Totally Bounded Metric Spaces, Their Model Theoretic Stability and Similarity Detecting Algorithms

Sági, Gábor and Al-Sabti, Karrar

Abstract: ***Many questions of theoretical computer science can be reduced to the following problem: let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space, let $A \subseteq X$ and let ε be a positive real number; for a given input $x \in X$ find $a \in A$ (if any) for which $\varrho(a, x) \leq \varepsilon$. This problem is called the similarity detecting problem of $(\mathcal{X}, A, \varepsilon)$. Usually, A (or sometimes X) is finite but huge, and the challenge is to represent the metric space in such a way computer algorithms may handle it efficiently.***

This paper consists of two main parts. The first part has a theoretical character: we investigate model theoretic properties of certain metric spaces. In Theorem 2.4 we show, that relational structures associated to totally bounded metric spaces have some stability properties in the model theoretic sense (all relevant definitions will be recalled in the paper). This result allows us to build some metric spaces from their finite subspaces.

The second part of the paper is application oriented. Based on the first part and on some results of [9] in the second part we propose a similarity detecting algorithm. We associate a finite dimensional Euclidean space $\mathcal{Y}$ to a totally bounded metric space $\mathcal{X}$ and an "almost isometry" $f : X \to Y$ which preserve distances modulo a controlled amount of inaccuracy. After that, instead of working with $\mathcal{X}$, we can work with $\mathcal{Y}$. The main result of this part is the description of the above method.

In the special case, when $\mathcal{X}$ itself is a large dimensional Euclidean space (with its usual Euclidean metric), our method can be considered as a kind of dimension reduction. In this special case we are analyzing the time complexity of our proposed algorithm, as well.

Index Terms: ***Totally bounded metric spaces, similarity detecting algorithms, dimension reduction.***

1. Introduction

The present work has a practical and a theoretical motivation. We start by the practical motivation.

Many questions of theoretical computer science can be reduced to questions about certain metric spaces, for further details we refer to [2], [6] and the references therein. Usually, these spaces are finite, but huge and the problem is how to handle these spaces by computer algorithms in a tractable way. This is the case, for example, if the distance function of the metric space measures "similarity" of two objects and the problem is to find the elements of a database which are similar enough to a given input. Related problems can be effectively solved if one is able to represent the metric space in a suitable way, for example, if one is able to embed the metric space into a finite dimensional Euclidean space (endowed with its usual metric) with a function which is an "almost isometry", or if one is able to embed a compact subset of a (finite dimensional) Euclidean space having large dimension into a considerably smaller dimensional Euclidean space by an "almost isometry".

For the theoretical background we briefly recall investigations initiated in [10] and continued in [9]. For metric spaces $\mathcal{X} = \langle X, \varrho \rangle$, $\mathcal{Y} = \langle Y, \sigma \rangle$ and a positive real number ε, a function $f : X \to Y$ is defined to be an ε-map iff for all $y \in Y$, the diameter of $f^{-1}(y)$ is at most ε. Thus, if ε is small, then f is almost injective. In Theorem 10 of [9] the first author gave a new proof for the following well-known fact: if $\mathcal{X}$ is totally bounded (for further explanation see Definition 2.1 and the sentence immediately after it), then for all ε there exists a finite number n and a continuous ε-map $f_\varepsilon : X \to \mathbb{R}^n$, where $\mathbb{R}^n$ is the usual n-dimensional Euclidean space endowed with the Euclidean metric. Such ε-maps still exist even if $\mathcal{X}$ has infinite covering dimension (in this case, n depends on ε, of course). Contrary to the previously known proofs (see e.g. Chapter 8 of [8]), the proof technique in [9] is effective in the sense, that it allows to establish estimations for n in terms of ε and structural properties of $\mathcal{X}$.

Now we turn to the theoretical part of this paper. There is a well known method (which will be recalled in Section 2 below) that associates a first order relational structure to a metric space. This

Manuscript received April 29, 2020. This work was supported by the Hungarian National Foundation for Scientific Research grant K129211.

G. Sági (contact person) Alfréd Rényi Institute of Mathematics, Reáltanoda u. 13-15, H-1053 Budapest, Hungary and Budapest University of Technology and Economics, Department of Algebra, Egry J. u. 1, H-1111 Budapest, Hungary (e-mail: sagi@renyi.hu).

