The Expected Number of Maximal Points of the Convolution of Two 2-D Distributions

Josep Diaz
Department of CS, UPC, Barcelona, Spain
diaz@cs.upc.edu

Mordecai Golin
CSE Department, Hong Kong UST
golin@cse.ust.hk

Abstract

The *Maximal* points in a set S are those that are not *dominated* by any other point in S. Such points arise in multiple application settings and are called by a variety of different names, e.g., maxima, Pareto optimums, skylines. Their ubiquity has inspired a large literature on the *expected* number of maxima in a set S of n points chosen IID from some distribution. Most such results assume that the underlying distribution is uniform over some spatial region and strongly use this uniformity in their analysis.

This research was initially motivated by the question of how this expected number changes if the input distribution is perturbed by random noise. More specifically, let $\mathbf{B}_p$ denote the uniform distribution from the 2-dimensional unit ball in the metric L_p. Let $\delta\mathbf{B}_q$ denote the 2-dimensional L_q-ball, of radius δ and $\mathbf{B}_p + \delta\mathbf{B}_q$ be the convolution of the two distributions, i.e., a point $v \in \mathbf{B}_p$ is reported with an error chosen from $\delta\mathbf{B}_q$. The question is how the expected number of maxima change as a function of δ. Although the original motivation is for small δ, the problem is well defined for any δ and our analysis treats the general case.

More specifically, we study, as a function of n, δ, the expected number of maximal points when the n points in S are chosen IID from distributions of the type $\mathbf{B}_p + \delta\mathbf{B}_q$ where $p, q \in \{1, 2, \infty\}$ for $\delta > 0$ and also of the type $\mathbf{B}_\infty + \delta\mathbf{B}_q$ where $q \in [1, \infty)$ for $\delta > 0$.

For fixed p, q we show that this function changes "smoothly" as a function of δ but that this smooth behavior sometimes transitions unexpectedly between different growth behaviors.

2012 ACM Subject Classification Theory of computation → Randomness, geometry and discrete structures

Keywords and phrases maximal points, probabilistic geometry, perturbations, Minkowski sum

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2019.35

Category RANDOM

Related Version https://arxiv.org/abs/1807.06845

Funding *Josep Diaz*: TIN2017-86727-C2-1R

1 Introduction

Let S be a set of 2-dimensional points. The "largest" points in S are the *maximal points* of S and are a well-studied object. More formally

▶ **Definition 1.** *For $u \in \Re^2$ let $u.x$ ($u.y$) denote the x (y) coordinate of u. For $u, v \in \Re^2$, u is* dominated *by v if $u \neq v$, $u.x \leq v.x$ and $u.y \leq v.y$. If $S \subset \Re^2$ then*

$$\mathrm{MAX}(S) = \{u \in S : u \text{ is not dominated by any point in } S \setminus \{u\}\}.$$

$\mathrm{MAX}(S)$ *are the* maximal points *of S. See Fig. 1.*

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019).
Editors: Dimitris Achlioptas and László A. Végh; Article No. 35; pp. 35:1–35:14
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

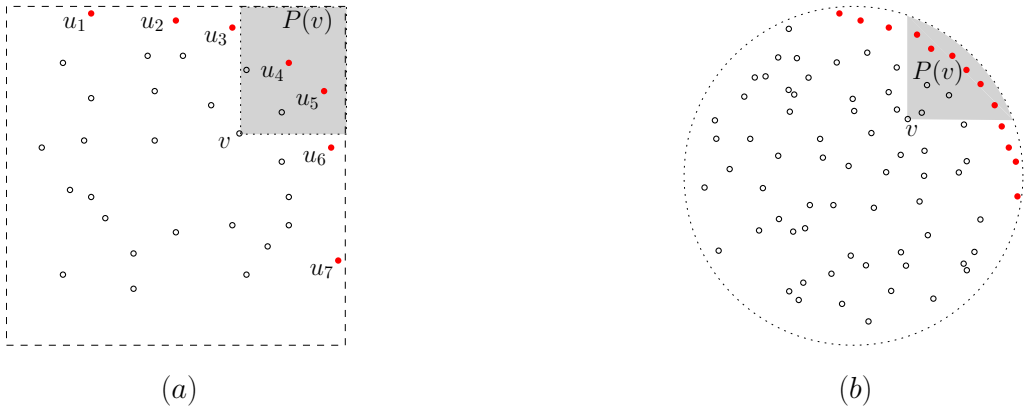

Figure 1 The diagram shows $\mathrm{MAX}(S_n)$ for two point sets S_n. In both (a) and (b) the circles denote the points in S_n and the (red) filled circles are $\mathrm{MAX}(S_n)$. If the points are considered as being drawn from region D, $P(v)$, as introduced in Def. 2, denotes the region in D that dominates v. In (a), D is the dotted square; in (b), D is the dotted circle.

