On the largest empty axis-parallel box amidst n points

Adrian Dumitrescu* Minghui Jiang†

January 5, 2012

Abstract

We give the first efficient $(1-\varepsilon)$-approximation algorithm for the following problem: Given an axis-parallel d-dimensional box R in $\mathbb{R}^d$ containing n points, compute a *maximum-volume empty axis-parallel d-dimensional box* contained in R. The minimum of this quantity over all such point sets is of the order $\Theta\left(\frac{1}{n}\right)$. Our algorithm finds an empty axis-aligned box whose volume is at least $(1-\varepsilon)$ of the maximum in $O((8ed\varepsilon^{-2})^d \cdot n \log^d n)$ time. No previous efficient exact or approximation algorithms were known for this problem for $d \geq 4$. As the problem has been recently shown to be NP-hard in arbitrarily high dimensions (*i. e.*, when d is part of the input), the existence of an efficient exact algorithm is unlikely.

We also present a $(1-\varepsilon)$-approximation algorithm that, given an axis-parallel d-dimensional cube R in $\mathbb{R}^d$ containing n points, computes a *maximum-volume empty axis-parallel hypercube* contained in R. The minimum of this quantity over all such point sets is also shown to be of the order $\Theta\left(\frac{1}{n}\right)$. A faster $(1-\varepsilon)$-approximation algorithm, with a milder dependence on d in the running time, is obtained in this case.

Keywords: Largest empty box, largest empty hypercube, discrepancy of a set of points, van der Corput point set, Halton-Hammersley point set, approximation algorithm, data mining.

1 Introduction

Given a set S of n points in the unit square $U = [0,1]^2$, let $A(S)$ be the maximum area of an empty axis-parallel rectangle contained in U, and let $A(n)$ be the minimum value of $A(S)$ over all sets S of n points in U. For any dimension $d \geq 2$, given a set S of n points in the unit hypercube $U_d = [0,1]^d$, let $A_d(S)$ be the maximum volume of an empty axis-parallel hyperrectangle (d-dimensional axis-parallel box) contained in U_d, and let $A_d(n)$ be the minimum value of $A_d(S)$ over all sets S of n points in U_d. For simplicity we sometimes omit the subscript d in the planar case ($d = 2$). For a fixed d, it is known [22] that $A_d(n)$ is of the order $\Theta(\frac{1}{n})$.

We first introduce some notations and definitions. Throughout this paper, a *box* is an *open* axis-parallel hyperrectangle contained in the unit hypercube $U_d = [0,1]^d$, $d \geq 2$. Given a set S of points in U_d, a box B is *empty* if it contains no points in S, *i. e.*, $B \cap S = \emptyset$. Some small examples are illustrated in Fig. 1 (Section 2) and Fig. 2 (Section 6).

Given an axis-parallel rectangle R in the plane containing n points, the problem of computing a maximum-area empty axis-parallel sub-rectangle contained in R is one of the oldest problems

*Department of Computer Science, University of Wisconsin–Milwaukee, WI 53201-0784, USA. Email: dumitres@uwm.edu. Supported in part by NSF CAREER grant CCF-0444188 and NSF grant DMS-1001667. Part of the research by this author was done at Ecole Polytechnique Fédérale de Lausanne.

†Department of Computer Science, Utah State University, Logan, UT 84322-4205, USA. Email: mjiang@cc.usu.edu. Supported in part by NSF grant DBI-0743670.

studied in computational geometry. For instance, this problem arises when a rectangular shaped facility is to be located within a similar region which has a number of forbidden areas, or in cutting out a rectangular piece from a large similarly shaped metal sheet with some defective spots to be avoided [20]. In higher dimensions, finding the largest empty axis-parallel box has applications in data mining, in finding large gaps in a multi-dimensional data set [15]. Last but not least, an efficient $(1-\varepsilon)$-approximation algorithm is useful for testing various constructions, etc.

Several algorithms have been proposed for the planar problem over time [1, 2, 3, 8, 12, 19, 20, 21]. For instance, an early algorithm by Chazelle, Drysdale and Lee [8] runs in $O(n \log^3 n)$ time and $O(n \log n)$ space. The fastest known algorithm, proposed by Aggarwal and Suri in 1987 [1], runs in $O(n \log^2 n)$ time and $O(n)$ space. A lower bound of $\Omega(n \log n)$ in the algebraic decision tree model for this problem has been shown by McKenna *et al.* [19].

For any dimension d, there is an obvious brute-force algorithm running in $O(n^{2d+1})$ time and $O(n)$ space. No significantly faster algorithms, *i. e.*, with a fixed degree polynomial running time in $\mathbb{R}^d$, were known. Confirming this state of affairs, Backer and Keil [5, 6] recently proved that the problem is NP-hard in high dimensions (*i. e.*, when d is part of the input). They also reported an exact algorithm running in $O(n^d \log^{d-2} n)$ time, for any $d \geq 3$. In particular, the running time of their algorithm for $d = 3$ is $O(n^3 \log n)$. Previously, Datta and Soundaralakshmi [13] had reported an $O(n^3)$ time exact algorithm for the $d = 3$ case, but their analysis for the running time seems incomplete. Specifically, the $O(n^3)$ running time depends on an $O(n^3)$ upper bound on the number of maximal empty boxes (see discussions in the next paragraph), but they only gave an $\Omega(n^3)$ lower bound.

Here we present the first efficient $(1-\varepsilon)$-approximation algorithm for finding an axis-parallel empty box of maximum volume, in any dimension d, whose running time is nearly linear for small d and increases only by an $O(d \cdot \log n/\varepsilon^2)$ factor when one goes up one dimension. The algorithm finds an axis-parallel box whose volume is at least $(1-\varepsilon)$ times the largest volume of an empty axis-parallel d-dimensional box contained in R.

An empty box of maximum volume must be maximal with respect to inclusion. A *maximal empty* box is sometimes also called *restricted* in the literature [13, 20]. Thus the maximum-volume empty box in U_d is restricted. Naamad *et al.* [20] have shown that in the plane, the number of maximal empty rectangles is $O(n^2)$, and that this bound is tight. It was conjectured by Datta and Soundaralakshmi [13] that the maximum number of maximal empty boxes is $O(n^d)$ for each (fixed) d. The conjecture has been recently confirmed by Backer and Keil [5] (for $d \geq 3$). Here we extend (Theorem 7, Section 6) the lower bound constructions with $\Omega(n^d)$ maximal empty boxes for $d = 2$ in [20] and $d = 3$ in [13] to arbitrary d. Independently and simultaneously with us [14], Backer and Keil have also obtained this result [5]. Hence the maximum number of maximal empty boxes is $\Theta(n^d)$ for each fixed d. This means that any algorithm for computing a maximum-volume empty box based on enumerating maximal empty boxes is bound to be inefficient in the worst case. However, as it is the case with the algorithm of Backer and Keil, their algorithm is much faster in the case when there are only a few maximal empty boxes. On the other hand, at the expense of giving an $(1-\varepsilon)$-approximation, our algorithm does not enumerate all maximal empty boxes, and achieves efficiency by enumerating all large canonical boxes (to be defined) instead.

Our results are:

(I) In Section 2 we revisit the estimate $A_d(n) = \Theta\left(\frac{1}{n}\right)$ for $d \geq 2$. More precisely: $A_d(n) \geq \frac{1}{n+1}$, and $A_d(n) \geq \left(\frac{5}{4} - o(1)\right) \cdot \frac{1}{n}$. From the other direction we have $A_2(n) < 4 \cdot \frac{1}{n}$, and $A_d(n) < (2^{d-1} \prod_{i=1}^{d-1} p_i) \cdot \frac{1}{n}$ for any $d \geq 3$. Here p_i is the ith prime. While these results were obtained independently by us, the estimate $A_d(n) = \Theta\left(\frac{1}{n}\right)$ for a fixed $d \geq 2$ was previously established

by Rote and Tichy [22]; however, they did not specify the multiplicative constants in the upper bounds as we do. Incidentally, we also remark that our lower bounds are slightly better (in any dimension d).

(II) In Section 3 we exploit the fact that the maximum volume is $\Omega\left(\frac{1}{n}\right)$ in our design of the first efficient $(1-\varepsilon)$-approximation algorithm for finding the largest empty box: Given an axis-parallel d-dimensional box R in $\mathbb{R}^d$ containing n points, there is a $(1-\varepsilon)$-approximation algorithm, running in $O((8ed\varepsilon^{-2})^d \cdot n \log^d n)$ time, for computing a maximum-volume empty axis-parallel box contained in R.

(III) In Section 4 we show that the $\Theta\left(\frac{1}{n}\right)$ estimate also holds for the maximum volume (or area) of an axis-aligned hypercube (or square) amidst n points in $[0,1]^d$. This strengthens the lower bounds for boxes mentioned previously; see (I) above.

(IV) In Section 5 we present a faster $(1-\varepsilon)$-approximation algorithm for finding the largest empty hypercube: Given an axis-parallel d-dimensional hypercube R in $\mathbb{R}^d$ containing n points, there is a $(1-\varepsilon)$-approximation algorithm, running in $O(d^2\varepsilon^{-1} \cdot n \log n + (4d\varepsilon^{-1})^{d+1} \cdot n^{1/d} \log n)$ time, for computing a maximum-volume empty axis-parallel hypercube contained in R.

(V) In Section 6 we derive an $\Omega(n^d)$ lower bound on the number of maximal empty boxes in d-space, for fixed d. This matches the recent $O(n^d)$ upper bound of Backer and Keil [5]. Following their idea, we further narrow the gap between the bounds (in the dependence of d) based on a new closed-form upper bound and a finer estimation.