K.Al-Sabti Budapest University of Technology and Economics, Department of Algebra, Egry J. u. 1, H-1111 Budapest, Hungary and University of Kufa, Faculty of Computer Science and Mathematics (e-mail: karrard.alsabti@uokufa.edu.iq).

method allows us to translate questions about metric spaces into questions about relational structures and these translated questions may be investigated by techniques of first order logic and model theory. From the results of [9] one can easily obtain that if $\mathcal{X}$ is a totally bounded metric space, then it's associated relational structure $\mathcal{A}(\mathcal{X})$ is Δ-stable in the model theoretic sense. We will reconstruct the proof of this fact below in Theorem 2.3. The main goal of section 2 is to investigate the converse, that is, in Theorem 2.4 we will show, that if $\mathcal{X}$ is dense in itself, ϱ is bounded, and $\mathcal{A}(\mathcal{X})$ is Δ-stable in a rather strong sense, then $\mathcal{X}$ is totally bounded.

On the basis of this background, in Section 3 we are proposing a method which can be used for similarity detecting, clustering and related problems. The structure of this paper is as follows. At the end of this section we are summing up our system of notation. Section 2 is devoted to theoretical investigations. Here the main result is Theorem 2.4 which provides a characterization for totally bounded metric spaces in terms of stability properties of their associated relational structures. In Section 3 we describe and analyze our similarity detecting algorithm with special emphasis for the case, when $\mathcal{X}$ itself is a large dimensional Euclidean space. Finally in Section 4 we provide some conclusion.

Notation

Our notation is mostly standard, but the following explanation may help.

Throughout $\mathbb{N}$ denotes the set of natural numbers. In addition, $\mathbb{R}$ and $\mathbb{R}^+$ denotes the set of real numbers, and the set of positive real numbers, respectively.

Let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space, $a \in X$ and let γ be a non-negative real number. As usual, the open γ-ball $B(\gamma, a)$ at a is the set

$$B(\gamma, a) = \{x \in X : \varrho(a, x) < \gamma\}.$$

If $\mathcal{E}$ is a Euclidean space, then the norm of any element x of $\mathcal{E}$ will be denoted by $||x||$ (the norm function of $\mathcal{E}$ and its usual metric are mutually definable from each other in the usual way).

2. Totally Bounded Metric Spaces and Model Tiheoretic Stability

Let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space. As it is well known, one can associate a relational structure $\mathcal{A}(\mathcal{X})$ to $\mathcal{X}$ in the following way. If d is a distance of $\mathcal{X}$, that is, $d \in ran(\varrho)$ then the binary relation R_d is defined to be

$$R_d = \{\langle a, b \rangle \in X^2 : \varrho(a, b) \leq d\}.$$

Thus, the relational structure

$$\mathcal{A}(\mathcal{X}) := \langle X, R_d \rangle_{d \in ran(\varrho)}$$

completely describes $\mathcal{X}$ in the sense, that $\mathcal{X}$ and $\mathcal{A}(\mathcal{X})$ are mutually definable from each other. Further, at the same time, $\mathcal{A}(\mathcal{X})$ can be treated as a model for an appropriate first order language.

Recall e.g. from [1], that the metric space $\mathcal{X}$ is defined to be *totally bounded* if and only if for all positive $\varepsilon \in \mathbb{R}$ there exists a finite family of ε-balls of $\mathcal{X}$ that covers X. For completeness, we recall the (rather standard) formal definitions below.

Definition 2.1: Let $\gamma \in \mathbb{R}^+$. A family $\{B_i : i \in I\}$ of γ-balls is defined to be a γ-net iff it covers X, that is,

$$X = \bigcup_{i \in I} B_i.$$

Using this terminology, a metric space $\mathcal{X}$ is defined to be totally bounded iff for all positive $\gamma \in \mathbb{R}$ there exists a finite γ-net in $\mathcal{X}$.

As it is well known, $\mathcal{X}$ is compact iff it's metric is totally bounded and complete (i.e. every Cauchy sequence is convergent in $\mathcal{X}$). For further details we refer to [1], as well. It is well known, that every finite metric space is compact.

Let $\mathcal{A} = \langle A, R_i \rangle_{i \in I}$ be any first order structure, let $X \subseteq Y \subseteq A$ be arbitrary (finite or infinite) and let Δ_1, Δ_2 be sets of first order formulas in the language of $\mathcal{A}$. As usual, $S(Y)$ denotes the set of types over Y and v denotes the unique free variable of formulas in elements of $S(Y)$. For further explanation for the notation we refer e.g. to [13]. According to Definition I.2.6 of [13], a type $p \in S(Y)$ is (Δ_1, Δ_2)-splitting over X iff there exist $\overline{b}, \overline{c} \in Y$ and $\varphi \in \Delta_2$ such that

$$tp_{\Delta_1}(\overline{b}/X) = tp_{\Delta_1}(\overline{c}/X)$$

but

$$\varphi(v, \overline{b}),\ \neg\varphi(v, \overline{c}) \in p.$$

This motivates the following "approximate version" of splitting in the context of metric spaces which we recall from [9] (see also [10]).

