Expected Case for Projecting Points

Sergio Cabello and Matt DeVos
Institute for Mathematics, Physics and Mechanics, Ljubljana, Slovenia
E-mail: cabello@imfm.uni-lj.si

Bojan Mohar
Faculty of Mathematics and Physics, Ljubljana, Slovenia
E-mail: bojan.mohar@uni-lj.si

Keywords: randomized algorithm, unit distance, closest pair

Received: June 14, 2005

Consider a set of n points in the plane with the property that any pair of points is at least at distance one. We study the expected concentration of the point set after projecting it onto a random graduated line. There is a lower bound of $\Omega(\sqrt{n \log n})$ given by Matoušek in [4], and we provide an upper bound of $O(n^{2/3})$.

Povzetek: Analizirana je gostota točk v ravnini z razdaljo najmanj ena.

1 Introduction

Let P be a set of n points in the plane. For a line $L \subset \mathbb{R}^2$, we can project the points P orthogonally onto L, which we denote by $\pi_L(P)$. Imagine that the line L is a graduated line, that is, a line decomposed into line segments (*cells*) of length one. For a cell $c \subset L$, let $Pop(P, c)$ be the population of the cell c after the projection, that is $Pop(P, c) = |\{p \in P \,|\, \pi_L(p) \in c\}|$. For a graduated line L, we say that its concentration $Conc(P, L)$ is the number of points that its most populated cell gets; that is,

$$Conc(P, L) = \max_{c \text{ a cell of } L} \{Pop(P, c)\}.$$

In a recent paper, Díaz et al. [3] consider the algorithmic problem of computing a graduated line that minimizes the concentration, that is, they are interested in $Conc(P) = \min_L Conc(P, L)$. However, an asymptotically equivalent problem was considered by Kučera et al. [4] when studying a map labelling problem.

Here we are interested in the expected concentration that a point set has when projecting onto a random graduated line. Let $L(\alpha)$ be a graduated line through the origin with angle α with respect to the x-axis, and such that the origin is the boundary of a cell. We are interested in the expected concentration $EConc(P)$ over all lines $L(\alpha)$

$$EConc(P) = \mathbb{E}_\alpha \left[Conc(P, L(\alpha))\right],$$

where α is chosen uniformly at random. Let us observe that, for an asymptotic bound on $EConc(P)$, it is equivalent to consider that the lines $L(\alpha)$ pass through some other point of $\mathbb{R}^2$ instead of the origin.

If the point set P is arbitrarily dense, then it may be that $Conc(P, L) \geq n/2$ for any line L, and so $EConc(P) = \Omega(n)$. However, the problem becomes non-trivial if we put restrictions to the density of the point set.

Definition 1. *A point set $P \subset \mathbb{R}^2$ is* 1-separated *if its closest pair is at least at distance 1.*

Our objective[1],[2] is to bound the value $EConc(P)$ for any 1-separated point set. Kučera et al. [4] have shown that $Conc(P) = O(\sqrt{n \log n})$ for any 1-separated point set P. More interestingly, they use Besicovitch's sets [1] for constructing 1-separated point sets P having $Conc(P) = \Omega(\sqrt{n \log n})$, which implies $EConc(P) = \Omega(\sqrt{n \log n})$.

We will show that for any 1-separated point set P we have $EConc(P) = O(n^{2/3})$. Therefore, it remains open to find tight bounds for $EConc(P)$.

The rationale behind considering projections onto random lines is the efficiency of randomized algorithms whose running time depends on the expected concentration. As an example, consider a set of disjoint unit disks and any sweep-line algorithm [2, Chapter 2] whose running time depends on the maximum number of disks that are intersected by the sweep line. Choosing the direction in which the line sweeps affects the running time, but computing the best direction, or an approximation, is expensive: Kučera et al. [4] claim that it can be done in polynomial time, and Díaz et al. [3] give a constant-factor approximation algorithm with $O(nt \log nt)$ running time, where t is the diameter of P. By choosing a random projection we avoid having to compute a good direction for projecting, and we get a randomized algorithm. The results in this paper become helpful for analyzing the expected running time of such randomized algorithms.

The rest of the paper is organized as follows. In Section 2 we introduce some relevant random variables and give some basic facts. In Sections 3 and 4 we bound

[1]Supported by the European Community Sixth Framework Programme under a Marie Curie Intra-European Fellowship.

[2]Supported in part by the Ministry of Education, Science and Sport of Slovenia, Research Program P1–0507–0101.