If B is a box, we refer to the side length of B in the ith coordinate as the extent in the ith coordinate of B. Throughout this paper, $\log n$ and $\ln n$ are the logarithms of n in base 2 and e, respectively, where e denotes Euler's constant.

2 Empty rectangles and boxes

2.1 Empty rectangles in the plane

The lower bound. We start with a very simple-minded lower bound; however, as it turns out, it is very close to optimal. One can immediately see that $A(n) = \Omega(\frac{1}{n})$, by partitioning the unit square with vertical lines through each point: out of at most $n+1$ resulting empty rectangles, the largest rectangle has area at least $\frac{1}{n+1}$. Thus we have:

Proposition 1.

$$A(n) \geq \frac{1}{n+1}. \tag{1}$$

Using the next two lemmas we slightly improve the trivial lower bound $A(n) \geq \frac{1}{n+1}$ in Proposition 1. Let $\xi = \frac{3-\sqrt{5}}{2}$ be the solution in $(0,1)$ of the quadratic equation $(1-x)^2 = x$.

Lemma 1. *Given 2 points in the unit square, there exists an empty rectangle with area at least $\frac{3-\sqrt{5}}{2}$. This bound is tight, i. e., $A(2) = \frac{3-\sqrt{5}}{2} = 0.3819\ldots$.*

Proof. Let $p_1, p_2 \in U$, and assume without loss of generality that $x(p_1) \leq x(p_2)$, and $y(p_1) \geq y(p_2)$. Write $x = x(p_1)$, and $y = y(p_2)$. Consider the three empty rectangles $(0,x) \times (0,1)$, $(0,1) \times (0,y)$, and $(x,1) \times (y,1)$. Their areas are x, y, and $(1-x)(1-y)$, respectively. If $x \geq \xi$ or $y \geq \xi$, we are

done, as one of the first two rectangles has area at least ξ. So we can assume that $x \leq \xi$ and $y \leq \xi$. Then it follows that

$$(1-x)(1-y) \geq (1-\xi)^2 = \xi,$$

so the third rectangle has area at least ξ, as required.

To see that this bound is tight, take $p_1 = (\xi, 1-\xi)$, $p_2 = (1-\xi, \xi)$, and check that the largest empty rectangle has area ξ. □

Lemma 2. *Given 4 points in the unit square, there exists an empty rectangle with area at least $\frac{1}{4}$. This bound is tight, i. e., $A(4) = \frac{1}{4}$.*

Proof. To see that $A(4) \leq \frac{1}{4}$, consider the 4 points $(\frac{1}{4}, \frac{1}{2})$, $(\frac{1}{2}, \frac{1}{4})$, $(\frac{1}{2}, \frac{3}{4})$, $(\frac{3}{4}, \frac{1}{2})$, and check that the largest empty rectangle has area $\frac{1}{4}$. Next we prove the lower bound. Let $S = \{p_1, p_2, p_3, p_4\}$ be a set of 4 points, and assume without loss of generality that they are in lexicographic order: $x(p_1) \leq x(p_2) \leq x(p_3) \leq x(p_4)$, and if $x(p_i) = x(p_j)$ for $i < j$, then $y(p_i) < y(p_j)$. We can also assume that $y(p_1) \leq y(p_2)$. Encode each possible such 4-point configuration by a permutation π of $1, 2, 3, 4$ as follows: for $i < j$, $\pi(i) < \pi(j)$ if and only if $y(p_i) \leq y(p_j)$. For example $\pi = (2, 4, 3, 1)$ encodes the configuration shown in Fig. 1(right).

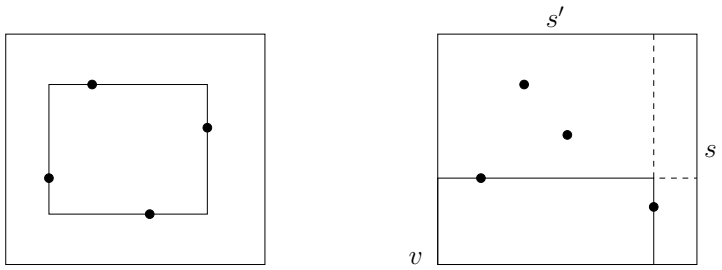

Figure 1: Left: $\pi = (2, 4, 1, 3)$ is special. Right: $\pi = (2, 4, 3, 1)$ is non-special; s is the right side of U, v is the lower left corner of U, and s' is the top side of U.

By our assumption $y(p_1) \leq y(p_2)$, there are only 12 permutations (types) out of the total of $4! = 24$ to consider, those with $\pi(1) < \pi(2)$. Two of these permutations, namely $(2, 4, 1, 3)$ and $(3, 4, 1, 2)$, are called *special*: the 4 points are in convex position and there is an empty rectangle $R \subset U$, with one of these points on each side of R. All the remaining 10 permutations are called *non-special*. We distinguish two cases:

Case 1: S is encoded by a special permutation. For each of the four sides s of U, let $P(s)$ be the largest empty rectangle containing s. See Fig. 1(left) for an example. We can assume that the area of each rectangle $P(s)$ is smaller than $\frac{1}{4}$, since else we are done. But then it follows that each of the four sides of R is longer than $1 - \frac{2}{4} = \frac{1}{2}$, so the area of R is larger than $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, so this case is settled.

Case 2: S is encoded by a non-special permutation. For each of the four vertices v of U, let $Q(v)$ be the largest empty rectangle having v as a vertex. A routine verification shows that for each of the 10 non-special permutations there is a side s of U and a vertex v of U such that (i) $P(s)$ and $Q(v)$ have a common boundary segment, and (ii) v is an endpoint of the side opposite to s. More precisely, if π is one of six permutations $(1, 2, 3, 4)$, $(1, 2, 4, 3)$, $(1, 3, 2, 4)$, $(1, 3, 4, 2)$, $(1, 4, 2, 3)$, $(1, 4, 3, 2)$, then s is the left side, and v is the lower-right corner; if π is one of four permutations $(2, 3, 1, 4)$, $(2, 3, 4, 1)$, $(2, 4, 3, 1)$, $(3, 4, 2, 1)$, then s is the right side, and v is the lower-left corner. See Fig. 1(right) for an example.

As in Case 1, we can assume that the area of $P(s)$ is smaller than $\frac{1}{4}$, thus its shorter side is smaller than $\frac{1}{4}$. By the same token, one of the sides of $Q(v)$ is longer than $1-\frac{1}{4}=\frac{3}{4}$, hence the other side must be shorter than $\frac{1}{3}$, since otherwise the area of $Q(v)$ would exceed $\frac{1}{4}$. Let s' be the side of U adjacent to s and disjoint from v. Consequently, the rectangle R' with side s' and adjacent to $Q(v)$ has the other side longer than $1-\frac{1}{3}=\frac{2}{3}$. Observe that R' has at most two points in its interior. By Lemma 1 and Observation 1, R' contains an empty rectangle of area at least

$$\frac{2}{3}\cdot\xi=\frac{3-\sqrt{5}}{3}=0.254\ldots>\frac{1}{4},$$

as required. This concludes the analysis of the second case.

Thus in both cases, there is an empty rectangle of area at least $\frac{1}{4}$. □

The following observation we need is immediate from invariance under scaling with respect to any of the coordinate axes.

Observation 1. *Assume that $A(n)\geq z$ holds for some n and z. Then, given n points in an axis-aligned rectangle R, there is an empty rectangle contained in R of area at least $z\cdot\mathrm{area}(R)$.*

Theorem 1. *Given n points in the unit square, there exists an empty rectangle with area at least $(\frac{5}{4}-o(1))\cdot\frac{1}{n}$. That is, $A(n)\geq(\frac{5}{4}-o(1))\cdot\frac{1}{n}$.*

Proof. Write $n=5k+r$, for some $k\in\mathbb{N}$ and $r\in\{0,1,2,3,4\}$. Partition U into $k+1$ rectangles of equal width. There exists at least one rectangle R' with at most 4 points in its interior. By Lemma 2 and Observation 1, R' contains an empty rectangle of area at least

$$\frac{1}{4}\cdot\frac{1}{k+1}\geq\frac{5}{4}\cdot\frac{1}{n+5}=\left(\frac{5}{4}-o(1)\right)\cdot\frac{1}{n},$$

as claimed. □

The lower bound derived in the proof, $\frac{5}{4}\cdot\frac{1}{n+5}$, is better than $\frac{1}{n+1}$ for all $n\geq 16$. For $n=5k+4$, the resulting bound is $\frac{5}{4}\cdot\frac{1}{n+1}$. An alternative partition, yielding the same bound in Theorem 1, can be obtained by dividing U into rectangles with vertical lines through every 5th point of the set. Slightly better lower bounds, particularly for small values of n can be obtained by constructing different partitions tailored for specific values of k,r (with a number of points other than 4 in a few of the rectangles), and using estimates on $A(2)$, $A(6)$, etc. For instance, from Lemma 2 we can derive[1] that $A(6)\geq 3-2\sqrt{2}=0.1715\ldots$. Incidentally, we remark that a suitable 6-point construction gives from the other direction that $A(6)<0.2$.

The upper bound. To derive an upper bound of $O(\frac{1}{n})$ on the area of the largest empty rectangle, we use the well-known van der Corput point set defined below. Refer to [4, 25] and [9, Lemma 4A, p. 11] for related results.