Definition 2.2: Let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space, let $A \subseteq B \subseteq X$ be arbitrary (finite or infinite), let p be a Δ-type over B in $\mathcal{A}(\mathcal{X})$ and let ε, δ be non-negative real numbers. Then we say, that p is (ε, δ)-splitting over A iff there exist $c_0, c_1 \in B$ such that for all $a \in A$ we have

$$|\varrho(a, c_0) - \varrho(a, c_1)| < \delta$$

but whenever b realizes p, we have

$$|\varrho(b, c_0) - \varrho(b, c_1)| \geq \varepsilon.$$

Keeping the notation introduced in the above definition, intuitively $p = tp_{\Delta}^{\mathcal{A}(\mathcal{X})}(b/B)$ is (ε, δ)-splitting over A if and only if there exists $c_0, c_1 \in A$ such that c_0 and c_1 are "indiscernible from the viewpoint of A modulo δ", but b "distinguishes

them modulo ε".

Jumping back to model theory, let κ be a cardinal. An increasing sequence of types $\langle p_i : i < \kappa \rangle$ is defined to be a splitting chain iff for all $i < \kappa$, p_{i+1} is splitting over the domain of p_i. Further, by Lemma I.2.7 of [13], if a first order theory T is λ-stable for some infinite cardinal λ, then in each model of T, the length of any splitting chain of types is smaller than λ. An easy compactness argument yields, that if $\mathcal{A}$ is a stable structure, and Δ is a finite set of formulas, then the length of each (Δ, Δ)-splitting chain of Δ-types in $\mathcal{A}$, is finite.

In the context of metric spaces, the following analogous result has been established in Theorem 5 of [9]: if $\mathcal{X}$ is a totally bounded metric space and Δ is the set of atomic formulas of the language of $\mathcal{A}(\mathcal{X})$ then for all $\varepsilon \in \mathbb{R}^+$ there exist $\delta \in \mathbb{R}^+$ and $N \in \mathbb{N}$ such that if $\langle p_i, i < m \rangle$ is a strictly increasing sequence of (ε, δ)-splitting Δ-types in $\mathcal{A}(\mathcal{X})$, then $m \leq N$; particularly, each (ε, δ)-splitting sequence of Δ-types in $\mathcal{A}(\mathcal{X})$ has finite length (in fact, N is a common upper bound for them). Thus, if $\mathcal{X}$ is totally bounded, then $\mathcal{A}(\mathcal{X})$ shows some stability properties. In fact, according to the next theorem, in this case, $\mathcal{A}(\mathcal{X})$ is Δ-stable. However, to state and prove the next theorem, we need to recall the following definition from [9]:

According to Definition 6 of [9], if $a \in X$ and $\varepsilon, \delta \in \mathbb{R}^+$ then $A \subseteq X$ is defined to be an (ε, δ)-basis for a iff for any $B \subseteq X - \{a\}$ with $A \subseteq B$, the type

$$tp^{\mathcal{X}}(a/B)$$

does not (ε, δ)-split over A.

Theorem 2.3: If $\mathcal{X} = \langle X, \varrho \rangle$ is a totally bounded metric space, then

(1) For all $\varepsilon \in \mathbb{R}^+$ there exist $\delta \in \mathbb{R}^+$ and a finite set $A \subseteq X$ such that A is an (ε, δ)-basis for all $a \in X$ (we emphasize, that A does not depend on a).

(2) $\mathcal{A}(\mathcal{X})$ is Δ-stable.

Proof. To show (1), assume $\mathcal{X} = \langle X, \varrho \rangle$ is a totally bounded metric space.Then (1) is the same, as Theorem 9 [9].

