Applied Mathematical Sciences, Vol. 10, 2016, no. 33, 1619 - 1641
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2016.6263

Some Remarks on an Efficient Algorithm to Find a Centroid in a k-Dimensional Real Space

Alfio Giarlotta

Dept. of Economics and Business
University of Catania, Italy

Pietro Ursino

Dept. of Science and High Technology
University of Insubria, Italy

Abstract

Given a finite configuration of points in a metric space, a Steiner center (respectively, a centroid) is the point of the space (respectively, of the configuration) that minimizes the sum of the distances from all its elements. Working on the k-dimensional real space endowed with the Manhattan distance, we study the approximate algorithm that takes a point of minimum distance from the Steiner center of a configuration as its centroid. This algorithm is more efficient than all known approximate algorithms to obtain a centroid, and has applications in bio-informatics and economics, where data bases are quite large. Experimental results show that both the magnitude and the frequency of the error drastically decrease as the configuration becomes very large. We determine upper bounds to the magnitude of the error as a difference and as a ratio, and provide arguments to justify the negligible frequency of the error for large configurations.

Mathematics Subject Classification: 51M16, 65Y20, 90C90

Keywords: Minimum-distance point; centroid; Steiner center; Manhattan distance

1 Introduction

Given a metric space S and a finite collection A of points in S, a typical problem in discrete geometry is to find a point $\mathbf{s} \in S$ which minimizes the length of the A-star centered at $\mathbf{s}$, that is, the sum of the distances of all points in A from $\mathbf{s}$. If the point $\mathbf{s}$ that solves this minimization problem is assumed to belong to A, then $\mathbf{s}$ is called a *centroid* of A (or a 1-*median*) of A. On the other hand, if the solution $\mathbf{s}$ is allowed not to be in A, then $\mathbf{s}$ is called a *Steiner center* of A.

In most of the cases considered in the literature, the base space S is $\mathbb{R}^k$ endowed with the standard Euclidean distance. Alternative base spaces of interest are directed trees [5], and metric spaces with the Hamming distance [10]. The literature also examines other extensions of this problem, e.g., the case in which the given collection A is a subset of an Euclidean space having continuously many points [1]. See also some recent work oriented toward applications in bio-informatics, in which the base space S is an arbitrary metric or pseudometric space [3, 10].

If the base space S is the Euclidean space $\mathbb{R}^k$, then the problem of finding a centroid (or the Steiner center) of a finite collection A of points in S is known as the "Fermat-Weber location problem" [4]. More general versions of the Fermat-Weber problem are also considered in the literature. For example, given an integer number $d \geq 1$ and a finite collection A of points in $\mathbb{R}^k$, the "d-median problem" consists of finding d points (called medians) in a way such that the sum of the distances of each point of A from the closest median point is minimized.[1] Of course, the classical Fermat-Weber location problem is the 1-median problem in the Euclidean space $\mathbb{R}^k$.

Several (exact or approximate) algorithms for the 1-median problem in $\mathbb{R}^k$ endowed with a suitable distance are well-known. The goal of this paper is to analyze a special approximate algorithm, which solves the 1-median problem in $\mathbb{R}^k$ endowed with the Manhattan distance. The choice of this distance is motivated by the complexity of the related algorithm. In fact, from a computational point of view, d-median problems in $\mathbb{R}^k$ with the Euclidean distance are rather lengthy and difficult to solve, whereas if we endow $\mathbb{R}^k$ with the Manhattan distance, then the same problems can be solved more effectively by means of a linear algorithm.

The algorithm examined in this paper is called *minimum distance algorithm* (*md-algorithm*, for short): given a configuration A of points in $\mathbb{R}^k$ endowed with the Manhattan distance, this algorithm takes as a centroid of A a point in A having minimum distance from the Steiner center of A. (Henceforth, we call a point of minimum distance an *md-point*.)

The advantage of the md-algorithm is that it is linear. To prove this,

[1] For an even more general version of this problem, see [12].

observe that since the computation of a point minimizing the distance is obviously linear, it suffices to show the computation of Steiner center is linear as well. The latter fact motivates the choice of the Manhattan distance, because in this case the computation of the distance can be done coordinate by coordinate. It follows that the Steiner center can be determined by choosing the median point for each set of (odd) coordinates, which can be done in linear time (since the number of coordinates is fixed).

The main drawback of the md-algorithm is that it is approximate, since an md-point may well fail to be a centroid. However, experimental results show that this failure is rather unlikely for large configurations, in the sense that both the magnitude and the frequency of the error tend to decrease as the number of points increases. Here we provide some theoretical motivations to support this empirical evidence.

Specifically, our ideal goal is to obtain a good estimation of:

(i) the largest *magnitude* of the error, in terms of both the difference and the ratio between (a) the length of the A-star centered at an md-point, and (b) the length of the A-star centered at a centroid of A;

(ii) the *frequency* of the error, i.e., the relative number of times that an md-point fails to be a centroid.

For what concerns (i), in this paper we determine an upper bound to the magnitude of the error, in terms of both the difference and the ratio between (a) and (b). It turns out that these bounds are independent of the dimension k of the space, being only a function of the size of the given configuration.

Concerning (ii), observe that any result about the frequency of the error depends on the type of distribution of points. In the absence of constraints induced by the specific problem at hand, it is natural to assume a uniform distribution of points. Therefore, in this paper we examine special configurations of points in $\mathbb{R}^k$ – called *sampled configurations* – which are uniformly distributed over a bounded part of $\mathbb{R}^k$, and whose points have integer-valued coordinates. We prove that under suitable hypotheses, an md-point of a sampled configuration is always a centroid. Despite the seeming specialty of our case analysis, this result sheds some light on configurations with a large number of points. Indeed, as the number of points of a configuration A in $\mathbb{R}^k$ increases, the following two facts happen in average: (1) the points of A tend to be uniformly distributed over (a bounded part of) $\mathbb{R}^k$; and (2) md-points of A tend to have their distance from the Steiner center of A uniformly distributed over their k coordinates. Therefore, if we reason in terms of expected value, then the case of sampled configurations is less special than it might seem at a first sight, and provides a theoretical motivation of the experimental evidence.

We close this section with a brief overview of the existing algorithms which solve the 1-median problem, in order to provide the reader with some arguments supporting the choice of the md-algorithm over the others. Two categories of algorithms of this kind have been considered in the literature: exact and approximate. The *exact* algorithms that solve the 1-median problem in a generic metric space S are known to have complexity at least $O(n^2)$ (in fact, $O(n \log n)$ in case S is $\mathbb{R}^k$ endowed with the Manhattan distance), where n is the number of points of a configuration. On the other hand, there are *approximate* algorithms solving the 1-median problem with a complexity that is linear in the number of points. For example, in [9] the author describes such an algorithm, also providing an estimation of the error. Regrettably, the value of the linear constant is too high to make this algorithm useful in many concrete cases.

Another linear algorithm for the 1-median problem is given in [3]. This is a *mixed* algorithm, since it is approximate[2] in the general case, but it becomes exact if the base space S belongs to a certain category $\mathcal{M}$ of metric spaces. This algorithm has proven to be particularly effective in experimental results. Unfortunately, it is not practicable to employ it as an exact algorithm, since the computational cost to determine whether the base space S belongs to the category $\mathcal{M}$ is even higher than the computational cost of the algorithm for the search of a centroid.

In the particular case that the base space S is $\mathbb{R}^k$ endowed with the Manhattan distance, the md-algorithm seems to be a possible alternative to the above mentioned algorithms. In fact, the md-algorithm has the following features: (1) it is linear; (2) it provides some estimation of the error; (3) the probability of the error becomes negligible for uniformly distributed configurations with a large amount of points. Indeed, linearity of the md-algorithm follows from the fact that given a configuration A in $\mathbb{R}^k$, the search of the Steiner center of A can be performed by means of an algorithm of median searching repeated k times, and so the problem of finding an md-point of A has a linear cost as well.

