Discrete Comput Geom 25:519–524 (2001)
DOI: 10.1007/s00454-001-0014-2

Discrete & Computational
Geometry

The Discrepancy of Boxes in Higher Dimension*

B. Chazelle[1] and A. Lvov[2]

[1]Department of Computer Science, Princeton University,
Princeton, NJ 08544, USA
chazelle@cs.princeton.edu
and
NEC Research Institute, 4 Independence Way,
Princeton, NJ 08540, USA

[2]Program in Applied and Computational Mathematics,
Princeton University, Princeton, NJ 08544, USA
lvov@math.princeton.edu

Abstract. We prove that the red–blue discrepancy of the set system formed by n points and n axis-parallel boxes in $\mathbf{R}^d$ can be as high as $n^{\Omega(1)}$ in any dimension $d = \Omega(\log n)$. This contrasts with the fixed-dimensional case $d = O(1)$, where the discrepancy is always polylogarithmic. More generally we show that in any dimension $1 < d = O(\log n)$ the maximum discrepancy is $2^{\Omega(d)}$. Our result also leads to a new lower bound on the complexity of off-line orthogonal range searching. This is the problem of summing up weights in boxes, given n weighted points and n boxes in $\mathbf{R}^d$. We prove that the number of arithmetic operations is at least $\Omega(nd + n \log\log n)$ in any dimension $d = O(\log n)$.

1. Introduction

A set system formed by n points and n axis-parallel boxes in $\mathbf{R}^d$ is characterized by its incidence matrix A, where $A_{i,j} = 1$ if the ith box contains the jth point and $A_{i,j} = 0$ otherwise. The red–blue discrepancy of the set system is the minimum value of $\|Ax\|_\infty$ over all $x \in \{-1, 1\}^n$. We prove that in any dimension $d = \Omega(\log n)$ some set systems have discrepancy in $n^{\Omega(1)}$. Interestingly, our lower bound also holds for the Hamming cube $\{0, 1\}^d$. More generally we show that in any dimension $d = O(\log n)$ the maximum discrepancy is $2^{\Omega(d)}$.

It was already known [4] that in dimension $O(\log n / \log\log n)$ the discrepancy could be as high as $n^{\Omega(1/\log\log n)}$, but the dimension at which the discrepancy became polynomial

* This work was supported in part by NSF Grant CCR-96-23768, ARO Grant DAAH04-96-1-0181, and NEC Research Institute.

was left unresolved. We show that it is precisely $\Theta(\log n)$. Quite different from the number-theoretic construction of [4], our proof is purely probabilistic. It is interesting to contrast our result with the discrepancy of boxes in fixed dimension. Throughout this paper we assume that $d > 1$. The discrepancy of boxes in $\mathbf{R}^d$ is bounded by $O(\log n)^{d+1/2}\sqrt{\log\log n}$ [6].

Using a complexity result from [4], a simple consequence of our bounds is that the complexity of off-line orthogonal range searching is at least $\Omega(nd + n\log\log n)$ in any dimension $d = O(\log n)$. Given n weighted points and n boxes in $\mathbf{R}^d$, off-line orthogonal range searching is the task of computing the added weight of all the points in each box.

2. The Discrepancy of Boxes

Throughout this paper we assume that $d > 1$; the case $d = 1$ is trivial and can be ignored. Also, the term box always refers to an axis-parallel box. We state our main result and an immediate corollary.

Theorem 2.1. *For any n large enough and any dimension $d = O(\log n)$, there exists a set system of n points and n boxes in $\mathbf{R}^d$, whose red–blue discrepancy is $2^{\Omega(d)}$.*

Corollary 2.2. *For any n large enough and any dimension $d = \Omega(\log n)$, there exists a set system of n points and n boxes in $\mathbf{R}^d$, whose red–blue discrepancy is $n^{\Omega(1)}$.*

Recall [7] that the red–blue discrepancy is always in $O(\sqrt{n})$, and so the remaining open problem is to determine the precise constant behind the $\Omega()$ notation. As we shall see, the theorem is in fact stronger than stated, since it holds for points and boxes in the Hamming cube $\{0, 1\}^d$. Theorem 2.1 follows easily from the lemma below.