$EConc(P)$ using the first and second moments, respectively.

2 Preliminaries

Let $P = \{p_1, \ldots, p_n\}$ be a 1-separated point set, and let $d_{i,j} = d(p_i, p_j)$. We use the notation $[n] = \{1, \ldots, n\}$. Without loss of generality, we can restrict ourselves to graduated lines passing through the origin. Let $L(\alpha)$ be the line passing through the origin that has angle α with the x-axis, and let $p^*(\alpha)$ be the orthogonal projection of a point p onto $L(\alpha)$. Consider the following random variables for the angle α

$$X_{i,j}(\alpha) = \begin{cases} 1 & \text{if } d\big(p_i^*(\alpha), p_j^*(\alpha)\big) \le 1, \\ 0 & \text{otherwise;} \end{cases}$$

$$X_i(\alpha) = \sum_{j=1}^{n} X_{i,j}(\alpha);$$

$$X_{max}(\alpha) = \max\{X_1(\alpha), \ldots, X_n(\alpha)\};$$

$$X(\alpha) = \sum_{i=1}^{n} X_i(\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n} X_{i,j}(\alpha),$$

where α is chosen uniformly at random from the values $[0, \pi)$. In words: $X_{i,j}$ is the indicator variable for the event that $p_i^*(\alpha)$ and $p_j^*(\alpha)$ are at distance at most one in the projection; X_i is the number of points (including p_i itself) whose projection is at distance at most one from $p_i^*(\alpha)$; X_{max} is the maximum among $X_1, \ldots, X_n$; and X counts twice the number of pairs of points at distance at most one in the projection. It is clear that $\mathbb{P}[X_{i,i} = 1] = 1$ for any $i \in [n]$. Otherwise we have the following result.

Lemma 1. *If $i \neq j$, then*

$$\mathbb{P}[X_{i,j} = 1] = \frac{2 \arcsin 1/d_{i,j}}{\pi}.$$

Proof. Assume without loss of generality that p_i is placed at the origin and p_j is vertically above it, on the y-axis. See Figure 1. We may also assume that the line $L(\alpha)$ passes through p_i. Because $d_{i,j} \ge 1$, there are values α such that $X_{i,j}(\alpha) \neq 1$. The angles that make $X_{i,j}(\alpha) = 1$ are indicated in the figure. In particular, if β is the angle indicated in the figure, and we choose α uniformly at random, then $\mathbb{P}[X_{i,j} = 1] = \frac{2\beta}{\pi}$. The angle β is such that $\sin\beta = \frac{1}{d_{i,j}}$, and so $\beta = \arcsin\frac{1}{d_{i,j}}$. We conclude that $\mathbb{P}[X_{i,j} = 1] = \frac{2\beta}{\pi} = \frac{2\arcsin 1/d_{i,j}}{\pi}$. □

The first observation, which is already used for the approximation algorithms described by Díaz et al. [3], is that, asymptotically, we do not need to care for the graduation, but only for the orientation of the line. In particular, the random variables X_i contain all the information that we need asymptotically.

Lemma 2. *We have*

$$\frac{EConc(P))}{2} \le \mathbb{E}\left[X_{max}(\alpha)\right] \le 2\, EConc(P).$$

3 Using the first moment

Using that the closest pair of P is at least one apart, we get the following result.

Lemma 3. *For every $i \in [n]$, we have*

$$\sum_{j \in [n]\setminus\{i\}} \frac{1}{d_{i,j}} = O(\sqrt{n}).$$

Proof. Without loss of generality, assume that $i = n$. Let n_d be the number of points in P whose distance from p_n is in the interval $[d, d+1)$. We have

$$\sum_{j\in[n-1]} \frac{1}{d_{i,j}} = \sum_{d=1}^{\infty} \left(\sum_{d_{i,j}\in[d,d+1)} \frac{1}{d_{i,j}} \right) \tag{1}$$

$$\le \sum_{d=1}^{\infty} \left(\sum_{d_{i,j}\in[d,d+1)} \frac{1}{d} \right) \tag{2}$$