The underlined features of the md-algorithm are particularly important in the light of its possible usage in applied sciences, such as bio-informatics and economics. Indeed, all data bases (of proteins, genes, stocks, etc.) used in these kinds of research do have a huge size, hence for practical purposes the search of a minimum distance point would yield a solution of the 1-median problem with a high likelihood. In fact, we conducted some experiments in $\mathbb{R}^k$ endowed with the Manhattan distance, and compared the results given by the md-algorithm versus those given by the algorithms proposed in [3, 9]. The evidence is overwhelmingly in favor of the md-algorithm, in terms of the speed

[2]Regrettably, an estimation of the size and the likelihood of the possible error is not given in [3].

of computation, the magnitude of error, and the probability of error.[3]

Furthermore, the advantages of the md-algorithm over the others are not limited to the case in which the base space is $\mathbb{R}^k$ endowed with the Manhattan distance. In fact, the md-algorithm is also effective in a generic metric space S, provided that there is a computationally convenient isometric embedding of S into $\mathbb{R}^k$. (Here by "computationally convenient" we mean that it has a complexity bounded by $O(n)$, where n is the number of points.) Such a perspective is justified by the existence of embeddings from suitable metric spaces into $\mathbb{R}^k$, which are either isometric or "nearly" isometric (i.e., with a suitably bounded distortion). More specifically, in [11] and [2] the authors exhibit, respectively, an isometric embedding from an ultrametric space U into $\mathbb{R}^k$, and an embedding with bounded distortion from a generic metric space S into $\mathbb{R}^k$ (endowed with the Euclidean distance, however extendable to $\mathbb{R}^k$ endowed with the Manhattan distance). An algorithmic version of the latter embedding is given in [13], where the author shows that the computational complexity of such an algorithm is logarithmic in the number n of points. Regrettably, the dimension k of the space $\mathbb{R}^k$ in which to embed S turns out to be equal to $\log n$, and so the md-algorithm would have complexity $O(n \cdot \log n)$ in this context. On the other hand, it would be interesting to determine whether the special types of base spaces used in applications (e.g., spaces with Hamming distance in bio-informatics) have properties such that the embedding into $\mathbb{R}^k$ is either isometric or "nearly" isometric. This interest carries out to other applications as well, such as those in economics and management. For some recent sources of the related literature, see, e.g., [14, 15] and references therein.

The paper is organized as follows. In the first two sections we provide an estimation of the magnitude of the error of the md-algorithm, in terms of difference (Section 2) and ratio (Section 3): both upper bounds only depend on the number of points of a configuration, and not on the dimension of the space in which they are embedded. In Section 4 we deal with sampled configurations in $\mathbb{R}^k$, giving some evidence that an md-point tends to be a centroid whenever dealing with uniformly distributed large configurations.

2 Magnitude of the error: a tight upper bound for the difference

For readers' convenience, first we establish the basic terminology and notation.

[3]The authors wish to thank A. Ferro and A. Pulvirenti for providing experimental evidence in this sense.

Definition 2.1 For each $k \geq 1$, the *Manhattan distance* in $\mathbb{R}^k$ is the metric $\Lambda^{(k)} \colon \mathbb{R}^k \to \mathbb{R}$ defined by

$$\Lambda^{(k)}(\mathbf{a}, \mathbf{b}) \ := \ \sum_{i=1}^{k} |a_i - b_i|$$

for each $\mathbf{a} = (a_1, \ldots, a_k), \mathbf{b} = (b_1, \ldots, b_k) \in \mathbb{R}^k$. Whenever there is no risk of confusion, we simplify notation, and write $\Lambda(\mathbf{a}, \mathbf{b})$ in place of $\Lambda^{(k)}(\mathbf{a}, \mathbf{b})$. Unless otherwise specified, by $\mathbb{R}^k$ we mean the metric space $\big(\mathbb{R}^k, \Lambda\big)$. A *configuration* in $\mathbb{R}^k$ is a finite multi-set in $\mathbb{R}^k$ (i.e., a finite subset of $\mathbb{R}^k$ such that some of the points can be repeated more than once). We denote by $\mathcal{P}_n^{(k)}$ the set of all configurations in $\mathbb{R}^k$ having $n \geq 3$ points.

Next, we define injective types of configurations.

Definition 2.2 A configuration $A \in \mathcal{P}_n^{(k)}$ is called *pure* if no repetition of points is allowed. Whenever A is pure, we use the notation $A \subseteq \mathbb{R}^k$ to emphasize the fact that A is a subset of $\mathbb{R}^k$. In particular, $A \subseteq \mathbb{R}^k$ is *totally pure* if for any $\mathbf{a} = (a_1, \ldots, a_k), \mathbf{b} = (b_1, \ldots, b_k) \in A$ with $\mathbf{a} \neq \mathbf{b}$, we have $a_j \neq b_j$ for each $j \in \{1, \ldots, k\}$.

Some results of this paper are only proved for pure (or totally pure) configurations of points. For our purposes, this fact does not affect the generality of results, since in applications one deals with random configurations in $\mathbb{R}^k$, so repetition of points or even coordinates is probabilistically negligible. However, whenever an extension to arbitrary configurations can be easily obtained, we shall add some comments on the way the argument can be adapted to the general setting.

We also emphasize from the outset that we only consider configurations with an *odd* number of points. Again, this assumption causes no loss of generality for the goals of this paper, due to the large size of the data bases in applications. In fact, whenever the number of points is even, we can make it odd by simply deleting one entry (hence paying a negligible price in terms of loss of data). The assumption that the size of a configuration is odd is needed in order to ensure the uniqueness of the Steiner center.

Definition 2.3 Let A be a configuration in $\mathbb{R}^k$. Define a map $L_A^{(k)} \colon A \to \mathbb{R}$ by

$$L_A^{(k)}(\mathbf{a}) \ := \ \sum_{\mathbf{x} \in A} \Lambda^{(k)}(\mathbf{a}, \mathbf{x})$$

for each $\mathbf{a} \in A$. The number $L_A^{(k)}(\mathbf{a})$ is called the *length of the A-star centered at* $\mathbf{a}$. Whenever there is no risk of confusion, we drop the superscript and write $L_A(\mathbf{a})$ in place of $L_A^{(k)}(\mathbf{a})$. Using the notion of A-star, we can associate to each $A \in \mathcal{P}_n^{(k)}$ some special points as follows.

- A *centroid* of A is a point $\mathbf{c} \in A$ such that the length of the A-star centered at $\mathbf{c}$ is minimum among all points in A. We denote by $\mathsf{Centr}(A)$ the set of all centroids in A. Since by definition we have $L_A(\mathbf{c}) = L_A(\mathbf{c}')$ for all $\mathbf{c}, \mathbf{c}' \in \mathsf{Centr}(A)$, we write C_A for the length of the A-star centered at any of the centroids of A.

- The *Steiner center* of A is the unique point $\mathbf{s} \in \mathbb{R}^k$ (not necessarily in A) which minimizes the total length of the A-star centered at it. The length of the A-star of the Steiner center of A is denoted by S_A.

- A *minimum distance point* of A (for short, *md-point*) is a point $\mathbf{m} \in A$ which has minimum distance from the Steiner center of A. We denote by $\mathsf{Min}(A)$ the set of all such points in A. Observe that if $\mathbf{m}, \mathbf{m}' \in \mathsf{Min}(A)$, then we usually have $L_A(\mathbf{m}) \neq L_A(\mathbf{m}')$. We denote by M_A the maximum length of an A-star among the points in $\mathsf{Min}(A)$, i.e., $M_A := \max\{L_A(\mathbf{m}) : \mathbf{m} \in \mathsf{Min}(A)\}$.

Remark 2.4 As an immediate consequence of Definition 2.1, the chain of inequalities

$$S_A \leq C_A \leq M_A$$

holds for each configuration $A \subseteq \mathbb{R}^k$.

The goal of this paper is to show that under suitable assumptions an md-point of $A \subseteq \mathbb{R}^k$ is a centroid of A, i.e., the inclusion $\mathsf{Min}(A) \subseteq \mathsf{Centr}(A)$ holds. Note that however, due to the fact that we consider random configurations in $\mathbb{R}^k$, both centroid and md-point are probabilistically unique, so it suffices to show that these two points coincide in most cases of interest.

2.1 Configurations in $\mathbb{R}$

To start, we analyze the simplified setting in which the base space is $\mathbb{R}$, and the configuration $A \subseteq \mathbb{R}$ is pure. In particular, the Manhattan distance $\Lambda^{(1)} = \Lambda$ becomes the Euclidean distance. Let $A \subseteq \mathbb{R}$ be such that $|A| = n = 2p+1 \geq 3$, where $p \in \mathbb{N} \setminus \{0\}$.

