minimum distortion c is at least

$$S \cdot \frac{C \cdot \mathrm{avg}(G)}{\Delta}$$

where S is a constant scaling factor and $\mathrm{avg}(G)$ is the average distance between pairs of nodes in G. The expansion constant C can be bounded from below by the "eigenvalue gap", i.e., the difference between the first eigenvalue (equal to Δ) and the second eigenvalue of the adjacency matrix of G. The eigenvalues of G can be computed efficiently using power method, which requires much less space than earlier methods (notably SDP).

Unfortunately, the eigenvalue gap of G has refused to be large. Specifically, it was about 2.7 for $d = 4, 8, 12, 16$. For comparison, it is equal to 2 for a d-dimensional hypercube H,which is clearly embeddable into l_1 with no distortion. Since $\mathrm{avg}(H) = d/2$, $\Delta(H) = d$, it follows that $S \leq 1$. Since $\mathrm{avg}(G) \leq d/2$ and $\Delta(G) = 3d - 1$, the resulting distortion lower bound is weaker that $3/2$.

References

[CM02] G. Cormode and S. Muthukrishnan. The string edit distance matching problem with moves. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2002.

[CPSV00] G. Cormode, M. Paterson, C. Sahinalp, and U. Vishkin. Communication complexity of document exchange. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2000.

[DL76] D. Dobkin and R. Lipton. Multidimensional search problems. *SIAM Journal on Computing*, 5:181–186, 1976.

[DL97] M. M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Algorithms and Combinatorics 15. Springer-Verlag, Berlin etc., 1997.

[LLR94] 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.

[Mat] J. Matoušek. Lectures on discrete geometry. *Springer, in press.*

[MS00] S. Muthukrishnan and C. Sahinalp. Approximate nearest neighbors and sequence comparison with block operations. *Proceedings of the Symposium on Theory of Computing*, 2000.

A Another proof for the 3/2 lower bound for embeddings into $(l_2)^2$

Proof: We first provide a natural embedding of $K_{2,n}$ metric into the square of l_2, with distortion 3/2. Although this result is implied by the earlier theorem, the construction provides a good intuition for the proof of the lower bound.

An embedding with distortion 3/2 can be obtained as follows. Consider the following vectors $x_1, x_2, ... x_n$, such that $x_i = (x_{i1}, x_{i2} ...)$.

$$\begin{aligned} x_{n+1} &= (\frac{\sqrt{2}}{2}, 0, 0, ...) \\ x_{n+2} &= (-\frac{\sqrt{2}}{2}, 0, 0, ...) \\ x_{ij} &= \delta(i+1, j) \end{aligned}$$

where $1 \leq i \leq n, j \geq 1$ and δ is Kronecker's symbol. Then, we have for $1 \leq i, j \leq n, i \neq j$:

$$\begin{array}{lcl} 2 \leq & \|x_{n+1} - x_{n+2}\|^2 = \left(2 * \frac{\sqrt{2}}{2}\right)^2 = 2 & \leq 3/2 * 2 \\ 2 \leq & \|x_i - x_j\|^2 = 1 + 1 = 2 & \leq 3/2 * 2 \\ 1 \leq & \|x_{n+1} - x_i\|^2 = \frac{1}{2} + 1 = 3/2 & \leq 3/2 * 1 \\ 1 \leq & \|x_{n+2} - x_i\|^2 = \frac{1}{2} + 1 = 3/2 & \leq 3/2 * 1 \end{array}$$

Now we proceed with the lower bound. The proof is by contradiction. Suppose the statement is false, i.e., there exists an $\epsilon > 0$ such that for any n, $K_{2,n}$-metric can be embedded with distortion at most $3/2 - \epsilon$. As before, we label the $2+n$ vertices of $K_{2,n}$ as $A_1, A_2, B_1, B_2, ..., B_n$, where there edges are only between the vertices A_i and B_j, $1 \leq i \leq 2, 1 \leq j \leq n$.