Let C_n be the van der Corput set of n points [10, 11], with coordinates $(x(k),y(k))$, $0\leq k\leq n-1$, constructed as follows [7, 18]: Let $x(k)=k/n$. If $k=\sum_{j\geq 0}a_j2^j$ is the binary representation of k, where $a_j\in\{0,1\}$, then $y(k)=\sum_{j\geq 0}a_j2^{-j-1}$. Observe that all points in C_n lie in the unit square $U=[0,1]^2$.

[1] Let $\ell(v)$ and $\ell'(v)$ denote the two orthogonal lines incident to a vertex v of U. It is easy to see that there exists a vertex v of U such that when $\ell(v)$ and $\ell'(v)$ are translated towards the interior of the square, they hit two distinct points out of the six contained in U. By balancing the areas of the two rectangles swept by these two lines, say R_1 and R_2, with the area of the largest empty sub-rectangle inside the rectangle $U\setminus(R_1\cup R_2)$ as guaranteed by Lemma 2 and Observation 1, we get that $A(6)\geq x$, where $x=3-2\sqrt{2}$ is the solution of the quadratic equation $x=(1-x)^2/4$.

Theorem 2. *For the van der Corput set of n points, $C_n \subset U$, the area of the largest empty axis-parallel rectangle is less than $4/n$.*

Proof. Let B be any open empty axis-parallel rectangle inside the unit square. We next show[2] that the area of B is less than $4/n$. Following the presentation in [18, p. 39], a *canonical* interval is an interval of the form $[u \cdot 2^{-v}, (u+1) \cdot 2^{-v})$ for some positive integer v and an integer $u \in [0, 2^v - 1]$.

Let $I_y = [t \cdot 2^{-q}, (t+1) \cdot 2^{-q})$ be the longest canonical interval contained in the projection of the empty rectangle B onto the y-axis (recall that B is open, so this projection is an open interval). Then the side length of B along y must be less than $2 \cdot 2^{-q+1}$ because otherwise the projection would contain a longer canonical interval of length 2^{-q+1}.

Let $k = \sum_{j \geq 0} a_j 2^j$ be the binary representation of an integer k, $0 \leq k \leq n-1$. In the van der Corput construction, a point in C_n with x-coordinate k/n has its y-coordinate in the canonical interval I_y if and only if $t \cdot 2^{-q} \leq \sum_{j \geq 0} a_j 2^{-j-1} < (t+1) \cdot 2^{-q}$, which happens exactly when $\sum_{j=0}^{q-1} a_j 2^{-j-1} = t \cdot 2^{-q}$. In this case, since a_j, $j = 0, \ldots, q-1$, are uniquely determined from the previous equation, $k \bmod 2^q = \sum_{j=0}^{q-1} a_j 2^j$ is a constant $z = z(t,q)$. It then follows that the side length of B along x is at most $2^q/n$. Therefore the area of B is less than $2 \cdot 2^{-q+1} \cdot 2^q/n = 4/n$, as required. □

Corollary 1. $A(n) < 4 \cdot \frac{1}{n}$.

2.2 Empty boxes in higher dimensions

As in the planar case, $A_d(n) \geq \frac{1}{n+1}$ is immediate, by partitioning the hypercube U_d with n axis-parallel hyperplanes, one through each of the n points. By projecting the n points on one of the faces of U_d, and proceeding by induction on d, the lower bound in Theorem 2 carries over here too. Thus we have:

Proposition 2. $A_d(n) \geq \frac{1}{n+1}$. *Moreover,* $A_d(n) \geq \left(\frac{5}{4} - o(1)\right) \cdot \frac{1}{n}$.

We next show that, modulo a constant factor depending on d, this estimate is also best possible. To this end we use the generalization of the van der Corput set to higher dimensions given by Halton and Hammersley [16, 17]. Let H_n be the Halton-Hammersley set of n points [16, 17], with coordinates $(x_0(k), x_1(k), \ldots, x_{d-1}(k))$, $0 \leq k \leq n-1$, constructed as follows [7, 18]: Let p_i be the ith prime number. Each integer k has a unique base-p_i representation $k = \sum_{j \geq 0} a_{i,j} p_i^j$, where $a_{i,j} \in [0, p_i - 1]$. Let $x_0(k) = k/n$, and let $x_i(k) = \sum_{j \geq 0} a_{i,j} p_i^{-j-1}$, $1 \leq i \leq d-1$. Then all points in H_n are inside the unit hypercube $U_d = [0,1]^d$.

Theorem 3. *For the Halton-Hammersley set of n points, $H_n \subset U_d$, the volume of the largest empty axis-parallel box is less than $(2^{d-1} \prod_{i=1}^{d-1} p_i)/n$, where p_i is the ith prime.*

Proof. Let B be any open empty box inside the unit hypercube. We next show that the volume of B is less than $(2^{d-1} \prod_{i=1}^{d-1} p_i)/n$. Generalizing the planar case, a *canonical* interval of the axis x_i, $1 \leq i \leq d-1$, is an interval of the form $[u \cdot p_i^{-v}, (u+1) \cdot p_i^{-v})$ for some positive integer v and an integer $u \in [0, p_i^v - 1]$. Note that $p_1 = 2$.

First consider each axis x_i, $1 \leq i \leq d-1$. Let $I_i = [t_i \cdot p_i^{-q_i}, (t_i+1) \cdot p_i^{-q_i})$ be a longest canonical interval (there could be more than one for $i \geq 2$) contained in the projection of the empty box B

[2]The argument we use here is similar to that used for bounding the geometric discrepancy of the van der Corput set of points.

onto the axis x_i. Then the side length of B along x_i must be less than $2 \cdot p_i^{-q_i+1}$ because otherwise the projection would contain a longer canonical interval of length $p_i^{-q_i+1}$.

Next consider the axis x_0. Let $k = \sum_{j \geq 0} a_{i,j} p_i^j$ be the base-p_i representation of an integer k, $0 \leq k \leq n-1$ and $1 \leq i \leq d-1$. In the Halton-Hammersley construction, a point in H_n with x_0-coordinate k/n has its x_i-coordinate in the canonical interval I_i if and only if $t_i \cdot p_i^{-q_i} \leq \sum_{j \geq 0} a_{i,j} p_i^{-j-1} < (t_i+1) \cdot p_i^{-q_i}$, which happens exactly when $\sum_{j=0}^{q_i-1} a_{i,j} p_i^{-j-1} = t_i \cdot p_i^{-q_i}$. In this case, $k \bmod p_i^{q_i} = \sum_{j=0}^{q_i-1} a_{i,j} p_i^j$ is a constant $z_i = z_i(t_i, q_i)$.

Note that the $d-1$ integers $p_i^{q_i}$, $1 \leq i \leq d-1$, are relatively prime. By the Chinese remainder theorem, it follows that a point in H_n with x_0-coordinate k/n has its x_i-coordinate in the canonical interval I_i for all $1 \leq i \leq d-1$ if and only if $k \bmod \prod_{i=1}^{d-1} p_i^{q_i} = z$ for some integer $z = z(t_1, q_1; \ldots; t_{d-1}, q_{d-1})$. Therefore the side length of B along x_0 is at most $(\prod_{i=1}^{d-1} p_i^{q_i})/n$. Consequently, the volume of B is less than $(\prod_{i=1}^{d-1} 2 \cdot p_i^{-q_i+1}) \cdot (\prod_{i=1}^{d-1} p_i^{q_i})/n = (2^{d-1} \prod_{i=1}^{d-1} p_i)/n$. □

Corollary 2. $A_d(n) < (2^{d-1} \prod_{i=1}^{d-1} p_i) \cdot \frac{1}{n}$.

It is known that $(\prod_{i=1}^{x} p_i)/x^x \to 1$ as $x \to \infty$; see *e. g.*, [23]. Thus asymptotically in d, $A_d(n)$ is bounded from above by $(2d-2)^{d-1}/n$, roughly.

3 A $(1-\varepsilon)$-approximation algorithm for finding the largest empty box

Let R be an axis-parallel d-dimensional box in $\mathbb{R}^d$ containing n points. In this section, we present an efficient $(1-\varepsilon)$-approximation algorithm for computing a maximum-volume empty axis-parallel box contained in R.

Theorem 4. *Given an axis-parallel d-dimensional box R in $\mathbb{R}^d$ containing n points, there is a $(1-\varepsilon)$-approximation algorithm, running in*

$$O\left(\left(\frac{8ed}{\varepsilon^2}\right)^d \cdot n \cdot \log^d n\right)$$

time, for computing a maximum-volume empty axis-parallel box contained in R.

Our algorithm is based on a non-trivial adaptation of the well-known grid method in computational geometry; see for example [24]. This method adopts the standard real-RAM model, and uses the floor function. However, the floor function can be disposed off with a slight increase in the running time; see the remark after Lemma 7.

Overview of the algorithm. By a direct generalization of Observation 1, we can assume w.l.o.g. that $R = [0,1]^d$. Let S be the set of n points contained in R. The algorithm generates a finite set $\mathcal{B}$ of large *canonical boxes*. For each such canonical box $B_0 \in \mathcal{B}$, a corresponding *canonical grid* is considered, and B_0 is placed with its lowest corner at each such grid position and tested for emptiness and containment in R. The one with the largest volume amongst these is returned in the end. We now proceed with the details and first set a few parameters.