Definition 2.5 For all $a \in A$, denote $(\leftarrow, a) := \{x \in A : x < a\}$ and $(a, \rightarrow) := \{x \in A : x > a\}$. (Thus, the equality $A = (\leftarrow, a) \cup \{a\} \cup (a, \rightarrow)$ holds.) The *median* of A is the unique point $\gamma \in A$ such that $|(\leftarrow, \gamma)| = |(\gamma, \rightarrow)| = p$. Further, let $\Delta_a := |a - \gamma|$ be the distance of $a \in A$ from the median γ of A.

Note that since $|A| = n \geq 3$ by hypothesis, the length $L_A(a) = \sum_{x \in A} |a - x|$ of the A-star centered at each $a \in A$ is always positive. We will show that the function $L_A \colon A \to \mathbb{R}^+$ attains its minimum at γ, which is therefore the unique centroid (and Steiner center) of A. To this end, we will use the following technical fact.

Lemma 2.6 *Let $A \subseteq \mathbb{R}$ be a pure configuration such that $|A| = n = 2p+1 \geq 5$. Denote by γ the median of A. For each $a, b \in A$ such that $a \neq b$ and $0 < \Delta_a \leq \Delta_b$, we have:*

$$\Delta_b - (n-2)\Delta_a \ < \ L_A(b) - L_A(a) \ < \ (n-2)\Delta_b - \Delta_a. \tag{1}$$

If $a = \gamma$, then we obtain:

$$\Delta_b \ \leq \ L_A(b) - L_A(\gamma) \ < \ (n-2)\Delta_b. \tag{2}$$

Furthermore, these bounds are best possible.

Proof. Let a, b, γ be pairwise distinct. To prove (1), we distinguish the following two cases:

(i) $(a-\gamma)(b-\gamma) > 0$;

(ii) $(a-\gamma)(b-\gamma) < 0$.

In case (i), one can show that the following inequalities hold:

$$3(\Delta_b - \Delta_a) \ \leq \ L_A(b) - L_A(a) \ \leq \ (n-2)(\Delta_b - \Delta_a). \tag{3}$$

The proof of these inequalities is left to the reader.[4] By (3), it suffices to show that (1) holds in case (ii). Therefore, assume without loss of generality that

$$a < x_1 < \cdots < x_r < \gamma < y_1 < \cdots < y_s < b$$

for some $x_1, \ldots, x_r, y_1, \ldots, y_s \in A$, where $r, s \in \mathbb{N}$ are such that $0 \leq r, s \leq p-1$. For each $i = 1, \ldots, r$ and $j = 1, \ldots, s$, let $\varepsilon_i := x_i - a$ and $\delta_j := b - y_j$. Since

$$L_A(a) \ = \ \sum_{x \in (\leftarrow, a)} (a - x) + \Delta_a + p(\Delta_a + \Delta_b) + \sum_{i=1}^{r} \varepsilon_i - \sum_{j=1}^{s} \delta_j + \sum_{x \in (b, \rightarrow)} (x - b)$$

and

$$L_A(b) \ = \ \sum_{x \in (\leftarrow, a)} (a - x) + \Delta_b + p(\Delta_a + \Delta_b) - \sum_{i=1}^{r} \varepsilon_i + \sum_{j=1}^{s} \delta_j + \sum_{x \in (b, \rightarrow)} (x - b),$$

we obtain

$$L_A(b) - L_A(a) \ = \ (\Delta_b - \Delta_a) + 2\sum_{j=1}^{s} \delta_j - 2\sum_{i=1}^{r} \varepsilon_i.$$

[4] In fact, we will never directly use this part of the result, which is only mentioned for the sake of completeness.

This difference is maximized for $r = 0$, $s = p - 1$, and each δ_j approaching its maximum value Δ_b; its least upper bound (which becomes its maximum if $n = 3$) is $(n-2)\Delta_b - \Delta_a$. On the other hand, $L_A(b) - L_A(a)$ is minimized for $r = p - 1$, $s = 0$, and each ε_i approaching its maximum value Δ_a; its greatest lower bound (which becomes its minimum if $n = 3$) is $\Delta_b - (n-2)\Delta_a$. This proves the inequality (1), and also shows that the given bounds are best possible.

To finish the proof, we assume that $a = \gamma$, and prove that (2) holds. Note that the inequalities (2) can be viewed as the limit case of (1), as Δ_a goes to 0. It is easy to check that the inequalities

$$\Delta_b \ \leq \ L_A(b) - L_A(\gamma) \ < \ (n-2)\Delta_b$$

hold, where the second inequality is strict because the configuration is pure. □

The next two results are immediate consequences of Lemma 2.6.

Corollary 2.7 *For each $a \in A$, we have $L_A(\gamma) \leq L_A(a)$.*

Corollary 2.8 *For each pure configuration $A \subseteq \mathbb{R}$, the median of A is the unique centroid of A, and is also the Steiner center of A.*

Remark 2.9 The previous results hold also in case that A is a multi-set (i.e., the configuration is not pure). To prove this fact, it suffices to split repeated points having the same value a within an interval of the type $[a, a + \varepsilon)$ for a suitably chosen $\varepsilon > 0$, thus obtaining a pure configuration on which we can argue as above.

2.2 Configurations in $\mathbb{R}^k$

Here we extend the previous results to a multidimensional setting. Let $A \subseteq \mathbb{R}^k$ be a *totally pure* configuration with $n = 2p + 1 \geq 5$ points ($p \in \mathbb{N}$). For each $j \in \{1, \ldots, k\}$, denote by $A_j := \{a_j : \mathbf{a} \in A\} \subseteq \mathbb{R}$ the set of the j-th coordinates of the points of A. Since $A \subseteq \mathbb{R}^k$ is totally pure, we have $|A_j| = n$ for each $j \in \{1, \ldots, k\}$. In this setting, the Steiner center of A is $\mathbf{s} = (s_1, \ldots, s_k)$, where s_j is the median of A_j for each $j \in \{1, \ldots, k\}$. Of course, $\mathbf{s} \notin A$ in general. On the other hand, in case $\mathbf{s} \in A$, it follows that $|\mathsf{Centr}(A)| = 1$, and the (unique) centroid of A coincides with the Steiner center $\mathbf{s}$ of A. The next result is an immediate consequence of the definition of Manhattan distance.

Lemma 2.10 *For each $\mathbf{a} = (a_1, \ldots, a_k) \in A$, we have $L_A(\mathbf{a}) = \sum_{j=1}^{k} L_{A_j}(a_j)$.*

We can extend Lemma 2.6 to $\big(\mathbb{R}^k, \Lambda\big)$ as follows.

Lemma 2.11 *Let $A \subseteq \mathbb{R}^k$ be a totally pure configuration, and* $\mathbf{s}$ *its Steiner center. For each* $\mathbf{a} \in A \setminus \{\mathbf{s}\}$*, we have:*

$$\Lambda(\mathbf{a},\mathbf{s}) \;\leq\; L_A(\mathbf{a}) - L_A(\mathbf{s}) \;<\; (n-2)\,\Lambda(\mathbf{a},\mathbf{s}).$$

Proof. Let $\mathbf{a} = (a_1,\dots,a_k) \in A$. Without loss of generality, assume that the inequality $s_j \leq a_j$ holds for each $j = 1,\dots,k$. Thus, we have $\Lambda(\mathbf{a},\mathbf{s}) = \sum_{j=1}^{k} \Delta_j$, where $\Delta_j := a_j - s_j \geq 0$. For each j such that $\Delta_j > 0$ (note that there is at least one such j), Lemma 2.6 implies that

$$\Delta_j \;\leq\; L_{A_j}(a_j) - L_{A_j}(s_j) \;<\; (n-2)\Delta_j.$$

Thus the claim follows from Lemma 2.10. □

Finally, we determine a tight upper bound for the difference of the A-star centered at two distinct points of a totally pure configuration in $\big(\mathbb{R}^k, \Lambda\big)$.

Theorem 2.12 *Let $A \subseteq \mathbb{R}^k$ be a totally pure configuration with $n \geq 5$ points. For each* $\mathbf{a},\mathbf{b} \in A$*, with* $\mathbf{a} \neq \mathbf{b}$*, we have:*

$$L_A(\mathbf{b}) - L_A(\mathbf{a}) \;<\; (n-2)\,\Lambda(\mathbf{b},\mathbf{s}) - \Lambda(\mathbf{a},\mathbf{s}).$$

Further, this upper bound is best possible.