Lemma 2.3. *For any n large enough, there exists a set system of n points and n boxes in $\mathbf{R}^d$, where $d = \Theta(\log n)$, whose red–blue discrepancy is $\Omega(n^{0.0477})$.*

The theorem is trivially implied by the lemma for $d \geq c\log n$, for some constant $c > 0$. Suppose now that $d < c\log n$. Set $n_0 = 2^{d/c}$ so that we can apply the lemma with respect to n_0 and d. We can pad the set system to be n-by-n by adding $n - n_0$ artificial points and boxes with no enclosure relationships. The lower bound of $\Omega(n_0^{0.0477})$ is also $\Omega(2^{\Omega(d)})$. We can assume that d is large enough since a logarithmic lower bound is already known [3] for $d = 2$. Thus the lower bound can be expressed more simply as $2^{\Omega(d)}$.

Proof of Lemma 2.3. The hereditary discrepancy [5] of the set system defined by A, which is denoted by $\mathrm{herdisc}(A)$, is defined as the maximum discrepancy of any submatrix of A. By a simple padding argument it is clear that a lower bound on the hereditary discrepancy implies a similar bound on the red–blue discrepancy. We proved in [4] that if $M = AA^T$, then

$$\mathrm{herdisc}(A) \geq \tfrac{1}{4}\, c_0^{n\,\mathrm{tr}\,M^2/\mathrm{tr}^2\,M}\sqrt{\frac{\mathrm{tr}\,M}{n}}, \tag{1}$$

for some constant $0 < c_0 < 1$. So, to achieve a red–blue discrepancy lower bound of $n^{\Omega(1)}$, it suffices to exhibit a probabilistic construction of m points and n boxes in $\mathbf{R}^{O(\log n)}$ with the following characteristics: for some constant $c \approx 1.0955$,

(i) $m = \Theta(n)$ and $\operatorname{tr} M = \Theta(n^c)$ with probability at least $1/2$;
(ii) $\mathbf{E} \operatorname{tr} M^2 = O(n^{2c-1})$.

Indeed, after appropriate padding and rescaling, we immediately derive from these conditions the existence of a suitable n-by-n set system that, in view of (1), implies Lemma 2.3. □

As we said earlier, both the point set and the boxes live in the Hamming cube $\{0, 1\}^d$. For the proof, we define a few parameters whose meaning we explain below (all logarithms are to base two):

$$\begin{cases} w = \dfrac{1 - 2p + p^9}{1 - 2p - (1 + 2p)p^2 \log e}, & \text{where} \quad p = 0.153, \\ c = 2 - (1 - p)w, \\ G = n^{c-1}. \end{cases}$$

We assume that both $d \stackrel{\text{def}}{=} w \log n$ and pd are integral: this is of no consequence as rounding off to the nearest integer produces lower-order errors of no significance to our results. The m points are chosen by picking each element of the Hamming cube $\{0, 1\}^d$ independently with probability n^{1-w}. (Note that $w \approx 1.067867$, so $n^{1-w} < 1$.) The expected number of points is n. In fact, by Chebyshev's inequality, we have

Lemma 2.4. *With probability* $> 1/2$, *the number* m *of points is* $\Theta(n)$.

A box is specified by a word of length d, over the alphabet $\{0, 1, *\}$, containing exactly pd stars. For example, in dimension 5, the word $0 * 1 * *$ denotes the three-dimensional box $x_1 = 0$, $x_3 = 1$. We construct the n boxes by specifying G groups of parallel boxes. Each group is defined by selecting the location of the stars first (the *star pattern*), and then taking all the corresponding boxes. To specify the star pattern, we choose pd coordinates uniformly at random (without replacement) and make them stars. In our previous example, the group of parallel boxes consists of $0 * 0 * *$, $0 * 1 * *$, $1 * 0 * *$, and $1 * 1 * *$. The number of boxes is precisely $2^{(1-p)d} G = n$.