Parameters. We assume that $0 < \varepsilon < 1$, and $d \geq 3$, which covers all cases of interest. To somewhat simplify our calculations we also assume that $n \geq 12$. Let us choose parameters

$$\delta = \frac{\varepsilon}{2d}, \quad m = \left\lceil \frac{1}{\delta} \right\rceil = \left\lceil \frac{2d}{\varepsilon} \right\rceil, \quad \text{and } a = \frac{1}{1-\delta}. \tag{2}$$

Let k be the unique positive integer such that

$$a^{k-1} \leq n + 1 < a^k. \tag{3}$$

We next derive some inequalities that follow from this setting. By assumptions $0 < \epsilon < 1$ and $d \geq 3$, we have $\delta = \frac{\varepsilon}{2d} \leq \frac{1}{6}$, and $m \geq 2d/\varepsilon \geq 2d \geq 6$. Then a simple calculation shows that

$$a = \frac{1}{1-\delta} \leq 1 + \frac{6}{5}\delta = 1 + \frac{3\varepsilon}{5d}. \tag{4}$$

It is also clear that $a = \frac{1}{1-\delta} > 1 + \delta$. So a satisfies

$$1 < 1 + \delta < a = \frac{1}{1-\delta} \leq 1 + \frac{6}{5}\delta \leq \frac{6}{5}. \tag{5}$$

Since $n \geq 12$ and $a \leq \frac{6}{5}$, it follows from the second inequality in (3) that $k \geq 15$. We now derive an upper bound on k as a function of n, d and ε. First observe that

$$\log a = \log \frac{1}{1-\delta} \geq \log(1+\delta).$$

We also have

$$\ln(1+\delta) \geq 0.9\delta \quad \text{for } \delta \leq \frac{1}{6}.$$

From (3) we deduce the following sequence of inequalities:

$$k - 1 \leq \frac{\log(n+1)}{\log a} \leq \frac{\log(n+1)}{\log(1+\delta)} = \frac{\log(n+1) \cdot \ln 2}{\ln(1+\delta)} \leq \frac{\log(n+1) \cdot \ln 2}{0.9\delta} \leq \frac{0.78 \log(n+1)}{\delta}. \tag{6}$$

From (6), a straightforward calculation (where we use $n \geq 12$ and $\delta \leq 1/6$) gives

$$k \leq \frac{0.78 \log(n+1)}{\delta} + 1 \leq \frac{0.78 \log(n+1) + 1/6}{\delta} \leq \frac{\log n}{\delta} = \frac{2d}{\varepsilon} \cdot \log n. \tag{7}$$

Canonical boxes and their associated grids. Consider the set $\mathcal{B}$ of *canonical boxes*, whose side lengths are elements of

$$\mathcal{X} = \left\{ \frac{a^i}{a^{k+1}},\ i = 0, 1, \ldots, k-1 \right\}. \tag{8}$$

For a given canonical box $B_0 \in \mathcal{B}$, with sides $X_1, \ldots, X_d \in \mathcal{X}$, consider the *canonical grid associated with* B_0 with points of coordinates

$$\left(\frac{i_1 X_1}{m}, \ldots, \frac{i_d X_d}{m} \right), \quad i_1, \ldots, i_d \geq 0 \tag{9}$$

contained in U_d.

Let B be a maximum-volume empty box in $R = U_d$, with $V_{\max} = \text{vol}(B)$. By the trivial inequality $A_d(n) \geq \frac{1}{n+1}$ of Proposition 2, we have $V_{\max} \geq \frac{1}{n+1}$. This lower bound is crucial in the

design of our approximation algorithm, as it enables us to bound from above the number of large canonical boxes (canonical boxes of smaller volume can be safely ignored in the computation).

Consider the following set $\mathcal{I}$ of $k+1$ intervals

$$\mathcal{I} = \left\{ \left[\frac{a^i}{a^{k+1}}, \frac{a^{i+1}}{a^{k+1}} \right), \ i = 0, 1, \ldots, k \right\}. \tag{10}$$

Observe that for each $i = 1, \ldots, d$, the extent in the ith coordinate of B is at least $\frac{a}{a^{k+1}} = \frac{1}{a^k}$, since otherwise we would have $\mathrm{vol}(B) < \frac{1}{a^k} < \frac{1}{n+1}$, a contradiction. Let Z_i be the extent in the ith coordinate of B, for $i = 1, \ldots, d$. By the above observation, for each $i = 1, \ldots, d$, Z_i belongs to one of the last k intervals in the set $\mathcal{I}$. That is, there exists an integer $y_i \in \{0, 1, \ldots, k-1\}$, such that

$$Z_i \in \left[\frac{a^{y_i+1}}{a^{k+1}}, \frac{a^{y_i+2}}{a^{k+1}} \right). \tag{11}$$

The next lemma shows that B contains an (empty) canonical box with side lengths

$$X_i = \frac{a^{y_i}}{a^{k+1}}, \ i = 1, \ldots, d, \tag{12}$$

at some position in the canonical grid associated with it. We call such a canonical box contained in a maximum-volume empty box, a *large canonical box*. Two key properties of large canonical boxes are proved in Lemma 4 and Lemma 5.

Lemma 3. *If for each $i = 1, \ldots, d$, the extent in the ith coordinate of B belongs to the interval as in* (11)*, then B contains an (empty) large canonical box B_0 with side lengths as in* (12) *at some position in the canonical grid associated with it.*

Proof. It is enough to prove the containment for each coordinate axis i. Let I and I_0 be the corresponding intervals of B and B_0, respectively. Assume for contradiction that the placement of I_0 with its left end point at the first canonical grid position in I is not contained in I. But then we have, by taking into account the grid cell lengths:

$$\frac{a^{y_i+1}}{a^{k+1}} \leq |I| < |I_0| + \frac{|I_0|}{m} \leq |I_0| + \delta \cdot |I_0| = (1+\delta) \frac{a^{y_i}}{a^{k+1}},$$

and consequently,

$$a < 1 + \delta.$$

We reached a contradiction to the 2nd inequality in (5), and the proof is complete. □

We now show that the (empty) large canonical box $B_0 \subset B$ from the previous lemma yields a $(1-\varepsilon)$-approximation for the empty box B of maximum volume.

Lemma 4.

$$\mathrm{vol}(B_0) \geq (1-\varepsilon) \cdot \mathrm{vol}(B).$$

Proof. By (12) and (11),

$$\mathrm{vol}(B_0) = \prod_{i=1}^{d} \frac{a^{y_i}}{a^{k+1}} = \frac{1}{a^{2d}} \prod_{i=1}^{d} \frac{a^{y_i+2}}{a^{k+1}} \geq \frac{1}{a^{2d}} \cdot \mathrm{vol}(B).$$

It remains to be shown that

$$\frac{1}{a^{2d}} \geq 1 - \varepsilon.$$

But this follows from our choice of a and from Bernoulli's inequality:

$$(1+x)^q \geq 1+qx, \text{ for any } x \geq -1, \text{ and any positive integer } q.$$

Indeed,

$$\frac{1}{a^{2d}} = \left(1-\frac{\varepsilon}{2d}\right)^{2d} \geq 1-2d\cdot\frac{\varepsilon}{2d} = 1-\varepsilon,$$

and the proof of Lemma 4 is complete. □

Observe that the number of canonical boxes in $\mathcal{B}$ is exactly k^d, and by (7) is bounded from above as follows:

$$k^d \leq \left(\frac{2d}{\varepsilon}\right)^d \cdot \log^d n. \tag{13}$$

We can prove however a better upper bound on the number of large canonical boxes.

Lemma 5. *The number of large canonical boxes in $\mathcal{B}$ is at most*

$$\left(\frac{2e}{\varepsilon}\right)^d \cdot \log^d n.$$

Proof. Recall that $\mathrm{vol}(B)$ satisfies

$$\frac{1}{a^k} < \frac{1}{n+1} \leq \mathrm{vol}(B) \leq \prod_{i=1}^{d} \frac{a^{y_i+2}}{a^{k+1}} = \frac{a^{2d}\prod_{i=1}^{d} a^{y_i}}{a^{dk+d}} = \frac{a^d \prod_{i=1}^{d} a^{y_i}}{a^{dk}},$$

for some integers $y_i \in \{0,1,\ldots,k-1\}$. Since $a>1$ we deduce that

$$dk-k-d \leq \sum_{i=1}^{d} y_i \leq dk-d, \tag{14}$$

and we want an upper bound on the number of solutions of (14), since a tuple $(y_1,\ldots,y_d)$ uniquely determines a box. Make the substitution $z_i = k-1-y_i$, for $i=1,2,\ldots,d$. So $z_i \in \{0,1,\ldots,k-1\}$, for $i=1,2,\ldots,d$. The above inequalities are equivalent to

$$0 \leq \sum_{i=1}^{d} z_i \leq k. \tag{15}$$

Let t be a nonnegative integer. It is well-known (see for instance [26]) that the number of nonnegative integer solutions of the equation $\sum_{i=1}^{d} z_i = t$ equals $\binom{t+d-1}{d-1}$, that is, when we ignore the upper bound on each z_i. Summing over all values of $t \in \{0,1,\ldots,k\}$, and using a well-known binomial identity (see for instance [26, p. 217]) yields that the number of solutions of (15), hence also of (14), is no more than

$$\sum_{t=0}^{k} \binom{t+d-1}{d-1} = \binom{k+d-1+1}{d-1+1} = \binom{k+d}{d}.$$

A well-known upper bound approximation for binomial coefficients

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k,$$

for positive integers n and k with $1 \le k \le n$, further yields that

$$\binom{k+d}{d} \le \left(\frac{e(k+d)}{d}\right)^d = e^d \left(\frac{k+d}{d}\right)^d. \tag{16}$$

We now check that

$$k + d \le \frac{\log n}{\delta}.$$

Recall inequality (6). A straightforward calculation (where we use $n \ge 12$, $d \ge 3$, and $\varepsilon \le 1$), gives

$$k + d \le \frac{0.78 \log(n+1)}{\delta} + d + 1 = \frac{0.78 \log(n+1) + \frac{d+1}{2d}\varepsilon}{\delta} \le \frac{\log n}{\delta} = \frac{2d}{\varepsilon} \cdot \log n, \tag{17}$$

as claimed. Substituting this upper bound into (16) yields

$$\binom{k+d}{d} \le e^d \left(\frac{2d}{d\varepsilon} \cdot \log n\right)^d = \left(\frac{2e}{\varepsilon}\right)^d \cdot \log^d n, \tag{18}$$

as required. This expression is an upper bound on the number of solutions of (14), hence also on the number of large canonical boxes in $\mathcal{B}$. □

Consider a grid G with cell lengths $x_1, x_2, \ldots, x_d$, superimposed so that the origin of U_d is a grid point. Denote the corresponding grid cells by index tuples $(i_1, i_2, \ldots, i_d)$, where $i_1, i_2, \ldots, i_d \ge 0$. Note that some of the grid cells on the boundary of U_d may be smaller. Given a grid G superimposed on U_d, let $M(G)$ be the number of cells (with nonempty interior) into which U_d is partitioned.