Proof. Let $\mathbf{a} = (a_1,\dots,a_n)$ and $\mathbf{b} = (b_1,\dots,b_n)$. For each $j \in \{1,\dots,k\}$, let $\Delta_j^{\mathbf{a}} := |a_j - s_j|$ and $\Delta_j^{\mathbf{b}} := |b_j - s_j|$. By Lemma 2.6, we obtain

$$L_A(\mathbf{b}) - L_A(\mathbf{a}) \;<\; \sum_{j=1}^{k} \big[(n-2)\Delta_j^{\mathbf{b}} - \Delta_j^{\mathbf{a}}\big] \;=\; (n-2)\,\Lambda(\mathbf{b},\mathbf{s}) - \Lambda(\mathbf{a},\mathbf{s})$$

which proves the claimed inequality. Finally, it is easy to exhibit a configuration of points such that this upper bound is attained as a limit: see Example 2.14 below. □

As a consequence of Theorem 2.12, we obtain a tight upper bound for the difference between: (i) the star centered at a centroid of A, and (ii) the star centered at an md-point of A. Without loss of generality, we assume that the Steiner center $\mathbf{s}$ of A does not belong to A (otherwise the problem is trivial); thus, for each $\mathbf{c} \in \mathsf{Centr}(A)$ and $\mathbf{m} \in \mathsf{Min}(A)$, we have $\Lambda(\mathbf{c},\mathbf{s})\,\Lambda(\mathbf{m},\mathbf{s}) > 0$.

Corollary 2.13 *Let $A \subseteq \mathbb{R}^k$ be a totally pure configuration with $n \geq 5$ points. For each* $\mathbf{c} \in \mathsf{Centr}(A)$ *and* $\mathbf{m} \in \mathsf{Min}(A)$*, we have*

$$L_A(\mathbf{m}) - L_A(\mathbf{c}) \;<\; (n-3)\,\Lambda(\mathbf{m},\mathbf{s}).$$

In particular, we have

$$M_A - C_A \;<\; (n-3)\,\Lambda(\mathbf{m},\mathbf{s}).$$

PROOF. Let $\mathbf{c} \in \mathsf{Centr}(A)$ and $\mathbf{m} \in \mathsf{Min}(A)$. By definition of $\mathsf{Min}(A)$, we have $\Lambda(\mathbf{c}, \mathbf{s}) \geq \Lambda(\mathbf{m}, \mathbf{s})$. Thus, the first inequality is an immediate consequence of Theorem 2.12. The second inequality follows from the definition of M_A and C_A. □

We end this section with an example that shows that the upper bound given in Corollary 2.13 can be attained as a supremum, hence the given inequality is best possible.

Example 2.14 We exhibit a totally pure configuration $A = \{\mathbf{a}^1, \ldots, \mathbf{a}^n\} \subseteq \mathbb{R}^k$ $(k \geq 2)$ with $n = 2p + 1 \geq 5$ points satisfying the following properties:

(i) the Steiner center $\mathbf{s}$ of A is the zero vector;

(ii) A has a unique centroid $\mathbf{c}$;

(iii) all points of A have the same Manhattan distance (equal to 1) from $\mathbf{s}$, hence $\mathsf{Min}(A) = A$;

(iv) there exists $\mathbf{m} \in A = \mathsf{Min}(A)$ such that $L_A(\mathbf{m}) - L_A(\mathbf{c})$ can be made arbitrarily close to the upper bound $n-3$, as determined by Corollary 2.13.

Select $p - 1$ positive real numbers $\varepsilon_2, \ldots, \varepsilon_p$ such that $\frac{1}{n} > \varepsilon_2 > \varepsilon_3 > \ldots > \varepsilon_p > 0$. Set:

- $\mathbf{a}^1 := \left(-\frac{1}{n}\,,\, -\frac{1}{n}\,,\, \cdots\,,\, -\frac{1}{n}\right)$
- $\mathbf{a}^i := \left(-\varepsilon_i\,,\, \frac{1}{n} + \frac{1/n-\varepsilon_i}{n-1}\,,\, \frac{1}{n} + \frac{1/n-\varepsilon_i}{n-1}\,,\, \ldots\,,\, \frac{1}{n} + \frac{1/n-\varepsilon_i}{n-1}\right)$ for each $i = 2, \ldots, p-1$
- $\mathbf{a}^p := \left(-\varepsilon_p\,,\, \frac{1}{n} + \frac{2/n-\varepsilon_p}{n-2}\,,\, \frac{1}{n} + \frac{2/n-\varepsilon_p}{n-2}\,,\, \cdots\,,\, \frac{1}{n} + \frac{2/n-\varepsilon_p}{n-2}\,,\, 0\right)$
- $\mathbf{a}^{p+1} := (0\,,\, 0\,,\, \cdots\,,\, 0\,,\, 1)$
- $\mathbf{a}^{p+2} := \left(\frac{1}{n}\,,\, \frac{1}{n}\,,\, \cdots\,,\, \frac{1}{n}\right)$
- $\mathbf{a}^j := (1 - (n-1)\varepsilon_{j-p-1}\,,\, -\varepsilon_{j-p-1}\,,\, -\varepsilon_{j-p-1}\,,\, \cdots\,, -\varepsilon_{j-p-1})$ for each $j = p+3, \ldots, 2p+1$.

Note that the Steiner center of A is $\mathbf{s} = \mathbf{0} = (0, \ldots, 0) \notin A$. Furthermore, for each $i = 1, \ldots, n$, we have $\Lambda(\mathbf{a}^i, \mathbf{0}) = 1$. The point $\mathbf{a}^1$ is the point whose total sum L_A is maximum among all points of A, i.e., $L_A(\mathbf{a}^1) = M_A$. On the other hand, the point $\mathbf{a}^{p+2}$ is the unique centroid of A. Furthermore, we have:

$$L_A(\mathbf{a}^1) - L_A(\mathbf{a}^{p+2}) \;=\; 2\left((p-1) - k\sum_{i=2}^{p}\varepsilon_i\right).$$

Thus, as each ε_i goes to 0, the difference $L_A(\mathbf{a}^1) - L_A(\mathbf{a}^{p+2})$ goes to $n-3$, as claimed.

In Figure 1 we give an instance of such a set $A \subseteq \mathbb{R}^5$ with $|A| = 9$. We denote by a circled star the points $\mathbf{a}^1 = \left(-\frac{1}{5}, -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{5}\right)$ and $\mathbf{a}^6 = \left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$, which are, respectively, the point of A with maximum L_A, and the unique centroid of A. The coordinates of the Steiner center $\mathbf{0} = (0, 0, 0, 0, 0)$ of A are denoted by a star.

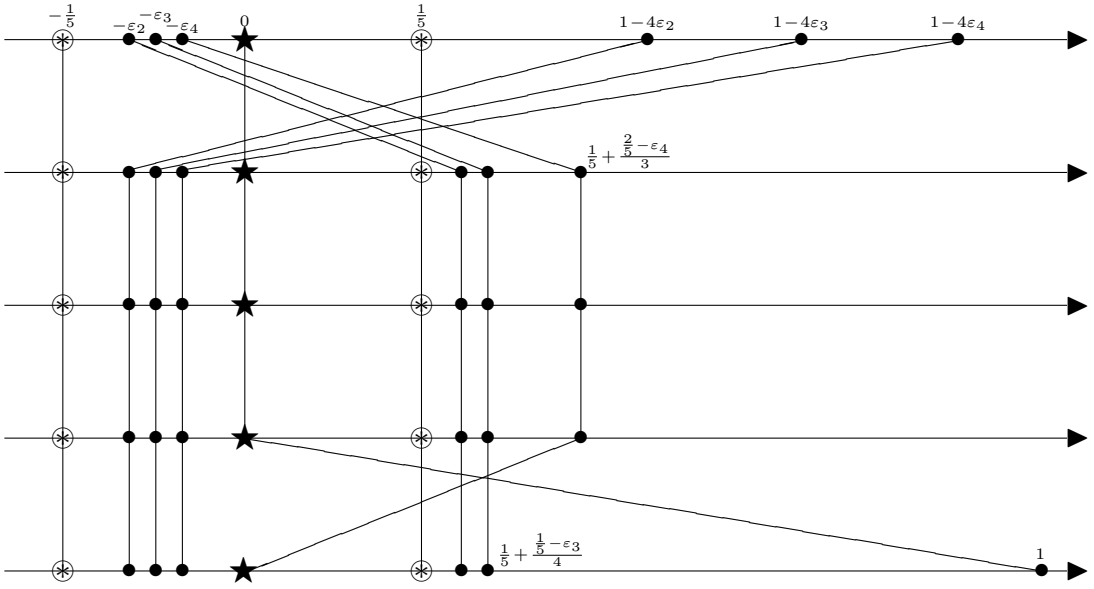

Figure 1: A totally pure configuration $A \subseteq \mathbb{R}^5$ such that $|A| = |\mathsf{Min}(A)| = 9$ and $|\mathsf{Centr}(A)| = 1$.