Each point in the set system belongs to exactly one box in each of the G groups, so that $\operatorname{tr} M = mG$. By Lemma 2.4, we have the following result, which implies that condition (i) is satisfied with probability $> 1/2$.

Lemma 2.5. *With probability* $> 1/2$, *the trace of* M *is* $\Theta(n^c)$.

We now turn to the trace of M^2 and bound it from above as a function of n. By definition,

$$\operatorname{tr} M^2 = O(\sigma_{1,1} + \sigma_{1,2} + \sigma_{2,1} + \sigma_{2,2}),$$

where $\sigma_{i,j}$ counts the number of pairs (I, J) such that $I \supseteq J$, where I is the intersection of i distinct boxes and J is a set of j distinct points. Next, we derive upper bounds on

all these numbers, beginning with

$$\mathbf{E}\sigma_{1,1} = \mathbf{E}\,\mathrm{tr}\,M = n^c. \tag{2}$$

The next derivations are straightforward:

$$\mathbf{E}\sigma_{1,2} = O(n^{2c-1}) \quad \text{and} \quad \mathbf{E}\sigma_{2,1} = O(n^{2c-1}). \tag{3}$$

Why? Any one of the 2^{pd} Hamming cube vertices lying in a given box belongs to the set system with probability n^{1-w}. There are n boxes, so

$$\mathbf{E}\sigma_{1,2} = O\left(n(2^{pd}n^{1-w})^2\right) = O(n^{3-2(1-p)w}),$$

which takes care of $\sigma_{1,2}$. Regarding $\sigma_{2,1}$, note that boxes within the same group are disjoint, so only pairs in distinct groups can contribute to $\sigma_{2,1}$. Fix two such groups. Any one of the 2^d points of the Hamming cube belongs to exactly one pair of boxes. Since such a point is picked with probability n^{1-w}, we have $\mathbf{E}\sigma_{2,1} = O(G^2 2^d n^{1-w}) = O(n^{2c-1})$ and, hence, (3).

Finally, we turn to the expectation of $\sigma_{2,2}$:

$$\mathbf{E}\sigma_{2,2} = O(n^{2c-w+((1+2p)/(1-2p))p^2 w \log e} \log n). \tag{4}$$

Again, fix two groups of parallel boxes, and let x be the number of stars common to both star patterns. As we just saw, any point of the Hamming cube belongs to exactly one pair of boxes, and this point can be paired with exactly $2^x - 1$ other points. Each point being picked with probability n^{1-w}, it follows that

$$\sigma_{2,2} = O(G^2 2^{d+x} n^{2-2w})$$

and, hence,

$$\mathbf{E}\sigma_{2,2} = O(n^{2c-w})\mathbf{E}2^x.$$

To bound the expectation of 2^x is easy. Using the notation

$$N_k \stackrel{\text{def}}{=} N(N-1)\cdots(N-k+1)$$

and the inequality $k! > (k/e)^k$, we find that

$$\begin{aligned}
\mathbf{E}\,2^x &= \sum_{k=0}^{pd} 2^k \binom{pd}{k}\binom{d-pd}{pd-k} \Big/ \binom{d}{pd} = \sum_{k=0}^{pd} \frac{2^k (pd)_k (d-pd)_{pd-k}}{k!(pd-k)!} \Big/ \binom{d}{pd} \\
&\le \sum_{k=0}^{pd} \frac{(2pd)^k (d-pd)_{pd}(pd)_k}{k!(d-2pd)^k (pd)!} \Big/ \binom{d}{pd} \le \sum_{k=0}^{pd} \frac{(2ep^2d^2)^k (d-pd)_{pd}}{(kd)^k (1-2p)^k (pd)!} \Big/ \binom{d}{pd} \\
&\le \sum_{k=0}^{pd} (1-p)^{pd} \left(\frac{2ep^2 d}{(1-2p)k}\right)^k.
\end{aligned}$$