$$= \sum_{d=1}^{\infty} \frac{n_d}{d}. \tag{3}$$

Observe that if we have two sequences $(a_i)_{i\in\mathbb{N}}$ and $(b_i)_{i\in\mathbb{N}}$ of nonnegative numbers such that $\sum_{i=1}^{j} a_i \le \sum_{i=1}^{j} b_i$ for all $j \in \mathbb{N}$, then $\sum_{i=1}^{\infty} \frac{a_i}{i} \le \sum_{i=1}^{\infty} \frac{b_i}{i}$. That is, the sum is maximized when the values concentrate on the smallest possible indexes. Let N_d be the maximum number of 1-separated points that you can have in an annulus of inner radius d and exterior radius $d+1$, and let D be the smallest value such that $n < \sum_{d=1}^{D} N_d$. We have $n = \sum n_d$ and $\sum_{i=1}^{j} n_i \le \sum_{i=1}^{j} N_i$ for all $j \in [D]$, and from (1) we conclude

$$\sum_{j\in[n-1]} \frac{1}{d_{i,j}} \le \sum_{d=1}^{\infty} \frac{n_d}{d} \le \sum_{d=1}^{D} \frac{N_d}{d}. \tag{4}$$

We need to estimate the values N_d. For the lower bound, placing points at distance one in the circle of radius d, we get $N_d = \Omega(d)$. For the upper bound, we can use a packing argument to show that any 1-separated point set inside the annulus has $O(d)$ points. Indeed, if we place a disk of radius $1/2$ centered in each point of a 1-separated point set inside the annulus, they must have disjoint interiors and cover an area of $\Theta(N_d)$. Moreover, all these disks are contained in an annulus of inner radius $d-1$ and exterior radius $d+2$, which has an area of $\Theta(d)$. We conclude that $N_d = \Theta(d)$, and therefore $D = O(\sqrt{n})$. Using (4) we get

$$\sum_{j\in[n-1]} \frac{1}{d_{i,j}} \le \sum_{d=1}^{D} \frac{N_d}{d} \le \sum_{d=1}^{O(\sqrt{n})} \frac{O(d)}{d} = O(\sqrt{n}).$$

□

Lemma 4. *For every $i \in [n]$ we have $\mathbb{E}[X_i] = O(\sqrt{n})$.*

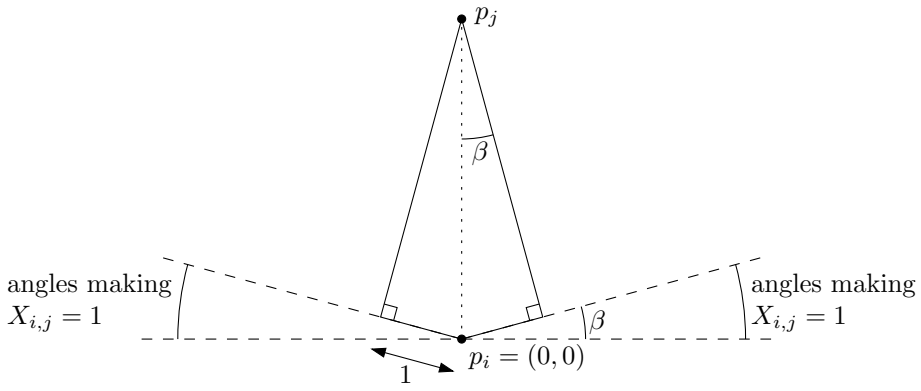

Figure 1: For Lemma 4. We consider what random lines $L(\alpha)$ through p_i that give $X_{ij} = 1$

Proof. Because $X_i = \sum_{j=1}^{n} X_{i,j}$ and the linearity of the expectation, we have

$$\begin{aligned}\mathbb{E}[X_i] &= \sum_{j=1}^{n} \mathbb{E}[X_{i,j}] = \sum_{j=1}^{n} \mathbb{P}[X_{i,j} = 1] \\ &= 1 + \sum_{j \in [n]\setminus\{i\}} \mathbb{P}[X_{i,j} = 1] \\ &= 1 + \sum_{j \in [n]\setminus\{i\}} \frac{2\arcsin(1/d_{i,j})}{\pi}.\end{aligned}$$

Observe that the function $\arcsin(x)$ is convex for $x \in [0,1]$, and therefore we have $\arcsin(x) \leq (\pi/2)x$ for all $x \in [0,1]$. We then have

$$\begin{aligned}\mathbb{E}[X_i] &= 1 + \sum_{j \in [n]\setminus\{i\}} \frac{2\arcsin(1/d_{i,j})}{\pi} \\ &\leq 1 + \sum_{j \in [n]\setminus\{i\}} \frac{1}{d_{i,j}},\end{aligned}$$

and using Lemma 3 we conclude that $\mathbb{E}[X_i] = O(\sqrt{n})$. □

Using the first moment method, we can show that for any 1-separated point set P it holds that $EConc(P) = O(n^{3/4})$. For this, consider a 1-separated point set P and its associated random variable X. We have $X = \sum X_i$, and because of Lemma 4 we conclude $\mathbb{E}[X] = O(n\sqrt{n})$.