To show (2), for all $n \in \mathbb{N}^+$ choose δ_n and a finite A_n such that A_n is an $(\frac{1}{n}, \delta_n)$-basis for all $a \in X$ (according to the previous paragraph, such δ_n and A_n exist). Now let $\mathcal{A}$ be any elementary extension of $\mathcal{A}(\mathcal{X})$ and let $Y \subseteq A$ with $|Y| \leq 2^{\aleph_0}$. We shall show, that there are at most $2^{\aleph_0}$ many Δ-types over Y in $\mathcal{A}$. Enlarging Y if necessary, we may assume $A_n \subseteq Y$ holds for all $n \in \mathbb{N}$. Since $\mathcal{A}$ is an elementary extension of $\mathcal{A}(\mathcal{X})$, it follows, that for any $a \in A$ and $Y \subseteq B \subseteq A - \{a\}$, the type

$$tp^{\mathcal{A}}(a/B)$$

does not $(\frac{1}{n}, \delta_n)$-split over Y (because for a fixed n, $(\frac{1}{n}, \delta_n)$-splitting is first order expressible in the language of $\mathcal{A}(\mathcal{X})$). As $A^* := \cup_{n \in \mathbb{N}} A_n$ is countable, there are at most $2^{\aleph_0}$ many types over A^*. Hence, it is enough to show, that for $a, b \in A - Y$,

$$\text{if } tp_\Delta(a/A^*) = tp_\Delta(b/A^*) \text{ then } tp_\Delta(a/Y) = tp_\Delta(b/Y).$$

Assume, seeking a contradiction, that $tp_\Delta(a/A^*) = tp_\Delta(b/A^*)$, but $tp_\Delta(a/Y) \neq tp_\Delta(b/Y)$. Then, there exists $c \in Y$ such that $\varrho(a, c) \neq \varrho(b, c)$. Fix $n \in \mathbb{N}^+$ with $\frac{1}{n} < |\varrho(a, c) - \varrho(b, c)|$. But this is impossible, because A_n (and hence A^* as well) is an $(\frac{1}{n}, \delta_n)$ basis for c. This contradiction completes the proof. ∎

Now we turn to show a kind of weak converse of the above theorem, which is the main result of this section (and the main theoretical result of the paper).

Theorem 2.4: Let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space such that $\mathcal{X}$ is dense in itself and the range of ϱ is bounded. Then the following are equivalent.

(1) $\mathcal{X} = \langle X, \varrho \rangle$ is totally bounded;

(2) $\mathcal{A}(\mathcal{X})$ is Δ-stable in the following, strong sense: for all $\varepsilon \in \mathbb{R}^+$ there exist $\delta \in \mathbb{R}^+$ and a finite set $A_{\varepsilon,\delta} \subseteq X$ such that

$$(*) \quad A_{\varepsilon,\delta} \text{ is an } (\varepsilon, \delta)\text{-basis for all } a \in X.$$

Proof. First we note that (1) implies (2) by Theorem 2.3 (we also note, that in this direction we don't use the assumptions that $\mathcal{X}$ is dense in itself and the range of ϱ is bounded).

To prove the converse, assume (2). Let $\varepsilon \in \mathbb{R}^+$ be arbitrary; we shall show, that there exists a finite 3ε-net in $\mathcal{X}$. Choose δ and $A_{\varepsilon,\delta}$ that satisfy (2). By assumption, the range of ϱ is bounded, say C is an upper bound for it. Consider the real $[0, C]$ interval and its $|A_{\varepsilon,\delta}|^{th}$ power

$$\mathcal{Y} := [0, C]^{|A_{\varepsilon,\delta}|}$$

as a subspace of $\mathbb{R}^{|A_{\varepsilon,\delta}|}$. Clearly, $\mathcal{Y}$ is a compact space, hence it has a finite δ-net

$$\{B(u_i, \delta) : \ i \in I\}.$$

For $c \in X$ define $t(c) \in \mathbb{R}^{A_{\varepsilon,\delta}}$ to be the vector

$$t(c) = \langle \varrho(c, x) : x \in A_{\varepsilon,\delta} \rangle.$$

Let

$$I' = \{i \in I : (\exists a \in X) t(a) \in B(u_i, \delta)\}$$

and for all $i \in I'$ choose $a_i \in X$ so that

$$t(a_i) \in B(u_i, \delta).$$

Finally, let $B = \{a_i : i \in I'\}$. Clearly, B is finite. Further, for any $c \in X$ there exists $i \in I$ such that

$$t(c) \in B(u_i, \delta).$$

Then c witnesses $i \in I'$ and in addition,

$$||t(c) - t(a_i)||_2 < \delta.$$

It follows, that for all $a \in A_{\varepsilon,\delta}$ we have

$$|\varrho(a, c) - \varrho(a_i, c)| < \delta.$$

Summing up we have shown, that for all $c \in X$ there exists $b \in B$ such that for all $a \in A_{\varepsilon,\delta}$ we have