The problems of finding and estimating the number of maximal points of a set in $\Re^2$ appear very often in many fields under different names: *maximal vectors*, *skylines*, *Pareto frontier/points* and others, see e.g. [5, 12, 15, 17, 18] for a more exhaustive history of the problems, uses in Computer Science and further references, Sections 1 and 2 in [7].

Let S_n denote a set of n points chosen Independently Identically Distributed (IID) from some 2-D distribution $\mathbf{D}$ and $M_n = |\mathrm{MAX}(S_n)|$ be the random variable counting the number of maximal points in S_n. Because maxima are so ubiquitous, understanding the expected number of maxima has been important in different areas and many properties of M_n have been studied. More specifically, if $\mathbf{D}$ is the uniform distribution drawn from an L_p ball with $p \geq 1$, then it is well known [2, 6, 12, 14], that

- If $p = \infty$, then $\mathbf{E}[M_n] = H_n \sim \ln n$.
  The same result holds if the points are drawn from some distribution $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$ where $\mathbf{X}$ and $\mathbf{Y}$ are any two 1-dimensional distributions that are independent of each other.
- If $p \geq 1$, then $\lim_{n\to\infty} \frac{\mathbf{E}[M_n]}{\sqrt{n}} = C_p$, where C_p is a constant dependent only upon p.
- Similar upper bounds to the above, i.e., that $\mathbf{E}[M_n] = O(\sqrt{n})$, derived using similar techniques, are known if $\mathbf{D}$ is a *uniform* distribution from ANY convex region [11].

It is also known [16] that if the n points are chosen IID from a 2-D Gaussian distribution then $\mathbf{E}[M_n] \sim \ln n$. There are also generalizations of these results (both the $\mathbf{B}_p$ ones and the Gaussian one) to higher dimensions. See [14] for a table containing most known results.

Surprisingly, given the importance of the problem, not much is known for other distributions. The motivation for this work is to extend the family of distributions for which $\mathbf{E}[M_n]$ can be derived.

Consider a point u originally generated from a uniform distribution over a unit L_p ball but measured or reported with an error, in the L_q metric, of at most δ. The actual reported point can be equivalently considered as being chosen from a new distribution which we denote by $\mathbf{B}_p + \delta\mathbf{B}_q$ (the next section provides formal definitions). The support of this distribution is the Minkowksi sum of the two balls but the distribution is not uniform over this support. Fig. 2 shows the support of $\mathbf{B}_p + \delta\mathbf{B}_q$, for different values of p and q.

Although the problem described above originally assumed small δ, it is well defined for all $\delta > 0$, which is the problem analyzed in this paper. More specifically, the motivation for the present work is twofold:

- Explain how $\mathbf{E}[M_n]$ changes when the distribution is perturbed.
  (Note: the perturbation size δ may be specified as a function of the sample size n.)
- Increase the families of distributions for which $\mathbf{E}[M_n]$ is understood.

The idea of analyzing how quantities change under perturbations could also be considered from the perspective of *smoothed analysis* [20, 21]. In the classic setting, smoothed analysis of the number of maxima would mean analyzing how, given a *fixed* set S_n, $\mathbf{E}[M_n]$ would change under small perturbations (as a function of the original set S_n). This was the approach in [9, 8] (see similar work for convex hulls in [10]). This paper differs in that it is the *Distribution* that is being smoothed (or convoluted) and not the point set. This paper also differs from recent work [22, 1] on the *most-likely* skyline and convex hull problems. Those papers assume each point has a given probability distribution and are attempting to find the subset of points that has the highest probability of being the skyline (or convex hull).

Outline of the paper. The next section defines the problem and states and explains our results. Sec. 3 describes key technical and conceptual ideas and tools used to achieve the main result. Sec. 4 describes how these tools are used to derive the result. Sec. 5 provides a review and a collection of open problems and possible extensions.

Due to space limitations, the proofs of many of the lemmas and theorems are not included. For the full proofs, please see the extended version of this paper [13] posted on Arxiv.

2 Definitions and Results

Let "$p \in [1, \infty)$" and "$p \geq 1$" both denote that p is a finite real number ≥ 1. $p = \infty$ also being permitted will be denoted by $p \in [1, \infty]$.

Recall: Let $\delta \geq 0$.
For $u \in \Re^2$, $\delta u = (\delta \cdot u.x, \delta \cdot u.y)$. For $u, v \in \Re^2$, $u + v = (u.x + v.x, u.y + v.y)$.
If $D \subseteq \Re^2$, $\delta D = \{(\delta u : u \in D\}$.
For $D_1, D_2 \subseteq \Re^2$, $D_1 + D_2 = \{u_1 + u_2 : u_1 \in D_1, u_2 \in D_2\}$ will denote the *Minkowski sum* of D_1 and D_2.
For $u \in \Re^2$, $u + D$ will denote $\{u\} + D$.