The function $(A/x)^x$ is maximized at $x = A/e$, therefore

$$\mathbf{E}2^x = O(n^{(\log e)p^2 w(1+2p)/(1-2p)} \log n),$$

hence (4). In view of (2)–(4),

$$\begin{aligned}\mathbf{E}\,\mathrm{tr}\,M^2 &= O\left(n^c + n^{2c-1} + n^{2c-w+((1+2p)/(1-2p))p^2 w\log e}\log n\right)\\ &= O\left(n^{2c-1} + n^{2c-1-p^9/(1-2p)}\log n\right) = O(n^{2c-1}),\end{aligned}$$

which establishes condition (ii), and hence Lemma 2.3 and Theorem 2.1. □

3. The Complexity of Orthogonal Range Searching

The construction of points and boxes can be used to prove a lower bound on the complexity of off-line orthogonal range searching. This is the problem of adding up weights in n boxes, given n weighted points. Specifically, we are given n axis-parallel boxes and n points in $\mathbf{R}^d$, fixed once and for all. The input to the problem is an assignment of reals (the weights) to the points, and the output is the sum of the weights of the points within each box. Equivalently, the problem is to compute Ax given x. From [4] we know that the size of any linear circuit with bounded coefficients for computing $x \mapsto Ax$ is

$$\Omega_\varepsilon\left(n\log\left(\mathrm{tr}\,M/n - \varepsilon\sqrt{\mathrm{tr}\,M^2/n}\right)\right),$$

for any positive constant ε. Setting ε small enough gives a lower bound of $\Omega(n\log n)$ for orthogonal range searching in dimension $\Omega(\log n)$. This is to be contrasted with the current $O(n\log\log n)$ lower bound for orthogonal range searching in fixed dimension [2].

The case of dimension $d = O(\log n)$ is handled as we did before. We create a problem of size $n_0 = 2^{\Theta(d)}$ in dimension d, with n_0 sufficiently small with respect to d that we can apply the previous result. The problem requires a circuit of size $\Omega(n_0\log n_0)$. We make about n/n_0 copies of it (with different weight assignments, of course) to boost the complexity to $\Omega(n\log n_0)$, which is $\Omega(nd)$. Since, for $d = 2$, the complexity is at least $\Omega(n\log\log n)$, we can safely conclude that the circuit complexity of orthogonal range searching in dimension $d = O(\log n)$ is $\Omega(nd + n\log\log n)$, as claimed.

Theorem 3.1. *Off-line orthogonal range searching in* $\mathbf{R}^d$ *has complexity* $\Omega(nd + n\log\log n)$ *for any dimension* $d = O(\log n)$.

Acknowledgment

We thank the referees for several useful comments.

References

[1] Beck, J., and Chen, W. W. L., *Irregularities of Distribution*, Cambridge Tracts in Mathematics, 89, Cambridge University Press, Cambridge, 1987.
[2] Chazelle, B., Lower bounds for off-line range searching, *Discrete Comput. Geom.*, **17** (1997), 53–65.
[3] Chazelle, B., *The Discrepancy Method*: *Randomness and Complexity*, Cambridge University Press, Cambridge, 2000.

[4] Chazelle, B., and Lvov, A., A trace bound for the hereditary discrepancy, *Proc*. 16*th Annual ACM Symp*. *Comput. Geom*. (2000).
[5] Lovász, L., Spencer, J., and Vesztergombi, K., Discrepancy of set systems and matrices, *European J. Combin*., **7** (1986), 151–160.
[6] Matoušek, J., *Geometric Discrepancy*: *An Illustrated Guide*, Algorithms and Combinatorics, 18, Springer-Verlag, New York, 1999.
[7] Spencer, J., Six standard deviations suffice, *Trans*. *Amer. Math. Soc*., **289** (1985), 679–706.

Received June 30, 2000, *and in revised form November* 9, 2000. *Online publication April* 6, 2001.