$$(**) \quad |\varrho(a, c) - \varrho(b, c)| < \delta.$$

We claim, that

$$\{B(b, 3\varepsilon) : b \in B\}$$

covers X, that is, it is a finite 3ε-net, as desired. To see this, let $c \in X$ be arbitrary. Then, there exists $b \in B$ such that $(**)$ holds. Further, as $\mathcal{X}$ is dense in itself, there exists $d \in X$ with $\varrho(c, d) < \varepsilon$. By $(*)$, $A_{\varepsilon,\delta}$ is an (ε, δ)-basis for d, hence

$$|\varrho(c, d) - \varrho(b, d)| < \varepsilon,$$

particularly, $\varrho(b, d) < 2\varepsilon$. But then,

$$\varrho(c, b) \leq \varrho(c, d) + \varrho(d, b) < 3\varepsilon,$$

that is, $c \in B(b, 3\varepsilon)$. Consequently,

$$\{B(b, 3\varepsilon) : b \in B\}$$

is a finite 3ε-net. As ε was arbitrary, this completes the proof. ∎

3. A Similarity Detecting Algorithm

In this section we provide an application inspired by the theoretical results presented in the previous section. This section is based on the investigations initiated in [11].

Let $\mathcal{X} = \langle X, \varrho \rangle$ be a metric space and let $A \subseteq X$ be a given set. We have the following intuitive picture in our mind: X is the set of all instances of an abstract data type and ϱ measures similarity between the elements of X: if $\varrho(x, y)$ is "small" for some $x, y \in X$ then we say, that x and y are "similar enough" to each other. More concretely, we fix $\varepsilon \in \mathbb{R}^+$ and consider it as an amount of inaccuracy one can tolerate. Then "x and y are similar enough" means $\varrho(x, y) < \varepsilon$.

More formally, the similarity detecting problem for $(\mathcal{X}, A, \varepsilon)$ is the following: given an input $x \in X$ find $a \in A$ such that $\varrho(x, a) < \varepsilon$. The problem is, that A may be huge and computing ϱ for two particular points may be slow.

As we mentioned, our goal in this section is to propose and analyze an algorithm that can be used to handle the above problem efficiently. To do so, we start by recalling the following notation.

If $\mathcal{X}$ is a totally bounded metric space, then $\nu(\mathcal{X}, \gamma)$ denotes the smallest cardinality κ for which there exists a κ-sized γ-net of $\mathcal{X}$.

For a given $\varepsilon \in \mathbb{R}^+$ let

$$N := 6 \cdot \nu(\mathcal{X}, \frac{\varepsilon}{30}) \cdot \nu(\mathcal{X}, \frac{\varepsilon}{12}).$$

Suppose $\mathcal{X}$ is a (finite or infinite) totally bounded metric space. Then, according to Theorem 10 of [9],

(1) if $\mathcal{X}$ does not contain isolated points, then there exist $n \leq N$ and an ε-map $f : X \to \mathbb{R}^n$ such that, for all $x, y \in X$ we have

$$||f(x) - f(y)|| \leq \sqrt{n}\varrho(x, y),$$

in particular, f is continuous.

(2) if $\mathcal{X}$ has countably many isolated points, then there exist $n \leq 1 + N$ and a continuous ε-map $f : X \to \mathbb{R}^n$.

(3) if $\mathcal{X}$ is compact, then there exist $n \leq 1 + N$ and a continuous ε-map $f : X \to \mathbb{R}^n$,

further the ε-map is effectively constructed in all cases above.

Now we can sketch our similarity detecting algorithm as follows (immediately after sketching the algorithm, we comment and further explain its crucial steps).

(Step 1) For a given $\varepsilon' \in \mathbb{R}^+$ find $n \in \mathbb{N}$ and a continuous ε'-map $f : X \to \mathbb{R}^n$;

(Step 2) compute $B := \{\langle f(a), a \rangle : a \in A\}$;

(Step 3) for an input $x \in X$ find $b =: \langle u, v \rangle \in B$ for which the usual Euclidean distance $||f(x) - u||$ is minimal;

(Step 4) if (for $b = \langle u, v \rangle$ computed in (Step 3) above) we have

$$||f(x) - u|| < \varepsilon',$$

then the output is v; otherwise there are no elements of A which are similar enough to x.

In (Step 1) above, n and f can be constructed in an algorithmic way, because in the proof of Theorem 10 of [9], the ε-maps have been constructed effectively for all ε. For (Step 2) we note, that

$$\{x : \exists y \langle x, y \rangle \in B\} = \{f(a) : a \in A\} \subseteq \mathbb{R}^n.$$

So, in (Step 3), instead of working with the original distance ϱ we are working with Euclidean distances that can be computed relatively quickly, provided that n is small enough.