3 Magnitude of the error: an upper bound for the ratio

In this section we determine an upper bound for the ratio M_A/C_A, which however fails to be tight in some cases. To this aim, we exhibit a tight upper bound for the ratio M_A/S_A, which only depends on the number n of points of A, and not on the dimension k of the base space. Since $M_A/C_A \leq M_A/S_A$, this upper bound also holds for M_A/C_A.

In our approach, we use a simplified version of the technique employed by Fekete and Meijer in [7] to determine an upper bound for the worst-case ratio of C_A/S_A; therefore, we will only sketch the related proof.[5] Since our result does not depend on the dimension of the base space $\mathbb{R}^k$, henceforth we work with a fixed $k \geq 2$. Further, we assume without loss of generality (see Remark 3.1)

[5]Note that the upper bound obtained in [7] is proved only for configurations in $\mathbb{R}^2$ and $\mathbb{R}^3$, but it can be easily extended to configurations in $\mathbb{R}^k$ for arbitrary $k \geq 2$: see [8].

that the Steiner center of the configuration A is the origin $\mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^k$, and that $S_A = L_A(\mathbf{0}) = 1$.

Remark 3.1 The two assumptions that the Steiner center $\mathbf{s}$ of A is $\mathbf{0}$, and the length of its A-star is $S_A = 1$ cause no loss of generality, because we are indeed working with equivalence classes of configurations in $\mathbb{R}^k$. Specifically, given a configuration A in $\mathbb{R}^k$, the equivalence class of A comprises all configurations in $\mathbb{R}^k$ obtained from A by re-scaling all points by a constant factor $\alpha \in \mathbb{R}$ and/or translating all k coordinates of each point by the same constants β_i, where $i \in \{1, \ldots, k\}$. To keep notation simple, henceforth we avoid any explicit mention to equivalence classes of configurations, implicitly selecting in each class the representative such that its Steiner center is $\mathbf{0}$, and the length of the A-star centered at $\mathbf{0}$ is equal to 1.

Definition 3.2 For each odd $n \geq 3$, let δ_n be the worst-case value of the ratio M_A/S_A for configurations with n points, i.e.,

$$\delta_n \ := \ \sup\{M_A/S_A : A \in \mathcal{P}_n^{(k)}\}.$$

A configuration $A \in \mathcal{P}_n^{(k)}$ is called *extremely minimal* (*e-minimal*, for short) if it attains the value δ_n; in this case, a point $\mathbf{a} \in \mathsf{Min}(A)$ such that $L_A(\mathbf{a})/S_A = M_A/S_A = \delta_n$ is a *witness* of the e-minimality of A.

Lemma 3.3 *For each odd $n \geq 3$, there exists an e-minimal configuration $A \in \mathcal{P}_n^{(k)}$.*

Proof. Similar to the proof of Lemma 6 in [7]. □

In what follows, we prove several preliminary facts about e-minimal configurations. These results will be needed toward our goal of obtaining a tight upper bound for the ratio M_A/S_A (Theorem 3.10). To start, we show that the e-minimality of a configuration implies that all of its points are equally spaced from the Steiner center.

Definition 3.4 A configuration $A \subseteq \mathbb{R}^k$ is *perfectly balanced* if all points of A have the same distance from the Steiner center of A, i.e., $\mathsf{Min}(A) = A$.

An instance of a perfectly balanced configuration has already been given in Example 2.14.

Lemma 3.5 *Each e-minimal configuration is perfectly balanced.*

PROOF. Toward a contradiction, assume that $A \in \mathcal{P}_n^{(k)}$ is an e-minimal configuration such that $\Lambda(\mathbf{a}, \mathbf{0}) < \Lambda(\mathbf{b}, \mathbf{0})$ for some $\mathbf{a}, \mathbf{b} \in A$. For $\varepsilon > 0$ small enough, replace $\mathbf{b}$ by $\mathbf{b}' = (1-\varepsilon)\mathbf{b}$ in a way such that in the new configuration $A' := A \setminus \{\mathbf{b}\} \cup \{\mathbf{b}'\}$, the point $\mathbf{b}'$ still fails to be a point of minimum distance from the Steiner center of A'. If follows that $S_{A'} = S_A - \varepsilon'$ for some small $\varepsilon' > 0$, but $M_{A'} \geq M_A - \varepsilon'$. Therefore, $M_{A'}/S_{A'} \geq (M_A - \varepsilon')/(S_A - \varepsilon') > M_A/S_A = \delta_n$, which contradicts the maximality of δ_n. □

The converse of Lemma 3.5 does not hold (cf. Example 2.14). The next two results summarize some properties satisfied by a witness of the e-minimality of a configuration.

Lemma 3.6 *Let $A \in \mathcal{P}_n^{(k)}$ be an e-minimal configuration, and $\mathbf{a} = (a_1, \ldots, a_k) \in A$ a witness of its e-minimality. For each $i \in \{1, \ldots, k\}$ such that $a_i \neq 0$, there is no $\mathbf{b} = (b_1, \ldots, b_k) \in A \setminus \{\mathbf{a}\}$ with $a_i b_i > 0$ and $0 < |b_i| \leq |a_i|$.*

PROOF. Toward a contradiction, assume without loss of generality that $\mathbf{a} \in \mathsf{Min}(A)$ is such that $L_A(\mathbf{a}) = M_A = \delta_n$, but there exists $\mathbf{b} \in A$ such $0 < b_1 \leq a_1$.[6] By Lemma 3.5, there exists $i \in \{2, \ldots, k\}$ such that $|a_i| \leq |b_i|$. Without loss of generality, let $0 \leq a_2 \leq b_2$. For $\varepsilon > 0$ small enough, if we replace the point $\mathbf{b}$ by the point $\mathbf{b}' = (b_1 - \varepsilon, b_2 + \varepsilon, b_3, \ldots, b_k)$, then we obtain a new configuration $A' \in \mathcal{P}_n^{(k)}$ such that $S_{A'} = S_A = 1$ and $\mathsf{Min}(A') = A'$, but $L_{A'}(\mathbf{a}') > L_A(\mathbf{a})$. It follows that $M_{A'}/S_{A'} > M_A/S_A = \delta_n$, a contradiction. □

Lemma 3.7 *For each odd $n \geq 3$, there exists an e-minimal configuration in $\mathcal{P}_n^{(k)}$ having a witness of its e-minimality with exactly one nonzero coordinate.*

PROOF. Let A be an arbitrary e-minimal configuration with an odd number $n \geq 3$ of points, and let $\mathbf{a} \in A$ be a witness of its e-minimality. If $\mathbf{a} = (a_1, \ldots, a_k)$ has exactly one nonzero coordinate, then we are immediately done. Otherwise, we can assume without loss of generality that there exists $h \geq 2$ such that $a_j > 0$ for $1 \leq j \leq h$, and $a_j = 0$ for $h+1 \leq j \leq k$. Set $\mathbf{a}^1 := (a'_1, \ldots, a'_k)$, with $a_1^1 := a_1 + a_2$, $a_2^1 := 0$, and $a_j^1 := a_j$ for all $j \in \{3, \ldots, k\}$. Then the configuration $A^1 := A \setminus \{\mathbf{a}\} \cup \{\mathbf{a}^1\}$ is still an e-minimal configuration in $\mathcal{P}_n^{(k)}$, having $\mathbf{a}^1$ as a witness of its e-minimality. (Note the Steiner center of the new configuration A^1 is equal to that of the original configuration A.). If $\mathbf{a}^1$ has exactly one nonzero coordinate (namely, a_1^1), then we are done. Otherwise, we have $h \geq 3$, and we can proceed by induction to eventually obtain an e-minimal configuration $A^{h-1} := A^{h-2} \setminus \{\mathbf{a}^{h-2}\} \cup \{\mathbf{a}^{h-1}\}$

[6]Note that since A is a multi-set, it is possible that $\mathbf{b}$ is a repetition of $\mathbf{a}$, in particular b_1 can be equal to a_1.

such that $\mathbf{a}^{h-1}$ is a witness of its e-minimality, and $\mathbf{a}^{h-1}$ has only one nonzero coordinate (namely, the first). □

Using the preceding lemmas, we now compute the star centered at a witness of e-minimality.