We claim that, for any value $t > 0$, if we have $X_{max}(\alpha) \geq t$, then $X(\alpha) \geq t^2/4$. Intuitively, if some $X_i = t$, then there are $\Theta(t^2)$ pairs of points at distance at most one from each other, and so contributing to X. The formal proof of the claim is as follows. Let i be an index such that $X_i(\alpha) \geq t$. Then, either to the right or to the left of $p_i^*(\alpha)$, the projection of p_i onto $L(\alpha)$, there are at least $t/2$ points $p_j^*(\alpha)$ at distance at most one from $p_i^*(\alpha)$. Assume that those points are to the left and let $\tilde{P} \subset P$ be the set of those points. We have $|\tilde{P}| \geq t/2$. For any $p_j, p_{j'} \in \tilde{P}$ we have $X_{j,j'}(\alpha) = 1$, and therefore we have $X_j(\alpha) \geq t/2$ for all $p_j \in \tilde{P}$. We conclude that

$$X(\alpha) \geq \sum_{p_j \in \tilde{P}} X_j(\alpha) \geq \sum_{p_j \in \tilde{P}} t/2 \geq t/2 \cdot |\tilde{P}| \geq t^2/4,$$

and the claim is proved.

We have shown that for any value $t > 0$ we have

$$[X_{max} \geq t] \subseteq [X \geq t^2/4],$$

and using Markov's inequality we conclude

$$\mathbb{P}[X_{max} \geq t] \leq \mathbb{P}[X \geq t^2/4] \leq \frac{4\mathbb{E}[X]}{t^2} \leq \frac{O(n\sqrt{n})}{t^2}.$$

Let $r = \lfloor n^{3/4} \rfloor$. Since X_{max} only takes natural numbers, we have

$$\begin{aligned}\mathbb{E}[X_{max}] &= \sum_{t=1}^{n} \mathbb{P}[X_{max} \geq t] \\ &= \sum_{t=1}^{r} \mathbb{P}[X_{max} \geq t] + \sum_{t=r+1}^{n} \mathbb{P}[X_{max} \geq t] \\ &\leq \sum_{t=1}^{r} 1 + \sum_{t=r+1}^{n} \frac{O(n\sqrt{n})}{t^2} \\ &\leq r + O(n\sqrt{n}) \int_r^n \frac{1}{t^2} dt \\ &\leq n^{3/4} + O(n\sqrt{n}) \left(\frac{1}{r} - \frac{1}{n}\right) \\ &= O(n^{3/4}).\end{aligned}$$

Using Lemma 2 it follows that $EConc(P) = O(n^{3/4})$. However, observe that this bound will be improved in next section.

We would like to point out that the random variables X_i do not have a strong concentration around their expectation. Therefore, we cannot use many of the results based on concentration of the measure that would reduce the bound on $EConc(P)$. To see this, consider the example in Figure 2. The point p_i is the center of a disc of

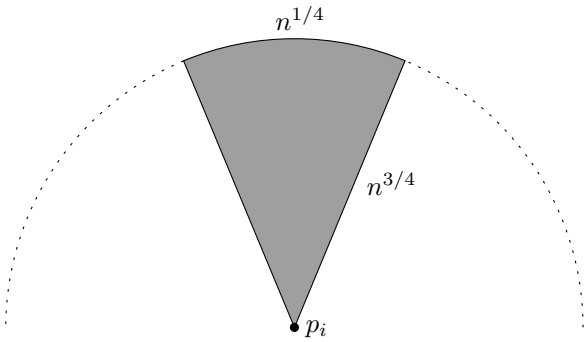

Figure 2: Example showing that X_i is not concentrated around its expectation.

radius $n^{3/4}$, and we consider a circular sector with arc-length $n^{1/4}$. This region is grey in the picture. Imagine that we place a densest 1-separated point set P inside the grey region. Asymptotically, since the region has area $\Theta(n)$, such a point set P has $\Theta(n)$ points. Consider the lines $L(\alpha + \pi/2)$ passing through p_i. If α is chosen uniformly at random, the line $L(\alpha)$ intersects the grey region with probability $n^{1/4}/(2\pi n^{3/4}) = \Theta(1/\sqrt{n})$, and in that case $X_i(\alpha + \pi/2) = \Theta(n^{3/4})$. We conclude that $\mathbb{E}[X_i] = \Theta(n^{1/4})$, but $\mathbb{P}[X_i = \Omega(n^{4/3})] = \Theta(1/\sqrt{n})$, and so X_i does not concentrate around its expectation.