A particularly important special case of the general similarity detecting problem is, when X is a subset of a large dimensional Euclidean space (and ϱ is the corresponding Euclidean distance), that is, if $X \subseteq \mathbb{R}^k$ for a large k. This special case will be called *dimension reduction*. We make the following remarks:

- Note, that (Step 1) and (Step 2) in the above

sketch are preparatory: they should be performed only once at the beginning; if we are searching similar objects many times, then the cost of (Step 3) will dominate.

• If the database A changes in time, then according to (Step 2) we can quickly modify B, as well.

Now we turn to study the dimension reduction problem, that is, the similarity detecting problem, when A is a subspace of a large dimensional Euclidean space $\mathcal{X} = \mathbb{R}^k$. Of course, the critical point in our algorithm is (Step 1) in which we have to find n and f. According to Theorem 10 of [9] they exist and f can be effectively constructed from n and from given $(\frac{\varepsilon}{30})$-nets and $(\frac{\varepsilon}{12})$-nets of A. We have to choose n as small as we can in order to make our algorithm more efficient. The rest of the present section is devoted to investigate the choice of n for dimension reduction.

Analyzing the proof of Theorem 10 of [9], one concludes, that n becomes smaller whenever one is able to find smaller $(\frac{\varepsilon}{30})$-nets and $(\frac{\varepsilon}{12})$-nets of A. More generally, for given $\delta \in \mathbb{R}^+$ and $k \in \mathbb{N}$ one has to find a k-sized δ-net of A (if such exists).

The natural approach would be, that for a given δ one tries to find δ-nets of A with as small cardinalities k as possible. However, this do not would be an efficient method, as such an approach would be equivalent to solve NP-hard problems of computational geometry, and cluster analysis; for further details in that direction we refer to [7] and the references therein. Instead, we propose to fix k and estimate δ for which there exists a k-sized δ-net in A; then consecutively increasing k we will obtain a decreasing sequence of the corresponding $\delta = \delta_k$ and we increase k until δ_k will be sufficiently small.

For a fixed k and a given finite $A \subseteq \mathbb{R}^m$ the well known k-center problem is to find a k-element subset $B \subseteq A$ such that

$$max_a min_b\{||b-a|| : a \in A\}$$

is as small as possible (for further details we refer e.g. to [5]). This problem is related to cluster analysis and to the inverse shortest path problem. According to [3], the k-center problem (already in the Euclidean plane $\mathbb{R}^2$) is known to be NP-complete as well (for more recent related investigations we refer to [12] and [14]). Hence, instead of exact solutions, it is more practical to search for suboptimal, approximate solutions. Indeed, there is a classical approximate solution for the k-center problem due to Gonzalez [4]; we will briefly recall this.

For a metric space $\mathcal{X} = \langle X, \varrho \rangle$ and $A \subseteq X, b \in X$ the standard definition of the distance of A and b is

$$\varrho(A,b) = inf\{\varrho(a,b) : a \in A\}.$$

Then the farthest path transversal sequence of a finite metric space is defined as follows: $a_0 \in X$ is arbitrary, and if a_j has already been defined for all $j < i$, then a_i is a point x in X for which

$$\varrho(\{a_j : j < i\}, x)$$

is maximal. Now we recall Gonzalez's idea presented in [4]: fix $k \in \mathbb{N}$ and let

$$r = \varrho(\{a_j : j < k\}, a_k).$$

Observe, that

(i) for any $i \neq j < k$ we have $\varrho(a_i, a_j) \geq r$

and

(ii) for any $x \in X$ there exists $i < k$ such that $\varrho(a_i, x) \leq r$.

By (ii), $\{B(a_i, r) : i < k\}$ is an r-net of size k. Further, Suppose $\{B(b_i, r') : i < k\}$ is another r'-net. Then, by the Pigeon-Hole principle, there would exist $i \neq j \leq k$ and $l < k$ such that $a_i, a_j \in B(b_l, r')$, therefore

$$r \leq \varrho(a_i, a_j) \leq 2 \cdot r'.$$

In another words, the minimal radius r' of a k-sized r'-net is at least $\frac{r}{2}$. Therefore r constructed above, is a 2-approximation of the minimal radius of a k-sized net.

Based on the above observations, our dimension reducing algorithm is built up from an initialization part and from a searching part; these may be summarized as follows (as before, comments and explanations will be provided immediately after describing these algorithms).