Let $x_1, x_2, ..., x_n, x_{n+1}, x_{n+2}$ be the corresponding vectors of $B_1, B_2, ..., B_n$ and A_1, A_2, such that the distortion is at most $3/2 - \epsilon$. This means that

$$\sigma \leq \|x_i - x_j\|^2 = \|x_i - x_j\|^2 \leq (3/2 - \epsilon)\sigma \qquad (1)$$

with $\sigma = 2$ if $1 \leq i < j \leq n$, or $\{i, j\} = \{n+1, n+2\}$ and $\sigma = 1$ otherwise.

We can rotate the vectors $x_1, x_2, ..., x_n, x_{n+1}, x_{n+2}$ so that $x_{n+1} = (a, 0, 0...)$ and $x_{n+2} = (-a, 0, 0...)$. From (1), we have that

$$\begin{aligned} 2 \quad &\leq \quad \|x_{n+1} - x_{n+2}\|^2 = \|(a, 0, 0...) - (-a, 0, 0...)\|^2 \\ &= \quad (2a)^2 \\ \Rightarrow a \quad &\geq \sqrt{2}/2 \end{aligned}$$

We have also

$$\begin{aligned} \|x_{n+1} - x_i\|^2 &= (a - x_{i1})^2 + (\|x_i\|^2 - x_{i1}^2) \qquad (2) \\ &= a^2 - 2ax_{i1} + \|x_i\|^2 \end{aligned}$$

for $1 \leq i \leq n$, and

$$\begin{aligned} \|x_{n+2} - x_i\|^2 &= (-a - x_{i1})^2 + (\|x_i\|^2 - x_{i1}^2) \qquad (3) \\ &= a^2 + 2ax_{i1} + \|x_i\|^2 \end{aligned}$$

for $1 \leq i \leq n$.

We can construct n new vectors $x'_1, x'_2, ..., x'_n$ from $x_1, x_2, ..., x_n$, such that the new vectors still satisfy (1), and have their first coordinates equal to zero. Specifically, the new vectors are constructed by adding a new 0 coordinate in front of the vectors, i.e.,

$$x'_i = (0, x_{i1}, x_{i2}, x_{i3}, ...)$$

Due to (2) and (3), the new vectors still satisfy equation (1). Therefore, from now on we assume vectors $x_1, x_2, ..., x_n$ have their first coordinate set to 0.

In this case

$$\|x_{n+1} - x_i\|^2 = \|x_{n+2} - x_i\|^2 = a^2 + \|x_i\|^2 \geq \frac{1}{2} + \|x_i\|^2$$

It follows that

$$(3/2 - \epsilon) \geq \|x_{n+1} - x_i\|^2 \geq \frac{1}{2} + \|x_i\|^2 \Rightarrow \|x_i\|^2 \leq (1 - \epsilon) \qquad (4)$$

Consider now the distance between two vectors x_i and x_j with $1 \leq i < j \leq n$:

$$\begin{aligned} \|x_i - x_j\|^2 &= \|x_i\|^2 + \|x_j\|^2 - 2x_i \cdot x_j \\ &\leq (1 - \epsilon) + (1 - \epsilon) + 2x_i \cdot x_j \end{aligned}$$

From (1), we have that $\|x_i - x_j\|^2 \geq 2$ for $1 \leq i < j \leq n$. Therefore, we have

$$\begin{aligned} 2 \leq \|x_i - x_j\|^2 \leq 2 - 2\epsilon - 2x_i \cdot x_j \quad &\Rightarrow \\ x_i \cdot x_j \leq -\epsilon \end{aligned}$$

Concluding, we have that there exists ϵ such that for any n, there exist n vectors $x_1, x_2, ..., x_n$ with norm less than 1 such that $x_i \cdot x_j \leq -\epsilon$. But this is not true, since it is known that if $n \geq 2/\epsilon + 1$, then there exist two distinct vectors x_i and x_j such that $x_i \cdot x_j > -\epsilon$. Instead of the reference, we attach an easy proof (provided to us by Venkat Guruswami):

$$\begin{aligned} 0 &\leq \|\sum_i x_i\|^2 \\ &= (\sum_i x_i) \cdot (\sum_i x_i) \\ &= \sum_i \|x_i\|^2 + 2 \sum_{i<j} x_i \cdot x_j \\ &\leq n - \epsilon 2 \binom{n}{2} \end{aligned}$$