4 Second moments

Lemma 5. *For every $i \in [n]$ we have $\mathbb{E}[X_i^2] = O(n)$.*

Proof. Assume without loss of generality that $d_{i,j} \geq d_{i,k}$ whenever $j > k$; that is, the points are indexed according to their distance from p_i. Like above, we assume that the line $L(\alpha)$ passes through p_i. We have

$$\begin{aligned}\mathbb{E}[X_i^2] &= \mathbb{E}\left[\sum_{j,k\in[n]} X_{i,j}X_{i,k}\right] \\ &\leq \mathbb{E}\left[2\sum_{j}\sum_{k\leq j} X_{i,j}X_{i,k}\right] \\ &= 2\sum_{j}\mathbb{E}\left[X_{i,j}\sum_{k\leq j} X_{i,k}\right]\end{aligned}$$

We claim that $\mathbb{E}\left[X_{i,j}\sum_{k\leq j} X_{i,k}\right] = O(1)$, and so the result follows.

To prove the claim, observe that if $X_{i,j}(\alpha) = 1$, then all the points p_k that have $X_{i,k}(\alpha) = 1$ need to be in the strip (or slab) of width two having $L(\alpha + \pi/2)$ as axis; see Figure 3, where this strip is in grey. Because of a packing argument, in this strip there are $O(d_{i,j})$ points p_k that satisfy $d_{i,j} \geq d_{i,k}$. Therefore, by the way we indexed the points, we conclude that, if $X_{i,j}(\alpha) = 1$, then $\left(\sum_{k\leq j} X_{i,k}\right)(\alpha) = O(d_{i,j})$. In any case, we always have $\left(X_{i,j}\sum_{k\leq j} X_{i,k}\right)(\alpha) = O(d_{i,j})$. Therefore

$$\begin{aligned}\mathbb{E}\left[\, X_{i,j}\sum_{k\leq j} X_{i,k}\,\right] &= \sum_{t=1}^{n} t\cdot\mathbb{P}\left[X_{i,j}\sum_{k\leq j} X_{i,k} = t\right] \\ &\leq \sum_{t=1}^{n} O(d_{i,j})\cdot\mathbb{P}\left[X_{i,j}\sum_{k\leq j} X_{i,k} = t\right] \\ &= O(d_{i,j})\sum_{t=1}^{n}\mathbb{P}\left[X_{i,j}\sum_{k\leq j} X_{i,k} = t\right] \\ &\leq O(d_{i,j})\cdot\mathbb{P}\left[X_{i,j} = 1\right] \\ &= O(d_{i,j})\frac{2\arcsin 1/d_{i,j}}{\pi} = O(1).\end{aligned}$$

This finishes the proof of the claim and of the lemma. □

Theorem 1. *For any 1-separated point set P we have $EConc(P) = O(n^{2/3})$.*

Proof. Let P be a 1-separated point set and consider the random variable $T(\alpha) = \left(\sum_i X_i^2\right)(\alpha)$. By Lemma 5 we have $\mathbb{E}[T] = \sum_i \mathbb{E}[X_i^2] = O(n^2)$. The rest of the proof resembles the argument in the previous section.

We claim that, for any value $t > 0$, if we have $X_{max}(\alpha) \geq t$, then $T(\alpha) \geq t^3/8$. The proof is as follows. Let i be an index such that $X_i(\alpha) \geq t$. Then, either to the right or to the left of $p_i^*(\alpha)$, the projection of p_i onto $L(\alpha)$, there are at least $t/2$ points $p_j^*(\alpha)$ at distance at most one from $p_i^*(\alpha)$. Assume that those points are to the left and let $\tilde{P} \subseteq P$ be the set of those points. We have $|\tilde{P}| \geq t/2$. For any $p_j, p_{j'} \in \tilde{P}$ we have $X_{j,j'}(\alpha) = 1$. Therefore for all $p_j \in \tilde{P}$ we have $X_j(\alpha) \geq t/2$, and $X_j^2(\alpha) \geq t^2/4$. We conclude that

$$\begin{aligned}T(\alpha) &\geq \sum_{p_j\in\tilde{P}} X_j^2(\alpha) \\ &\geq \sum_{p_j\in\tilde{P}} t^2/4 \geq t^2/4\cdot|\tilde{P}| \\ &\geq t^3/8,\end{aligned}$$

and the claim is proved.