Lemma 3.8 *If $\mathbf{a} \in A$ is a witness of the e-minimality of a configuration $A \in \mathcal{P}_n^{(k)}$, then the length of its A-star is $L_A(\mathbf{a}) = S_A + (n-2)\,\Lambda(\mathbf{a}, \mathbf{0})$.*

Proof. All e-minimal configurations in $\mathcal{P}_n^{(k)}$ share the same property of attaining the supremum δ_n. Thus, by Lemma 3.7 we can assume without loss of generality that the e-minimal configuration $A \subseteq \mathbb{R}^k$ has a witness $\mathbf{a} \in A$ of the type $\mathbf{a} = (a_1, 0, \ldots, 0)$, where $a_1 > 0$. In particular, we have $a_1 = \Lambda(\mathbf{a}, \mathbf{0}) > 0$, where $\mathbf{0}$ is the Steiner center of A. Lemma 3.6 yields that the first coordinate of all points $\mathbf{b} \in A \setminus \{\mathbf{a}\}$ is negative or zero, and their sum is equal to $-a_1$. Now an easy computation shows that the claim holds. □

It is useful to compare Lemma 3.8 to Lemma 2.11. In fact, given a configuration $A \in \mathcal{P}_n^{(k)}$ having $\mathbf{0}$ as its Steiner center, and an arbitrary point $\mathbf{a} \in A \setminus \{\mathbf{0}\}$, we have:

- if A it totally pure, then the inequality $L_A(\mathbf{a}) < S_A + (n-2)\,\Lambda(\mathbf{a}, \mathbf{0})$ holds by Lemma 2.11;
- if A is e-minimal, then the equality $L_A(\mathbf{a}) = S_A + (n-2)\,\Lambda(\mathbf{a}, \mathbf{0})$ holds by Lemma 3.8.

In particular, we obtain:

Corollary 3.9 *E-minimal configurations are not totally pure.*

We are finally ready to compute a tight upper bound for the ratio M_A/S_A, and, as a consequence, a bound for the ratio M_A/C_A.

Theorem 3.10 *Let $n \geq 3$ and $A \in \mathcal{P}_n^{(k)}$. We have*

$$\frac{M_A}{S_A} \;\leq\; 1 + \frac{n-2}{n} \tag{4}$$

and this inequality is best possible. In particular, $\frac{M_A}{C_A} \leq 1 + \frac{n-2}{n}$.

Proof. To prove that (4) holds, it suffices to show that $\delta_n = 1 + (n-2)/n$. First of all, observe that by Lemma 3.3 there exists an e-minimal configuration

$B \in \mathcal{P}_n^{(k)}$. Let $\mathbf{a} \in B$ be a witness of the e-minimality of $B \in \mathcal{P}_n^{(k)}$. Lemma 3.8 yields the chain of equalities

$$\delta_n = M_B = L_B(\mathbf{a}) = S_B + (n-2)\,\Lambda(\mathbf{a}, \mathbf{0}).$$

Since $S_B = L_B(\mathbf{0}) = 1$, Lemma 3.5 implies that $\Lambda(\mathbf{x}, \mathbf{0}) = 1/n$ for each $\mathbf{x} \in B$, in particular $\Lambda(\mathbf{a}, \mathbf{0}) = 1/n$. It follows that the equality $\delta_n = 1 + (n-2)/n$ holds. Finally, Example 3.11 given below shows that the upper bound for M_A/S_A is tight. □

Observe that Lemma 2.11 suffices to prove that the strict inequality $(M_A/S_A) < 1 + (n-2)/n$ holds for any totally pure configuration $A \subseteq \mathbb{R}^k$. On the other hand, the argument given in Theorem 3.10 is needed to show that the (weak) inequality holds in the general case.

We end this section by exhibiting an e-minimal configuration A in $\mathbb{R}^2$ such that $\mathbf{0}$ is the Steiner center of A, the equality $S_A = L_A(\mathbf{0}) = 1$ holds, and the upper bound $\delta_n = 1 + (n-2)/n$ is attained. Note that for $k \geq 3$, similar e-minimal configurations in $\mathbb{R}^k$ can be built by setting all the last $k-2$ coordinates of the points in A equal to zero.

Example 3.11 We construct a multi-set $A \subseteq \mathbb{R}^2$ having an odd number $n := 2p+1 \geq 3$ of points, which is an e-minimal configuration such that the the equality $\frac{M_A}{S_A} = 1 + \frac{n-2}{n}$ holds. The construction is trivial for $n = 3$, so consider the cases in which $p \geq 2$. Set $\mathbf{a}_1 := \left(\frac{1}{n}, 0\right)$, $\mathbf{a}_n := \left(-\frac{1}{2n}, \frac{1}{2n}\right)$, $\mathbf{a}_{p+1} := \left(-\frac{1}{2n}, -\frac{1}{2n}\right)$, $\mathbf{a}_i := \left(0, \frac{1}{n}\right)$ for each $i \in \{2, \ldots, p\}$, and $\mathbf{a}_j := \left(0, -\frac{1}{n}\right)$ for each $j \in \{p+2, \ldots, n-1\}$. One can readily check that the Steiner center of A is $\mathbf{0}$, $S_A = L_A(\mathbf{0}) = 1$, and $\Lambda(\mathbf{a}, \mathbf{0}) = \frac{1}{n}$ for each $\mathbf{a} \in A$. Furthermore, we have

$$\frac{M_A}{S_A} \;=\; M_A \;=\; L_A(\mathbf{a}_1) \;=\; 1 + \frac{n-2}{n}.$$

This proves that the upper bound in Theorem 3.10 is tight.

4 Frequency of the error: some probabilistic remarks

Experiments with large sets of data show that md-points turn out to be centroids in the majority of cases under examination. Our ideal goal would be to provide a full theoretical justification of this empirical evidence. In this section we make a first step in this direction, giving a partial justification based on an expected value argument. Roughly speaking, we show that under suitable hypothesis and whenever dealing with large configurations, points of minimum distance tend to be – in average – a good approximation of centroids.

More precisely, we analyze configurations of points with equally spaced integer coordinates, called *sampled*, and prove some related results about the coincidence of md-points and centroids. In view of the fact that the expected value of a random configuration is almost sampled, and this expected value tends to be sampled as the number of points diverges, the special results proved for sampled configurations become significant for the general case, and shed some light on the observed experimental evidence.

Definition 4.1 Let $A \in \mathcal{P}_n^{(k)}$ be a configuration of $n = 2p + 1 \geq 3$ points in $\mathbb{R}^k$ having integer coordinates and such that for each $\mathbf{a} = (a_1, \dots a_k), \mathbf{b} = (b_1, \dots, b_k) \in A$, we have:

- $a_i \in [-p, p]$ for each $i \in \{1, \dots, k\}$;
- if $\mathbf{a} \neq \mathbf{b}$, then $a_i \neq b_i$ for each $i \in \{1, \dots, k\}$.

We call A an *integer sampled configuration* (or, simply, a *sampled configuration*).

Note that a sampled configuration A is a totally pure configuration in $\mathbb{R}^k$, and it is a subset of $\mathbb{Z}^k$. Further, A has the property that for all integers $i, j \in \mathbb{Z}$ such that $1 \leq i \leq p$ and $-p \leq j \leq p$, there exists a unique $\mathbf{a} = (a_1, \dots, a_k) \in A$ such that $a_i = j$. Finally, observe that the Steiner center $\mathbf{s}$ of A is the origin $\mathbf{0} = (0, \dots, 0) \in \mathbb{R}^k$. (However, in some cases we keep using the notation $\mathbf{s} = (s_1, \dots, s_k)$ for the Steiner center of a given configuration of points, since our results do not depend on the fact that the Steiner center is the origin.)

Dealing with sampled configurations certainly eases computations, but does not eliminate the problem that the sets of centroids and md-points may be disjoint, as the next example shows.

Example 4.2 Consider the following sampled configuration in $\mathbb{R}^3$:

$$P \equiv (-1, -1, -1),\ Q \equiv (1, 1, 1),\ R \equiv (-2, 0, 0),\ S \equiv (0, -2, -2),\ T \equiv (2, 2, 2).$$

It is immediate to check that the unique point of minimum distance from the Steiner center $(0, 0, 0)$ is R, whereas the two centroids are P and Q.