Consider a (fixed) large canonical box, say B_0, with side lengths as in (12). The associated canonical grid, say G_0, has side lengths m times smaller in each coordinate. We now derive an upper bound on the number of canonical grid positions where B_0 is placed and tested for emptiness, according to (9).

Lemma 6. *The number of canonical grid positions for placing B_0 in G_0 is bounded as follows:*

$$M(G_0) \le 12 \cdot \left(\frac{2d}{\varepsilon}\right)^d \cdot n.$$

Proof. We have

$$M(G_0) \le \prod_{i=1}^{d} \left\lceil \frac{m \cdot a^{k+1}}{a^{y_i}} \right\rceil \le \prod_{i=1}^{d} \left(\frac{m \cdot a^{k+1}}{a^{y_i}} + 1 \right).$$

Observe that

$$\frac{m \cdot a^{k+1}}{a^{y_i}} + 1 = \frac{m \cdot a^{k+1} + a^{y_i}}{a^{y_i}} \le \frac{m \cdot a^{k+1} + a^{k-1}}{a^{y_i}} \le \frac{(m+1)a^{k+1}}{a^{y_i}}.$$

By substituting this bound in the product we get that

$$\begin{aligned} M(G_0) &\le \prod_{i=1}^{d} \frac{(m+1) \cdot a^{k+1}}{a^{y_i}} = (m+1)^d \prod_{i=1}^{d} \frac{a^{k+1}}{a^{y_i}} = (m+1)^d \cdot \frac{a^{kd+d}}{\prod_{i=1}^{d} a^{y_i}} \\ &\le (m+1)^d \cdot \frac{a^{kd+d}}{a^{kd-k-d}} = (m+1)^d \cdot a^{2d} \cdot a^k. \end{aligned} \tag{19}$$

For the last inequality above we used (14). We now bound from above each of the three factors in (19). For bounding the second and the third factors we use inequalities (4) and (3), respectively.

$$(m+1)^d = \left(\left\lceil \frac{2d}{\varepsilon} \right\rceil + 1\right)^d \leq \left(\frac{2d}{\varepsilon} + 2\right)^d = \left(\frac{2d}{\varepsilon}\right)^d \left(1 + \frac{\varepsilon}{d}\right)^d \leq \left(\frac{2d}{\varepsilon}\right)^d \cdot e^{\varepsilon} \leq e \left(\frac{2d}{\varepsilon}\right)^d.$$

$$a^{2d} \leq \left(1 + \frac{3\varepsilon}{5d}\right)^{2d} \leq e^{6\varepsilon/5} \leq e^{6/5}.$$

$$a^k = a \cdot a^{k-1} \leq a(n+1) \leq \frac{6}{5} \cdot (n+1) \leq \frac{13n}{10}, \text{ for } n \geq 12.$$

Substituting these upper bounds in (19) gives the desired bound:

$$M(G_0) \leq e^{11/5} \left(\frac{2d}{\varepsilon}\right)^d \cdot \frac{13n}{10} \leq 12 \cdot \left(\frac{2d}{\varepsilon}\right)^d \cdot n. \qquad \square$$

Testing canonical boxes for emptiness. Given a grid with cell lengths $x_1, x_2, \ldots, x_d$, let $n(i_1, i_2, \ldots, i_d)$ denote the number of points from S in cell $(i_1, i_2, \ldots, i_d)$; we refer to these as *cell counts*. For simplicity, we can assume w.l.o.g. that in all the grids generated by the algorithm, no point of S lies on a grid cell boundary. Indeed, the points of S on the boundary of $R = U_d$ can be safely ignored, and the above condition holds with probability 1 if instead of the given ε, the algorithm uses a value chosen uniformly at random from the interval $[(1 - \frac{1}{2d})\varepsilon,\ \varepsilon]$; see also the setting of the parameters in (2). Given a grid G, and integers $M_1, \ldots, M_d \geq 1$, a *grid box with array sizes* $M_1, \ldots, M_d$ is an axis-aligned box whose lower left corner is at a grid point and which spans M_i cells in dimension i, $i = 1, \ldots, d$. All the canonical boxes generated by our algorithm are in fact grid boxes.

The next four lemmas (7, 8, 9, 10) outline the method we use for efficiently computing the number of points in S in a rectangular box, over a sequence of boxes. In particular these boxes can be tested for emptiness within the same specified time.

Lemma 7. *Let G be a grid with cell lengths $x_1, x_2, \ldots, x_d$, superimposed on U_d, with $M(G)$ cells. Then the number of points of S lying in each cell, over all cells, can be computed in $O(d \cdot n + M(G))$ time.*

Proof. The number of points in each of the $M(G)$ cells is initialized to 0. For each point $p \in S$, its cell index tuple (label) is computed in $O(d)$ time using the floor function for each coordinate, and the corresponding cell count is updated. $\square$

Remark. If the floor function is not an option, the number of points in each cell can be computed using binary search for each coordinate. The resulting time complexity is $O(n \cdot \log M(G))$.

Denote by $N(i_1, i_2, \ldots, i_d)$ the number of points in S in the box with lower left cell $(0, 0, \ldots, 0)$, and upper right cell $(i_1, i_2, \ldots, i_d)$; we refer to the numbers $N(i_1, i_2, \ldots, i_d)$ as *corner box* numbers.

Lemma 8. *Given a grid G with cell lengths $x_1, x_2, \ldots, x_d$ placed at the origin, with $M(G)$ cells, and grid cell counts $n(i_1, i_2, \ldots, i_d)$, over all cells, the corner box numbers $N(i_1, i_2, \ldots, i_d)$, over all cells, can be computed in $O(2^d \cdot M(G))$ time.*

Proof. Define $N(i_1, i_2, \ldots, i_d) = 0$, if $i_j < 0$ for some j. Let $b = (b_1, b_2, \ldots, b_d) \in \{0, 1\}^d$ be a binary vector. Let the *parity* of b be $\pi(b) = b_1 \oplus b_2 \oplus \cdots \oplus b_d$, where $\oplus$ is the binary exclusive or operation.

By the inclusion-exclusion principle, the corner box numbers are given by the following formula with at most 2^d operations:

$$N(i_1, i_2, \ldots, i_d) = n(i_1, i_2, \ldots, i_d) + \sum_{\substack{b=(b_1,b_2,\ldots,b_d) \\ b\neq(0,0,\ldots,0)}} (-1)^{\pi(b)+1} N(i_1 - b_1, i_2 - b_2, \ldots, i_d - b_d).$$

Since G has $M(G)$ cells, the bound follows. □

Lemma 9. *Consider a grid G with cell lengths $x_1, x_2, \ldots, x_d$ placed at the origin, with $M(G)$ cells, and corner box numbers $N(i'_1, i'_2, \ldots, i'_d)$, over all cells. Let B_0 be a (canonical) grid box with array sizes $M_1, \ldots, M_d \geq 1$, and lower left cell $(i_1, i_2, \ldots, i_d)$. Then the number of points of S in B_0, denoted $N(B_0)$, can be computed in $O(2^d)$ time.*

Proof. Let $j_1 = i_1 + M_1 - 1, \ldots, j_d = i_d + M_d - 1$ be the upper right cell of B_0. By the inclusion-exclusion principle, the corner box number $N(j_1, j_2, \ldots, j_d)$ can be computed as follows:

$$N(j_1, j_2, \ldots, j_d) = N(B_0) + \sum_{\substack{b=(b_1,b_2,\ldots,b_d) \\ b\neq(0,0,\ldots,0)}} (-1)^{\pi(b)+1} N(j_1 - b_1 M_1, j_2 - b_2 M_2, \ldots, j_d - b_d M_d).$$

Hence $N(B_0)$ can be extracted from the above formula with at most 2^d operations. □

Let $Q(i_1, i_2, \ldots, i_d)$ be the number of points in S in the canonical box with array sizes $M_1, \ldots, M_d \geq 1$, and lower left cell $(i_1, i_2, \ldots, i_d)$.