Initializing part.

Input: a finite set $A \subseteq \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}^+$.

1. Choose an arbitrary $a_0 \in A$ and let $k = 1, \varepsilon' = 1 + \varepsilon$.
2. While $\varepsilon' > \varepsilon$ and $k < n$ Do
3. Let $r = max_{x \in A} \varrho(\{a_j : j < k\}, x)$.
4. Let $a_k \in A$ be such that $r = \varrho(\{a_j : j < k\}, a_k)$.
5. Let $\varepsilon' := 2r$.
6. Let $k = k + 1$.
7. End(Do).
8. Let $f : \mathbb{R}^n \to \mathbb{R}^k$ be the function that maps each $x \in \mathbb{R}^n$ onto

$$f(x) := \langle \varrho(x, a_0), ..., \varrho(x, a_{k-1}) \rangle.$$

9. Compute a list enumerating

$$B = \{\langle f(a), a \rangle : a \in A\}.$$

In order to keep notation simpler, we will denote the list enumerating B by B, as well.

Searching part.

Input: $x \in \mathbb{R}^n$.

1. let $m = 1$;
2. While $m \leq length(B)$ Do
3. let $\langle u, v\rangle$ be the m^{th} element of B;
4. let $d := ||f(x) - u||$;
5. if $d \leq \varepsilon/2$ then v is an output End(if);
6. if $d > \sqrt{k}\varepsilon$ then v is not an output End(if);
7. if $\varepsilon/2 < d \leq \sqrt{k}\varepsilon$ then
8. if $||x - v|| \leq \varepsilon$ then v is an output End(if);
9. if $||x - v|| > \varepsilon$ then v is not an output End(if);
10. End(if);
11. let $m := m + 1$;
12. End(Do).

We conclude this section by making some remarks on the Initializing and on the Searching parts.

Remarks on the Initializing part.

In step 1, ε' may be chosen to be an arbitrary real number greater than ε.

Suppose the algorithm has already computed $\{a_j : j < k\}$ for some k. Then, according to step 3, $\{B(r, a_j) : j < k\}$ is an r-net. Let $f_k : \mathbb{R}^n \to \mathbb{R}^k$ be the function that maps each $x \in \mathbb{R}^n$ onto

$$f_k(x) := \langle \varrho(x, a_0), ..., \varrho(x, a_{k-1})\rangle$$

(so, according to step 8, f is f_k for the last (largest) value of k). It follows from Lemma 1 of [9], that if $x, y \in A$ are such that $\varrho(x, y) > 2r$ then $||f_k(x) - f_k(y)|| > r$, or equivalently,

$$||f_k(x) - f_k(y)|| \leq r \text{ implies } \varrho(x, y) \leq 2r.$$

Thus, intuitively, the smaller is r, the "intermediate function" f_k is "more injective". According to step 5, ε' can be regarded as an estimation of "non-injectivity" of f_k and the algorithm is keep going whenever the value of ε' exceeds ε (the tolerable amount of inaccuracy given in the input). At the end we have

$$(*) \quad \text{if } ||f(x) - f(y)|| \leq \varepsilon/2 \text{ then } ||x - y|| \leq \varepsilon.$$

Further, by Lemma 1 of [9], for all k and $x, y \in \mathbb{R}^n$ we have

$$||f_k(x) - f_k(y)|| \leq \sqrt{k} \cdot ||x - y||,$$

particularly,

$$(**) \quad \text{if } ||f_k(x) - f_k(y)|| > \sqrt{k}\varepsilon \text{ then } ||x - y|| > \varepsilon.$$

So $(*)$ and $(**)$ can be summarized in the following three cases: let $x, y \in \mathbb{R}^n$ be arbitrary and let $d := ||f(x) - f(y)||$.

- if $d \leq \varepsilon/2$ then $||x - y|| \leq \varepsilon$;
- if $d > \sqrt{k}\varepsilon$ then $||x - y|| > \varepsilon$;
- if $\varepsilon/2 < d \leq \sqrt{k}\varepsilon$ then we have to compute $||x-y||$ in order to determine whether $||x-y|| \leq \varepsilon$. This is what we are doing in the Searching part.

As A is a finite set, the Initializing part always terminates. In fact, because of Step 2, it terminates after at most n many iterations of steps 3-6. If $k = n$ after Initializing, then this method is unable to reduce the dimension.