Next, we compute the length of the star centered at any point of a sampled configuration.

Lemma 4.3 *Let $A \subseteq \mathbb{Z}^k$ be a sampled configuration with $n = 2p + 1$ points. For each $\mathbf{a} = (a_1, \dots, a_k) \in A$, we have*

$$L_A(\mathbf{a}) \ = \ 2k\binom{p+1}{2} + \sum_{j=1}^{k} |a_j|^2 .$$

PROOF. For each $j \in \{1, \dots, k\}$, let $d_j := |a_j|$. An easy computation shows that the equalities

$$L_{A_j}(\mathbf{a}) \;=\; \sum_{i=1}^{p-d_j} i + \sum_{i=1}^{p+d_j} i \;=\; \binom{p+d_j+1}{2} + \binom{p-d_j+1}{2} \;=\; 2\binom{p+1}{2} + d_j^2$$

hold. Since $L_A(\mathbf{x}) = \sum_{j=1}^{k} L_{A_j}(\mathbf{x})$ for each $\mathbf{x} \in \mathbb{R}^k$, the claim follows. □

Lemma 4.3 computes the length of the star centered at an arbitrary point of A as the sum of two quantities:

(i) a fixed part $2k\binom{p+1}{2}$, which is the same for all points of A, and is a function only of the number n of points (since the dimension k of the ground space $\mathbb{R}^k$ is fixed);

(ii) a variable part $\sum_{j=1}^{k} |a_j|^2$, which grows with the square of the coordinates of a point, but is however a constant from the point of view of complexity (since k is fixed).

Lemma 4.3 will be used below in the proof of the main result of this section (Theorem 4.7). The following consequence, which is not relevant for our problem, is however worth being mentioned.

Remark 4.4 An alternative reading of Lemma 4.3 is that whenever dealing with sampled configurations, a centroid of a configuration under the Manhattan distance *is* an md-point under the Euclidean metric. This is an interesting instance of a problem over a certain metric space which can be solved by looking at a similar problem in a different metric space.

As a significant application of Lemma 4.3, consider the following example.

Example 4.5 Let $A \subseteq \mathbb{Z}^k$ be a sampled configuration with $n = 2p + 1 \geq 5$ points. Assume that A contains the following two points, having exactly two (resp. one) nonzero coordinates: $(r, s, 0, \dots, 0)$ and $(h, 0, 0, \dots, 0)$. Lemma 4.3 yields

$$L_A(r, s, 0, \dots, 0) \;=\; 2p(p+1)+|r|^2+|s|^2 \qquad \text{and} \qquad L_A(h, 0, 0, \dots, 0) \;=\; 2p(p+1)+|h|^2\,.$$

Note that since all sampled configurations in $\mathbb{R}^k$ have by definition $\mathbf{0} = (0, 0, \dots, 0)$ as their Steiner center, the above equalities hold for points of A whose coordinates in absolute value are permutations of the coordinates of $(|r|, |s|, 0, \dots, 0)$ and $(|h|, 0, 0, \dots, 0)$, respectively. In particular, if $(h, 0, 0, \dots, 0)$ is an md-point but not a centroid, and $(r, s, 0, \dots, 0)$ is a centroid but not an md-point, then $h, r, s \in \mathbb{Z}$ are such that $|h| < |r| + |s|$ and $|h|^2 > |r|^2 + |s|^2$.

To show that in a sampled configuration md-points with evenly distributed coordinates are indeed centroids, we need another technical lemma.

Lemma 4.6 *Let $f\colon \mathbb{R}^k \to \mathbb{R}$ be the function defined by $f(x_1,\ldots,x_k) := \sum_{i=1}^k x_i^2$ for each $(x_1,\ldots,x_k) \in \mathbb{R}^k$, and subject to the following constraints:* (i) $\sum_{i=1}^k x_i = h$, *where $h > 0$ is a parameter;* (ii) $x_i \geq 0$ *for each $i \in \{1,\ldots,k\}$. Then, we have:*

(a) *the minimum value of f subject to* (i) *and* (ii) *is h^2/k, which is attained at the unique point $\left(\frac{h}{k},\ldots,\frac{h}{k}\right)$;*

(b) *the maximum value of f subject to* (i) *and* (ii) *is h^2, which is attained at the k points $(h,0,\ldots,0,0)$, $(0,h,\ldots,0,0)$, …, $(0,0,\ldots,0,h)$.*

Proof. This is an easy exercise in the case $k = 2$. For the general case, proceed by induction on $k \geq 2$. □

As announced, we have:

Theorem 4.7 *Let A be a sampled configuration in $\mathbb{R}^k$ such that $|A| = n = 2p+1 \geq 5$. Assume that $\mathbf{s} = \mathbf{0} = (0,\ldots,0)$ is the Steiner center of A, and $\mathbf{a} = (a_1,\ldots,a_k)$ is an md-point such that $\Lambda(\mathbf{a},\mathbf{s}) = h > 0$. If $|a_i| = h/k$ for each $i \in \{1,\ldots,k\}$, then $\mathbf{a}$ is a centroid of A.*

Proof. Let $\mathbf{a} = (a_1,\ldots,a_k) \in \mathsf{Min}(A)$ be such that $\Lambda(\mathbf{a},\mathbf{s}) = h > 0$, and $|a_i| = h/k$ for each $i \in \{1,\ldots,k\}$. Further, let $\mathbf{c} = (c_1,\ldots,c_k) \in \mathsf{Centr}(A)$ be such that $\Lambda(\mathbf{c},\mathbf{s}) = q \geq h$. To prove the claim, it suffices to show that $L_A(\mathbf{a}) \leq L_A(\mathbf{c})$. The hypothesis and Lemma 4.3 imply

$$L_A(\mathbf{a}) - L_A(\mathbf{s}) \;=\; \sum_{i=1}^{k} |a_i|^2 \;=\; h^2/k\,,$$

whereas a joint application of Lemmas 4.3 and 4.6(a) yield

$$L_A(\mathbf{c}) \;-\; L_A(\mathbf{s}) \;=\; \sum_{i=1}^{k} |c_i|^2 \geq q^2/k.$$

It follows that

$$L_A(\mathbf{a}) - L_A(\mathbf{s}) \;=\; h^2/k \;\leq\; q^2/k \;\leq\; L_A(\mathbf{c}) - L_A(\mathbf{s})$$

hence $L_A(\mathbf{a}) \leq L_A(\mathbf{c})$. This proves that $\mathbf{a}$ is a centroid. □

Theorem 4.7 describes a very particular case in which an md-point is a centroid. One can observe that the set up of this result is way too special to be significative. However, we shall argue below that whenever the random configuration has an *average behavior*, then Theorem 4.7 becomes effective in showing that the md-algorithm works well. Here by "average behavior" we mean "taking the expected value" of a generic configuration. In order to effectively work in terms of expected value, next we introduce the notion of a quasi-sampled configuration.

Definition 4.8 Denote by $\mathsf{Samp}_n^{(k)}$ the set of all sampled configurations of $n = 2p + 1$ points in $\mathbb{R}^k$. A configuration $A = \{\mathbf{a}_1, \dots \mathbf{a}_n\}$ of n distinct points in $\mathbb{R}^k$ is called *quasi-sampled* if

$$\min_{X \in \mathsf{Samp}_n^{(k)}} \left(\max_{\mathbf{x}_i \in X} \left\{ |a_i^j - x_i^j| : j = 1, \dots, k \right\} \right) \; < \; \frac{1}{2}. \tag{5}$$

If A is quasi-sampled, then a sampled configuration A' that witnesses the inequality (5) is called a *sampled shadow* of A.

For instance, the (totally pure) configuration $A \subseteq \mathbb{R}^2$ given by

$$A := \{(-2.3, 0.2), (-0.75, 1.9), (0.4, -1.85), (1, 1), (2.02, -1.45)\}$$

is quasi-sampled, and its (unique) sampled shadow is the configuration

$$A' = \{(-2, 0), (-1, 2), (0, -2), (1, 1), (2, -1)\}.$$

We have defined quasi-sampled configurations in terms of their similarity to suitable integer-valued configurations, namely, their sampled shadows.[7] However, upon rescaling the base metric space, one can equivalently define a quasi-sampled configuration in terms of its closeness to a suitable "equally spaced" configuration, which is isometrically equivalent to a sampled configuration. This fact is especially important for our purposes, since we need to work on a fixed bounded part of the metric space, independently of the size of the random configuration.