Lemma 10. *Consider a grid G with cell lengths $x_1, x_2, \ldots, x_d$ placed at the origin, with $M(G)$ cells, and corner box numbers $N(i'_1, i'_2, \ldots, i'_d)$, over all cells. Then the numbers (counts) $Q(i_1, i_2, \ldots, i_d)$, over all cells, can be computed in $O(2^d \cdot M(G))$ time.*

Proof. There are $M(G)$ cells determined by G in U_d, and for each, apply the bound of Lemma 9. □

The last step in the proof of Theorem 4. For each canonical box, say B_0, there is a unique associated canonical grid, say G_0. The time taken to test B_0 for emptiness and containment in R when placed at all grid positions in G_0, is obtained by adding the running times in lemmas 7, 8, and 10:

$$O(d \cdot n + M(G_0) + 2^d \cdot M(G_0)) = O\left(2^d \cdot \left(\frac{2d}{\varepsilon}\right)^d \cdot n\right) = O\left(\left(\frac{4d}{\varepsilon}\right)^d \cdot n\right), \tag{20}$$

where we have used the upper bound on $M(G_0)$ in Lemma 6. By multiplying this with the upper bound on the number of large canonical boxes in Lemma 5, we get that the total running time of the approximation algorithm is

$$O\left(\left(\frac{2e}{\varepsilon}\right)^d \cdot \log^d n \cdot \left(\frac{4d}{\varepsilon}\right)^d \cdot n\right) = O\left(\left(\frac{8ed}{\varepsilon^2}\right)^d \cdot n \cdot \log^d n\right). \tag{21}$$

The proof of Theorem 4 is now complete.

4 Empty squares and hypercubes

For any dimension $d \geq 2$, given a set S of n points in the unit hypercube $U_d = [0,1]^d$, let $A'_d(S)$ be the maximum volume of an empty axis-parallel hypercube contained in U_d, and let $A'_d(n)$ be the minimum value of $A_d(S)$ over all sets S of n points in U_d. Then for any fixed dimension d, our next theorem shows that $A'_d(n) = \Theta\left(\frac{1}{n}\right)$, too:

Theorem 5. *For a fixed d, $A'_d(n) = \Theta\left(\frac{1}{n}\right)$. More precisely,*

$$\frac{1}{(n^{1/d}+1)^d} \leq A'_d(n) \leq \frac{1}{(\lfloor n^{1/d} \rfloor + 1)^d}. \tag{22}$$

Proof. We first prove the lower bound. Let S be a set of n points in the unit hypercube U_d. Let x be a positive number to be determined. Let X be an axis-parallel hypercube of side $1-x$ that is concentric with U_d. For each point $p \in S$, place an axis-parallel (open) hypercube of side x centered at p. If there is a point $q \in X$ that is not covered by the union of the n hypercubes, then the axis-parallel hypercube of side x centered at q is an empty hypercube contained in U_d.

The volume[3] of X is $(1-x)^d$. The total volume of n hypercubes of side x is nx^d. Let x be the positive solution to the following equation

$$(1-x)^d = nx^d.$$

The solution is $x = \frac{1}{n^{1/d}+1}$. For this value of x, either the n small hypercubes cover X with no interior overlap among themselves, or there they don't cover X. In either case, there exists an open axis-parallel hypercube of side length x, centered at a point in X, and empty of points in S. Consequently,

$$A'_d(n) \geq x^d = \frac{1}{(n^{1/d}+1)^d}.$$

We next prove the upper bound. Let $k = \lfloor n^{1/d} \rfloor$. Note that $n \geq k^d$. Partition the unit hypercube U_d into a $(k+1) \times \ldots \times (k+1)$ axis-aligned d-dimensional grid of cell length $1/(k+1)$. Place a point at each of the k^d grid vertices in the interior of U_d. Then any axis-parallel hypercube contained in U_d whose side is longer than $1/(k+1)$, must be non-empty. Consequently,

$$A'_d(n) \leq \frac{1}{(k+1)^d} = \frac{1}{(\lfloor n^{1/d} \rfloor + 1)^d}.$$

It remains to show that (22) implies that for a fixed d, we have $A'_d(n) = \Theta\left(\frac{1}{n}\right)$. The following inequalities are straightforward:

$$n^{1/d} + 1 \leq 2n^{1/d} \quad \Rightarrow \quad (n^{1/d}+1)^d \leq 2^d n,$$

$$n^{1/d} \leq \lfloor n^{1/d} \rfloor + 1 \quad \Rightarrow \quad n \leq (\lfloor n^{1/d} \rfloor + 1)^d.$$

Putting them together yields

$$\frac{1}{2^d n} \leq \frac{1}{(n^{1/d}+1)^d} \leq A'_d(n) \leq \frac{1}{(\lfloor n^{1/d} \rfloor + 1)^d} \leq \frac{1}{n},$$

as claimed. □

[3]For $d = 2$ volume is replaced by area throughout this proof.

5 A $(1-\varepsilon)$-approximation algorithm for finding a largest empty hypercube

Let R be an axis-parallel d-dimensional hypercube in $\mathbb{R}^d$ containing n points. In this section, we present an efficient $(1-\varepsilon)$-approximation algorithm for computing a maximum-volume empty axis-parallel hypercube contained in R. An exact algorithm for this problem running in $O(n^{d/2}\log n)$ time was devised by Backer and Keil [6]; observe that this algorithm is faster than the exact algorithm for finding a maximum-volume empty axis-parallel box mentioned in the introduction. With approximation algorithms the situation is analogous, and we are able to obtain a faster $(1-\varepsilon)$-approximation algorithm for finding the largest hypercube:

Theorem 6. *Given an axis-parallel d-dimensional hypercube R in $\mathbb{R}^d$ containing n points, there is a $(1-\varepsilon)$-approximation algorithm, running in*

$$O\left(\frac{d^2}{\varepsilon}\cdot n\log n+\left(\frac{4d}{\varepsilon}\right)^{d+1}\cdot n^{1/d}\log n\right)$$

time, for computing a maximum-volume empty axis-parallel hypercube contained in R.

The overall structure of the algorithm is similar to that for finding the largest empty box. We can assume w.l.o.g. that $R=U_d=[0,1]^d$, $n\geq 12$, and $d\geq 3$. Recall that, by Theorem 5, the volume of a largest empty hypercube in U_d is at least $(n^{1/d}+1)^{-d}$. We set the parameters δ, m and a as in equation (2). Inequalities (4) and (5) also follow. Let now k be the unique positive integer such that

$$a^{k-1}\leq n^{1/d}+1<a^k. \tag{23}$$

Thus

$$a^{k-1}\leq n^{1/d}+1\leq 2n^{1/d}.$$

Since $n\geq 12$ and $d\geq 3$ we have

$$k-1\leq\frac{1+\frac{1}{d}\log n}{\log a}\leq\frac{(\frac{1}{3}+\frac{1}{d})\log n}{\log a}\leq\frac{2}{3}\cdot\frac{\log n}{\log a}\leq\frac{2\log n}{3}\cdot\frac{\ln 2}{0.9\delta}\leq\frac{3\log n}{5\delta}.$$

It follows that

$$k\leq 1+\frac{3\log n}{5\delta}\leq\frac{\log n}{\delta}=\frac{2d}{\varepsilon}\cdot\log n. \tag{24}$$

Consider the set $\mathcal{H}$ of k *canonical hypercubes* whose sides are elements of $\mathcal{X}$ (as in (8)):

$$\mathcal{X}=\left\{\frac{a^i}{a^{k+1}},\ i=0,1,\ldots,k-1\right\}. \tag{25}$$

For a given canonical hypercube $H_0\in\mathcal{H}$, with side $X\in\mathcal{X}$, consider the *canonical grid associated with H_0* with points of coordinates

$$\left(\frac{i_1X}{m},\ldots,\frac{i_dX}{m}\right),\quad i_1,\ldots,i_d\geq 0 \tag{26}$$

contained in U_d.

Consider the set $\mathcal{I}$ of $k+1$ intervals (as in (10)):

$$\mathcal{I}=\left\{\left[\frac{a^i}{a^{k+1}},\frac{a^{i+1}}{a^{k+1}}\right),\ i=0,1,\ldots,k\right\}. \tag{27}$$

Let H be a maximum-volume empty hypercube in $R = U_d$, with side length Z and $V_{\max} = \mathrm{vol}(H)$. Observe that $Z \geq \frac{a}{a^{k+1}}$: indeed, $Z < \frac{a}{a^{k+1}}$ would imply that

$$Z^d < \frac{1}{a^{kd}} < \frac{1}{(n^{1/d}+1)^d},$$

in contradiction to the lower bound in Theorem 5. This means that Z belongs to one of the last k intervals in the set $\mathcal{I}$. That is, there exists an integer $y \in \{0, 1, \ldots, k-1\}$, such that

$$Z \in \left[\frac{a^{y+1}}{a^{k+1}}, \frac{a^{y+2}}{a^{k+1}}\right). \tag{28}$$

Analogous to Lemma 3, we conclude that H contains a *large canonical hypercube*, say H_0, whose side is

$$X = \frac{a^y}{a^{k+1}}, \tag{29}$$

at some position in the canonical grid associated with it. Analogous to Lemma 4, we show that $\mathrm{vol}(H_0) \geq (1-\varepsilon) \cdot \mathrm{vol}(H)$: By (29) and (28),

$$\mathrm{vol}(H_0) = \left(\frac{a^y}{a^{k+1}}\right)^d = \frac{1}{a^{2d}} \cdot \left(\frac{a^{y+2}}{a^{k+1}}\right)^d \geq \frac{1}{a^{2d}} \cdot \mathrm{vol}(H) \geq (1-\varepsilon) \cdot \mathrm{vol}(H),$$

since the setting of a is the same as before. Analogous to Lemma 5, now (24) is the upper bound we need on the number of canonical hypercubes. The bound in Lemma 6 needs to be adjusted because k is chosen differently, and we have a different upper bound on the third factor in the product, a^k. From the definition of k in (23) and from (5) we deduce

$$a^k = a \cdot a^{k-1} \leq a(n^{1/d}+1) \leq 2an^{1/d} \leq \frac{12}{5} n^{1/d}.$$