It is straightforward to see, that the number of steps of the Initializing part is proportional with $|A|^2$ in the worst case. The precise number of required steps strongly depends on ε and the structure of A, hence, at that level of generality we cannot improve the estimation for the time complexity of the Initializing part. However, we conjecture, that in particular situations, by a careful choice of ε, the number of required steps may be kept in a reasonably small level. We are planning to implement and test our algorithm on real life databases (such investigations are in progress at the moment).

Remarks on the Searching part.

It may happen, that there are several $\langle u, v_0\rangle, ..., \langle u, v_m\rangle \in B$ for which $||f(x) - u|| \leq \frac{\varepsilon}{2}$ (where x is the input). According to the choice of k in the Initializing part, we have $\varrho(v_i, v_j) \leq 2r \leq \varepsilon$ for all $i, j < m$, where $r = max_{x\in A}\varrho(\{a_j : j \leq k\}, x)$.

The number of steps in the Searching part is proportional with $|A|$, but when we compute $||f(x) - u||$, we are using the distance function of the k-dimensional Euclidean space. As k may be substantially smaller than n, this method may be more efficient than checking the elements of A step-by-step with the distance function of the n-dimensional Euclidean space. Further, because the Searching part is essentially a minimum-searching problem, it seems possible to accelerate Step 1 further by applying well known methods of algorithm theory or operation research.

4. Conclusion

In section 2 we investigated stability properties of the relational structures associated to totally bounded metric spaces. More concretely, in Theorem 2.3 we reconstructed a proof for the statement, that the relational structure associated to a totally bounded metric space is Δ-stable. Further, in Theorem 2.4 we proved a partial converse of the previous statement: a dense in itself metric space with a bounded metric is totally bounded iff its associated relational structure satisfies a strong stability condition.

Inspired by these theoretical results, in Section 3 we proposed a similarity detecting algorithm. The similarity between objects had been described by a metric ϱ. Our method is based on two parts: the Initializing and the Searching parts. In the special case, when the metric ϱ is the usual Euclidean distance of a large dimensional Euclidean space $\mathbb{R}^n$, our method can be considered as a way of reducing dimension. In this particular case we analyzed the time complexity of the Initializing and the Searching parts.

In the future we are planning to implement our algorithm and test it in real life databases.

REFERENCES

[1] Engelking, R., *General Topology*, Heldermann Verlag, Berlin, (1989).

[2] Faloutsos, C. and Lin, K., *FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets*, ACM., Vol. 24, No. 2, 163–174, (1995).

[3] R. J. FOWLER, M. S. PATERSON AND S. L. TANIMOTO, *Optimal packing and covering in the plane are NP-complete,* Information Processing Letters, Vol. 12, No. 3, 133-137, (1981).

[4] T. F. GONZALEZ, *Clustering to minimize the maximum intercluster distance,* Theoretical Computer Science, 38: 293-306, (1985).

[5] S. HAR-PELED, *Geometric Approximation Algorithms,* American Mathematical Society, Boston, US, (2011).

[6] Hjaltason, G.R. and Samet, H., *Contractive embedding methods for similarity searching in metric spaces*, Technical report, Computer Science Department, Center for Automation Research, Institute for Advanced Computer Studies, University of Maryland, (2000).

[7] V. Marianov and H. A. Eiselt editors, *Foundation of Location Analysis*, Springer Verlag, US, (2011).

[8] J. R. Munkres, *Topology*, Prentice Hall, US, (2000).

[9] G. SÁGI, *Almost injective Mappings of Totally Bounded Metric Spaces into Finite Dimensional Euclidean Spaces,* Advances in Pure Mathematics, 9. pp. 555-566 (2019).

[10] G. Sági and D. Nyiri, *On embeddings of finitw mwtric spaces*, in: the Proceedings of the 13th International Scientific Conference on Informatics (editors: V. Novitzka, S. Korečko and A. Szakál), pp. 227–231, IEEE, (2015).

[11] G. Sági and K. Al-Sabti, *Totally bounded metric spaces and similarity detecting algorithms*, in: the Proceedings of the 15th International Scientific Conference on Informatics (editors: W. Steingartner, S. Korečko and A. Szakál), pp. 338–342, IEEE, (2019).

[12] J. Satyabrata and P. Supantha,*Covering and packing of rectilinear subdivision,* WALCOM: algorithms and computation, 381–393, Lecture Notes in Comput. Sci., 11355, Springer, Cham, (2019).

[13] Shelah, S., *Classification Theory*, North–Holland, Amsterdam (1990).

[14] R. Zahed and T. M. Chan, *A clustering-based approach to kinetic closest pair,* Algorithmica Vol. 80, No. pp. 10, 2742–2756, (2018).