Specifically, assume that experimental evidence shows that (the relevant part of) the phenomenon under examination takes values in a bounded portion of $\mathbb{R}^k$, say $[0, 1]^k$ without loss of generality. Then we can associate to this phenomenon "rescaled" forms of sampled and quasi-sampled configurations with an odd number n of points as follows. Partition the unit interval $[0, 1] \subseteq \mathbb{R}$ into n subintervals of equal size $1/n$. Then a totally pure configuration $A \subseteq [0, 1]^k$ of size n is sampled whenever all the n-element sequences of the k coordinates are permutations of the centers of the subintervals, and it is quasi-sampled if it is sufficiently close to a sampled one. The next definition provides the formal setting.

[7] In the continuous case, the sampled shadow of a quasi-sampled configuration is unique with probability one.

Definition 4.9 Let $A \subseteq [0,1]^k$ be a configuration with an odd number $n \geq 3$ of points. We say that A is *fractional sampled* if it is totally pure and for all $\mathbf{a} = (a_1, \ldots a_k), \mathbf{b} = (b_1, \ldots, b_k) \in A$, we have:

- $a_i \in \left\{\frac{1}{2n}, \frac{3}{2n}, \ldots, \frac{2n-1}{2n}\right\}$ for each $i \in \{1, \ldots, k\}$;
- if $\mathbf{a} \neq \mathbf{b}$, then $a_i \neq b_i$ for each $i \in \{1, \ldots, k\}$.

Furthermore, A is *fractional quasi-sampled* if it is totally pure and

$$\min_{X \in \mathsf{FSamp}_n^{(k)}} \left(\max_{\mathbf{x}_i \in X} \left\{|a_i^j - x_i^j| : j = 1, \ldots, k\right\}\right) \; < \; \frac{1}{2n} \tag{6}$$

where $\mathsf{FSamp}_n^{(k)}$ denotes the set of all fractional sampled configurations of $[0,1]^k$.

Fractional sampled and quasi-sampled configurations reproduce on a *fixed* bounded space – the k-dimensional unit cube $[0,1]^k$ – the same phenomenon as sampled and quasi-sampled configurations in $\mathbb{R}^k$. Here the key fact is that for the fractional case, the bounded size of the ground space is fixed, and does not vary along with the number of points in a (quasi-)sampled configuration (where the ground set is $[-p, p]^k$ for configurations with $n = 2p + 1$ points).

Next, we describe a general example to obtain a random configuration of points in $[0,1]^k$. This example is constructed by aggregating k – one per coordinate – *bucket-like* experiments: see, e.g., [6] pp. 170–172. In fact, using the structure of the Manhattan distance, we generate random configurations in $[0,1]$, whose aggregation yields a random configuration in $[0,1]^k$.

Example 4.10 Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ be a configuration of $n = 2p + 1 \geq 3$ points in $[0,1]^k$, where the generic element of A is $\mathbf{a}_i = \left(a_i^1, \ldots, a_i^k\right)$. We interpret each fixed set of coordinates $\{a_1^j, \ldots, a_n^j\}$ of the points in A, where $j \in \{1, \ldots, k\}$, as a bucket-like experiment. For the sake of illustration, let us consider the set of the first coordinates $\{a_1^1, \ldots, a_n^1\}$ of the points in A. Partition the unit interval $[0,1]$ in n subintervals of the same size $1/n$, the so-called *buckets* of the experiment. Let X_i^1 be the random variable that denotes the number of first coordinates falling into the i-th bucket. Since we can assume a uniform distribution of the first coordinates in the interval $[0,1]$, it follows that each X_i^1 has expected value $E[X_i^1] = 1$, that is, each of the n buckets will contain in average exactly one of the n coordinates. Now repeat the same experiment for each of the k coordinates, and obtain the same conclusion.

To make use of Example 4.10 in our setting, assume now that empirical evidence shows that the phenomenon under examination takes values in $[0,1]^k$. Now the bucket-like experiment can be interpreted as the fact that the n-point

configuration in $[0,1]$ given by the first coordinates of the points in A is indeed (fractional) quasi-sampled. Since we are using the Manhattan distance, we can independently repeat the same experiment k times, one per coordinate, thus producing a random configuration A in $[0,1]^k$, which in average will be (fractional) quasi-sampled.

By arguing in the same fashion as in Example 4.10, we can conclude that the distance of a generic point of a bounded configuration in $\mathbb{R}^k$ from its Steiner center is evenly distributed among its k coordinates whenever we reason in average, that is, in terms of expected value. In particular, this holds for its md-points. However, as the number of points in $[0,1]^k$ gets larger and larger, a (fractional) quasi-sampled configuration obviously tends to become (fractional) sampled. In this sense, Theorem 4.7 intuitively applies for large configurations of points, thus justifying the "average effectiveness" of the md-algorithm.

Acknowledgements. The authors thank S. Giuffrida and R. Re for some useful suggestions.

References

[1] K. Beurer, S. P. Fekete, J. S. B. Mitchell, On the continuous Fermat-Weber problem, *Oper. Res.*, **53** (2005), no. 1, 61-76. http://dx.doi.org/10.1287/opre.1040.0137

[2] J. Bourgain, On Lipschitz embedding of finite metric spaces in Hilbert spaces, *Israel J. Math.*, **52** (1985), no. 1-2, 46-52. http://dx.doi.org/10.1007/bf02776078

[3] D. Cantone, G. Cincotti, A. Ferro, A. Pulvirenti, An efficient approximate algorithm for the 1-median problem in metric spaces, *SIAM J. Optim.*, **16** (2005), no. 2, 434-451. http://dx.doi.org/10.1137/s1052623403424740

[4] R. Chandrasekaran, A. Tamir, Algebraic optimization: The Fermat-Weber location problem, *Math. Program.*, **46** (1990), 219-224. http://dx.doi.org/10.1007/bf01585739

[5] M. Chrobak, L. Larmore, W. Rytter, The k-median problem for directed trees, Chapter in *Mathematical Foundations of Computer Science*, LNCS 2136, Springer Berlin Heidelberg, 2001, 260-271. http://dx.doi.org/10.1007/3-540-44683-4_23

[6] T. H. Cormen, C. E. Leiserson, R. Rivest, C. Stein, *Introduction to Algorithms*, MIT Press, 1990.

[7] S. P. Fekete, H. Meijer, On minimum stars and maximum matchings, *Discr. and Comput. Geom.*, **23** (2000), 389-407.
http://dx.doi.org/10.1007/pl00009508

[8] A. Giarlotta, P. Ursino, An extension to $\mathbb{R}^k$ of a result by Fekete and Meijer, *Far East J. Math. Sci.*, **68** (2012), no. 1, 21-29.

[9] P. Indyk, *High-Dimensional Computational Geometry*, Ph.D. Dissertation, Stanford, 2001.

[10] N. C. Jones, P. Pevzner, *An Introduction to Bioinformatics Algorithms*, MIT Press, 2004.

[11] S. A. Shkarin, Isometric embedding of finite ultrametric spaces in Banach spaces, *Topology and its Applications*, **142** (2004), 13-17.
http://dx.doi.org/10.1016/j.topol.2003.12.002

[12] J.-H. Lin, J. S. Vitter, Approximation algorithms for geometric median problems, *Inform. Process. Lett.*, **44** (1992), no. 5, 245-249.
http://dx.doi.org/10.1016/0020-0190(92)90208-d

[13] N. Linial, Finite metric spaces-combinatorics, geometry and algorithms, *Proc. of the International Congress of Mathematicians*, Higher Ed. Press, Beijing, **3** (2002), 573-586.

[14] D. Saban, N. Stier-Moses, The Competitive Facility Location Problem in a Duopoly: Connections to the 1-Median Problem, *Lect. Notes in Comp. Sci.*, **7695** (2012), 539-545.
http://dx.doi.org/10.1007/978-3-642-35311-6_44

[15] B-F. Wang, J-H. Ye, P-J. Chen, On the Round-Trip 1-Center and 1-Median Problems, *Lect. Notes in Comp. Sci.*, **7157** (2012), 100-111.
http://dx.doi.org/10.1007/978-3-642-28076-4_12

Received: March 3, 2016; Published: May 4, 2016