The resulting bound analogous to that in Lemma 6 is now

$$M(G_0) \leq e^{11/5} \left(\frac{2d}{\varepsilon}\right)^d \cdot \frac{12}{5} n^{1/d} \leq 22 \cdot \left(\frac{2d}{\varepsilon}\right)^d \cdot n^{1/d}. \tag{30}$$

The time taken to test H_0 for emptiness and containment in R when placed at all relevant grid positions is now

$$O(d \cdot n + M(G_0) + 2^d \cdot M(G_0)) = O\left(dn + 2^d \cdot \left(\frac{2d}{\varepsilon}\right)^d \cdot n^{1/d}\right) = O\left(dn + \left(\frac{4d}{\varepsilon}\right)^d \cdot n^{1/d}\right).$$

By multiplying this with the upper bound in (24), on the number of canonical hypercubes, we get that the total running time of the approximation algorithm is

$$O\left(\frac{d^2}{\varepsilon} \cdot n \log n + \left(\frac{4d}{\varepsilon}\right)^{d+1} \cdot n^{1/d} \log n\right).$$

The proof of Theorem 6 is now complete.

6 The number of maximal empty boxes

In this section we prove the following theorem:

Theorem 7. *Let $U_d = [0,1]^d$ be the unit hypercube in $\mathbb{R}^d$. For any $n \geq 0$, there exist n points in U_d such that the number of maximal empty boxes in U_d is at least $(\lfloor \frac{n}{d} \rfloor + 1)^d$. On the other hand, the number of maximal empty boxes determined by any set of n points in U_d is at most $\binom{n+2d}{d} \cdot \binom{2d}{d} \cdot 2^{2d} + 2d$.*

For the proof of the lower bound, we use the following lemma:

Lemma 11. *For any d integers $n_i \geq 0$, $1 \leq i \leq d$, let $n = \sum_{i=1}^{d} n_i$, and let*

$$U'_d = [-(n_d+1), n_1+1] \times [-(n_1+1), n_2+1] \times \cdots \times [-(n_{d-1}+1), n_d+1].$$

Then there exist n points in U'_d such that the number of maximal empty boxes in U'_d is at least $\prod_{i=1}^{d}(n_i+1)$.

Proof. Let $\pm\vec{x}_1, \ldots, \pm\vec{x}_d$ be the positive and negative unit vectors along the d axes of $\mathbb{R}^d$. Partition these $2d$ vectors into d groups of orthogonal vectors,

$$\{+\vec{x}_1, -\vec{x}_2\}, \{+\vec{x}_2, -\vec{x}_3\}, \ldots, \{+\vec{x}_{d-1}, -\vec{x}_d\}, \{+\vec{x}_d, -\vec{x}_1\},$$

with one positive vector and one negative vector in each group. For convenience, let $\vec{x}_{d+1} = \vec{x}_1$. Then, for each group of two orthogonal vectors $\{+\vec{x}_i, -\vec{x}_{i+1}\}$, $1 \leq i \leq d$, place a sequence of n_i points in U'_d:

$$k\vec{x}_i - (n_i + 1 - k)\vec{x}_{i+1}, \quad 1 \leq k \leq n_i.$$

Next compose $n_i + 1$ pairs of half-spaces in $\mathbb{R}^d$:

$$x_i < k \quad \text{and} \quad x_{i+1} > -(n_i + 2 - k), \qquad 1 \leq k \leq n_i + 1.$$

There are $\prod_{i=1}^{d}(n_i+1)$ combinations of d pairs of half-spaces, one pair for each group of orthogonal vectors. We claim that for each combination, the intersection of the d pairs of half-spaces is a maximal empty box.

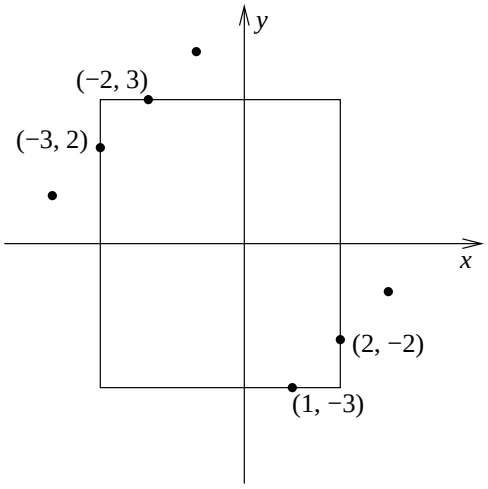

Figure 2: An example of the construction.

We refer to Fig. 2 for an example of the planar case. For $n_1 = 3$, $n_2 = 4$, and $n = 7$, the four unit vectors $\pm\vec{x}$ and $\pm\vec{y}$ are grouped into $\{+\vec{x}, -\vec{y}\}$ and $\{+\vec{y}, -\vec{x}\}$. The corresponding two sequences of points have the following (x, y)-coordinates:

$$(1, -3) \quad (2, -2) \quad (3, -1)$$
$$(-4, 1) \quad (-3, 2) \quad (-2, 3) \quad (-1, 4).$$

The following combination of two pairs of half-planes

$$x < 2 \quad \text{and} \quad y > -3$$
$$y < 3 \quad \text{and} \quad x > -3,$$

yields the maximal empty box $(-3, 2) \times (-3, 3)$.

We now prove our claim that the intersection of the d pairs of half-spaces in each combination is a maximal empty box. Observe that for each pair of half-spaces $x_i < a$ and $x_{i+1} > -b$ for the group $\{+\vec{x}_i, -\vec{x}_{i+1}\}$, we have property (i) that the intersection of the two half-spaces contains no points in this sequence, and property (ii) that each point that is on the boundary of one half-space is in the interior of the other half-space. Moreover, since the ith and $(i + 1)$th coordinates of the points in the other sequences are either zero or different in sign from the points in this sequence, we have property (iii) that each of the two half-spaces contains all points in the other sequences.

Consider the box B that is the intersection of the d pairs of half-spaces in any of the $\prod_{i=1}^{d}(n_i+1)$ combinations. By (i), B must be empty. By (ii) and (iii), each point that is on the boundary of some half-space is not only in the interior of the other half-space in the pair, but is also in the interior of the other $d-1$ pairs of half-spaces. This implies that for each face of B that is not flush with a bounding face of U'_d, there must be a point in the interior of the face. Hence B is maximal. In summary, for each of the $\prod_{i=1}^{d}(n_i+1)$ combinations, the intersection of the d pairs of half-spaces in the combination is a maximal empty box. □

By scaling and translating, the n points the hyperrectangle U'_d can be placed into the hypercube U_d. Choose $n_i \geq \lfloor n/d \rfloor$ for all $1 \leq i \leq d$ such that $n = \sum_{i=1}^{d} n_i$. Then Lemma 11 implies the lower bound in Theorem 7.

We note that the same lower bound was obtained independently and simultaneously by Backer and Keil [5], who also obtained a matching $O(n^d)$ upper bound when d is a constant. In the following, we borrow their deflation-inflation idea to derive a more precise upper bound as a closed-form function of n and d, then compare it with our closed-form lower bound to estimate their ratio as a function of d.

Fix an arbitrary order of the $2d$ orthogonal directions. Assume without of loss of generality that the n points have distinct coordinates along each axis and that there are no points on the boundary of the unit hypercube U_d. For each empty box B, we denote by $p(B)$ the number of points on the boundary of B, and denote by $q(B)$ the number of faces of B that are flush with the faces of U_d. Under our assumption of distinct coordinates, $p(B)$ can be equivalently defined as the number of faces of B that contain a point in their interior. Clearly $p(B) + q(B) \leq 2d$, with equality if and only if B is maximal.

We first consider the set $\mathcal{B}_d$ of maximal empty boxes with no face flush with any face of U_d. Let A be such a maximal empty box, which has one point in the interior of each face. Now consider the $2d$ faces of the box following the fixed order of the $2d$ orthogonal directions. For each face f, if it contains a single point in its interior, deflate the box by pushing the face f toward its opposite face until it contains another point on its boundary. Prior to the deflation such a point was either in the interior of one of the $(2d-2)$ faces adjacent to f, or on the shared boundary of two or more

of these faces. The original point of contact in the interior of f remains in the exterior of the box after deflation.

Initially $p(A) = 2d$, and after d such deflations, we obtain an empty box $A' \subset A$ such that $p(A') = d$. Note that A' is minimal in the sense that no box $A'' \subset A'$ satisfies $p(A'') = p(A')$. To recover the original box A from A', it suffices to inflate the box at the d faces in reverse order, by pushing each face away from its opposite face until it contains a point in its interior. Thus the number of maximal empty boxes A in $\mathcal{B}_d$ is at most the number of deflated boxes A' times the number of combinations of d faces chosen from the $2d$ faces, that is, at most $\binom{n}{d} \cdot \binom{2d}{d}$. We remark that our upper bound on $|\mathcal{B}_d|$ is slightly sharper than the $\binom{n}{d} \cdot 2^{2d}$ upper bound on $|\mathcal{B}_d|$ given by Backer and Keil [5].