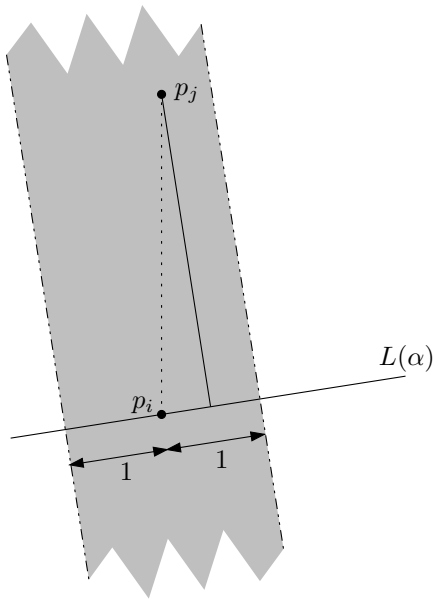

Figure 3: For the proof of Lemma 5. For any angle α we have $\left(X_{i,j} \sum_{k \le j} X_{i,k}\right)(\alpha) = O(d_{i,j})$.

We have shown that for any value $t > 0$ we have

$$[X_{max} \ge t] \subseteq [T \ge t^3/8],$$

and using Markov's inequality we conclude

$$\mathbb{P}[X_{max} \ge t] \le \mathbb{P}[T \ge t^3/8] \le \frac{8\mathbb{E}[T]}{t^3} \le \frac{O(n^2)}{t^3}.$$

Let $r = \lfloor n^{2/3} \rfloor$. Since X_{max} only takes natural numbers, we have

$$\begin{aligned}
\mathbb{E}[X_{max}] &= \sum_{t=1}^{n} \mathbb{P}[X_{max} \ge t] \\
&= \sum_{t=1}^{r} \mathbb{P}[X_{max} \ge t] + \sum_{t=r+1}^{n} \mathbb{P}[X_{max} \ge t] \\
&\le \sum_{t=1}^{r} 1 + \sum_{t=r+1}^{n} \frac{O(n^2)}{t^3} \\
&\le r + O(n^2) \int_r^n \frac{1}{t^3} dt \\
&\le n^{2/3} + O(n^2) \left(\frac{2}{r^2} - \frac{2}{n^2}\right) \\
&= O(n^{2/3}).
\end{aligned}$$

Using Lemma 2 it follows that $EConc(P) = O(n^{2/3})$. □

Trying to use the same ideas with higher moments of X_i does not help. Consider for example the 1-separated point set P consisting of all n points in a horizontal row of length n, and let p_1 be the leftmost point. We have $\mathbb{E}[X_1^3] = \Theta(n^2)$, and in general $\mathbb{E}[X_1^p] = \Theta(n^{p-1})$ for all naturals $p > 2$. From this we can only conclude weaker results of the type $EConc(P) = O(n^{p/(p+1)})$.

Conclusions

We have studied the expected concentration of projecting 1-separated point sets onto random lines, a parameter that is relevant for sweep-line algorithms when the direction for sweeping is chosen at random. We have shown that, if P consists of n points, the expected concentration $EConc(P)$ is $O(n^{2/3})$, while the best known lower bound is $\Omega(\sqrt{n \log n})$. Therefore, it remains to close this gap.

Acknowledgements

The authors are grateful to Jiří Matoušek for the key reference [4]. Sergio is also grateful to Christian Knauer for early discussions.

References

[1] A. S. Besicovitch. The Kakeya problem. *The American Mathematical Monthly*, 70:697–706, 1963.

[2] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, Germany, 2nd edition, 2000.

[3] J.M. Díaz and F. Hurtado and M. López and J.A. Sellarès. Optimal point set projections onto regular grids. In T. Ibaraki et al., editor, *14th Inter. Symp. on Algorithms and Computation*, volume 2906 of *LNCS*, pages 270–279. Springer Verlag, 2003.

[4] L. Kučera, K. Mehlhorn, B. Preis, and E. Schwarzenecker. Exact algorithms for a geometric packing problem. In *Proc. 10th Sympos. Theoret. Aspects Comput. Sci.*, volume 665 of *Lecture Notes Comput. Sci.*, pages 317–322. Springer-Verlag, 1993.