We next consider the set $\mathcal{B}'_d$ of maximal empty boxes with at least one face flush with some face of U_d. Backer and Keil [5] estimated $|\mathcal{B}'_d|$ to be $O(n^d)$ (when d is constant) by an inductive argument that depends on the value of $|\mathcal{B}_{d-1} \cup \mathcal{B}'_{d-1}|$. We show that a small variation of the same deflation-inflation idea gives a closed-form formula for $|\mathcal{B}_d \cup \mathcal{B}'_d|$.

Let B be any maximal empty box with $p(B) \geq 2$. Consider the $2d$ faces of the box in the order such that (i) the faces that are flush with the faces of U_d precede the faces that contain a point in their interior, and (ii) the faces in each of the two groups follow the fixed order of the $2d$ orthogonal directions. (We use such a particular order and require that $p(B) \geq 2$ to ensure that the box always has at least two points on its boundary and hence does not become flat after the sequence of d deflations. Note that the deflation of any face of the box that is flush with some face of U_d does not reduce the number of points on the boundary of the box. By our assumption of distinct coordinates, any two of the n points determine a non-degenerate box with the two points at two opposite corners of a main diagonal; this non-degenerate box is contained in any box that has the two points on its boundary.) For each face f, if it is flush with a face of U_d or contains a single point in its interior, deflate the box by pushing the face f toward its opposite face until it contains another point on its boundary (this point was either in the interior of one of the $(2d-2)$ faces adjacent to f, or on the shared boundary of two or more of these faces).

Initially $p(B) + q(B) = 2d$, and after d such deflations, we obtain an empty box $B' \subset B$ such that $p(B') + q(B') = d$. Note that B' is minimal in the sense that no box $B'' \subset B'$ satisfies $p(B'') = p(B')$ and $q(B'') = q(B')$. To recover the original box B from B', it suffices to inflate the box at the d faces in reverse order, by pushing each face away from its opposite face until it either contains a point in its interior or is flush with a face of U_d. Thus the number of maximal empty boxes B in $\mathcal{B}_d \cup \mathcal{B}'_d$ with $p(B) \geq 2$ is at most the number of deflated boxes B', times the number of combinations of d faces chosen from the $2d$ faces, times the number of partitions of the $2d$ faces into the two groups, that is, at most $\binom{n+2d}{d} \cdot \binom{2d}{d} \cdot 2^{2d}$. On the other hand, when $n > 0$, there are exactly $2d$ maximal empty boxes B with $p(B) = 1$. Thus the total number of maximal empty boxes is at most $\binom{n+2d}{d} \cdot \binom{2d}{d} \cdot 2^{2d} + 2d$. This completes the proof of our upper bound in Theorem 7.

A routine calculation (below) using estimates such as $\binom{n}{d} \leq n^d/d!$ and the Stirling's formula for the factorial shows that, when $n \gg d$, the ratio of the upper bound to the lower bound in Theorem 7 is approximately at most $\frac{(16e)^d}{\sqrt{2}\pi d}$.

Since $\binom{n}{d} \leq n^d/d!$ and $\binom{2d}{d} = (2d)!/(d!)^2$, we have

$$\binom{n+2d}{d} \cdot \binom{2d}{d} \leq (n+2d)^d \frac{(2d)!}{(d!)^3}.$$

By Stirling's formula, $d! = \sqrt{2\pi d}(d/e)^d(1 + O(1/d))$, hence

$$\frac{(2d)!}{(d!)^3} = \frac{\sqrt{2\pi 2d}(2d/e)^{2d}}{(\sqrt{2\pi d}(d/e)^d)^3}\big(1 \pm O(1/d)\big) = \frac{(4e/d)^d}{\sqrt{2}\pi d}\big(1 \pm O(1/d)\big).$$

Thus

$$\binom{n+2d}{d} \cdot \binom{2d}{d} \leq (n+2d)^d \frac{(4e/d)^d}{\sqrt{2}\pi d}\big(1 \pm O(1/d)\big).$$

Consequently,

$$\frac{\binom{n+2d}{d} \cdot \binom{2d}{d} \cdot 2^{2d} + 2d}{(\lfloor \frac{n}{d} \rfloor + 1)^d} \leq \frac{(n+2d)^d \frac{(4e/d)^d}{\sqrt{2}\pi d}\big(1 \pm O(1/d)\big)4^d + 2d}{(\lfloor \frac{n}{d} \rfloor + 1)^d},$$

which is approximately $\frac{(16e)^d}{\sqrt{2}\pi d}$ for large n.

7 Concluding remarks

Reducing the gap between the lower and upper bounds on $A_d(n)$, particularly in higher dimensions remains an interesting problem. Other questions can be asked regarding the computational complexity of computing maximum-volume empty ranges of other types than axis-aligned boxes. We list some specific questions and directions for further study:

(1) Is the dependence on d necessary in the upper bound on $A_d(n)$ as given by our Theorem 3, or is $A_d(n) \leq \frac{C}{n}$, where C is an absolute constant? As a preliminary question: Given d points in the unit hypercube $[0,1]^d$, is there always an empty box of volume C, where C is an absolute constant, or does $A_d(d)$ tend to zero with the dimension?

(2) In their article on the dispersion of point sequences, Rote and Tichy have also studied other empty ranges besides axis-parallel boxes, such as balls and arbitrary oriented rectangular boxes. Efficiently computing (or approximating) the largest (volume) empty ball, rectangular box, simplex, or convex polytope amidst n points in a bounded d-dimensional container is of obvious interest.

Acknowledgment The authors would like to thank the anonymous reviewers for careful reading and thoughtful suggestions.

References

[1] A. Aggarwal and S. Suri, Fast algorithms for computing the largest empty rectangle, in: *Proceedings of the 3rd Annual Symposium on Computational Geometry*, 1987, pp. 278–290.

[2] M. Atallah and G. Fredrickson, A note on finding the maximum empty rectangle, *Discrete Applied Mathematics*, **13** (1986), 87–91.

[3] M. Atallah and S. R. Kosaraju, An efficient algorithm for maxdominance, with applications, *Algorithmica*, **4** (1989), 221–236.

[4] J. Beck and W. Chen, *Irregularities of Distribution*, Cambridge University Press, 1987.

[5] J. Backer and M. Keil, The bichromatic rectangle problem in high dimensions, in: *Proceedings of the 21st Canadian Conference on Computational Geometry*, 2009, pp. 157–160.

[6] J. Backer and M. Keil, The mono- and bichromatic empty rectangle and square problems in all dimensions, in: *Proceedings of the 9th Latin American Symposium on Theoretical Informatics*, 2010, pp. 14–25.

[7] B. Chazelle, *The Discrepancy Method: Randomness and Complexity*, Cambridge University Press, 2000.

[8] B. Chazelle, R. Drysdale and D. T. Lee, Computing the largest empty rectangle, *SIAM J. Comput.*, **15** (1986), 300–315.

[9] W. Chen, Lectures on irregularities of point distribution, web edition, `http://rutherglen.ics.mq.edu.au/wchen/researchfolder/iod00.pdf`, 2000.

[10] J. G. van der Corput, Verteilungsfunktionen I., *Proc. Nederl. Akad. Wetensch.*, **38** (1935), 813–821.

[11] J. G. van der Corput, Verteilungsfunktionen II., *Proc. Nederl. Akad. Wetensch.*, **38** (1935), 1058–1066.

[12] A. Datta, Efficient algorithms for the largest empty rectangle problem, *Inform. Sci.*, **64** (1992), 121–141.

[13] A. Datta and S. Soundaralakshmi, An efficient algorithm for computing the maximum empty rectangle in three dimensions, *Inform. Sci.*, **128** (2000), 43–65.

[14] A. Dumitrescu and M. Jiang, On the largest empty axis-parallel box amidst n points, manuscript, 2009; `http://arxiv.org/abs/0909.3127v2`.

[15] J. Edmonds, J. Gryz, D. Liang, and R. Miller, Mining for empty spaces in large data sets, *Theoretical Computer Science*, **296(3)** (2003), 435–452.

[16] J. H. Halton, On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals, *Numer. Math.*, **2** (1960), 84–90.

[17] J. M. Hammersley, Monte Carlo methods for solving multivariable problems, *Ann. New York Acad. Sci.*, **86** (1960), 844–874.

[18] J. Matoušek, *Geometric Discrepancy: An Illustrated Guide*, Springer, 1999.

[19] M. McKenna, J. O'Rourke, and S. Suri, Finding the largest rectangle in an orthogonal polygon, in: *Proceedings of the 23rd Annual Allerton Conference on Communication, Control and Computing*, Urbana-Champaign, Illinois, October 1985.

[20] A. Namaad, D. T. Lee, and W.-L. Hsu, On the maximum empty rectangle problem, *Discrete Applied Mathematics*, **8** (1984), 267–277.

[21] M. Orlowski, A new algorithm for the largest empty rectangle problem, *Algorithmica*, **5** (1990), 65–73.

[22] G. Rote and R. F. Tichy, Quasi-Monte-Carlo methods and the dispersion of point sequences, *Mathematical and Computer Modelling*, **23** (1996), 9–23.

[23] S. M. Ruiz, A result on prime numbers, *Mathematical Gazette*, **81** (1997), 269–270.

[24] M. Smid, Closest-point problems in computational geometry, in: *Handbook of Computational Geometry*, J.-R. Sack and J. Urrutia, editors, pp. 877–932, North Holland, 1999.

[25] S. Srinivasan, On two-dimensional Hammersley's sequences, *Journal of Number Theory*, **10** (1978), 421–429.

[26] A. Tucker, *Applied Combinatorics*, 3rd edition, Wiley, 1995.