Balls and Unit Balls: Let $u \in \Re^2$, $r > 0$ and $p \in [1, \infty)$. Define:
- The L_p *ball of radius r around u* as $B_p(u, r) = \{(x, y) : |x - u.x|^p + |y - u.y|^p \leq r^p\}$.
- The L_∞ *ball of radius r around u* as $B_\infty(u, r) = \{(x, y) : \max(|x - u.x|, |y - u.y|) \leq r\}$.
- The respective *unit balls* as $B_p = B_p((0, 0), 1)$ and $B_\infty = B_\infty((0, 0), 1)$.

Set $a_p = \text{Area}(B_p)$ to be the area of the L_p unit ball. Then $a_\infty = 4, a_1 = 2, a_2 = \pi$. We use the fact that $a_p = \Theta(1)$.

Generation of a probability distribution: Let $\mathbf{D}$ be a distribution with support $D \subset \Re^2$. Then
- If $\delta \geq 0$, the distribution $\delta\mathbf{D}$ is generated by choosing a point u using $\mathbf{D}$ and then returning the point δu.
- Let $\mathbf{D}_1, \mathbf{D}_2$ be two distributions over $\Re^2$. Generate the *convolution* $\mathbf{D}_1 + \mathbf{D}_2$ by choosing a point u_1 from $\mathbf{D}_1$ and a point u_2 from $\mathbf{D}_2$ and returning $u_1 + u_2$.
- A set $S_n = \{u_1, \ldots, u_n\}$ is said to be *chosen from* $\mathbf{D}$ if each u_i is generated *independently and identically distributed* (IID) using the distribution $\mathbf{D}$.

Uniform distribution on unit balls: For all $p \in [1, \infty]$, $\mathbf{B}_p$ will denote the uniform distribution that selects a point uniformly from B_p. This distribution has support B_p with uniform density $1/a_p$ within B_p.

Convolution of two distributions: Let $\mathbf{B}_p + \delta\mathbf{B}_q$ be the convolution of distributions $\mathbf{B}_p$ and $\delta\mathbf{B}_q$.
$(\mathbf{B}_p + \delta\mathbf{B}_q)$'s *support* of this distribution is the Minkowski sum $B_p + \delta B_q$. Observe that the density of $\mathbf{B}_p + \delta\mathbf{B}_q$ is **not** uniform in $B_p + \delta B_q$. It is this non-uniformity that will cause complications in calculating $\mathbf{E}[M_n]$. The main result of this paper is

▶ **Theorem 1.** *Fix p, q so that either $p, q \in \{1, 2, \infty\}$ or $p = \infty$ and $q \geq 1$.*
Let S_n be n points chosen from the distribution $\mathbf{B}_p + \delta\mathbf{B}_q$ and $M_n = |\mathrm{MAX}(S_n)|$.
Let $\delta \geq 0$ be a function of n. Then $\mathbf{E}[M_n]$ behaves as below:

	(a)	(b)	(c)	(d)	(e)	(f)
	$\mathbf{D} =$	$0 \leq \delta$				$\delta = 1$
(i)	$\mathbf{B}_\infty + \delta\mathbf{B}_\infty$	$\Theta(\ln n)$				$\Theta(\ln n)$
		$\delta \leq \frac{1}{\sqrt{n}}$	$\frac{1}{\sqrt{n}} \leq \delta \leq 1$	$1 \leq \delta \leq \sqrt{n}$	$\sqrt{n} \leq \delta$	
(ii)	$\mathbf{B}_1 + \delta\mathbf{B}_1$	$\Theta(\sqrt{n})$	$\Theta\left(\frac{n^{1/3}}{\delta^{1/3}}\right)$	$\Theta\left(\delta^{1/3}n^{1/3}\right)$	$\Theta(\sqrt{n})$	$\Theta\left(n^{1/3}\right)$
(iii)	$\mathbf{B}_2 + \delta\mathbf{B}_2$	$\Theta(\sqrt{n})$	$\Theta\left(\frac{n^{2/7}}{\delta^{3/7}}\right)$	$\Theta\left(\delta^{3/7}n^{2/7}\right)$	$\Theta(\sqrt{n})$	$\Theta\left(n^{2/7}\right)$
		$\delta \leq \frac{1}{\sqrt{n}}$	$\frac{1}{\sqrt{n}} \leq \delta \leq \sqrt{n}$		$\sqrt{n} \leq \delta$	
(iv)	$\mathbf{B}_\infty + \delta\mathbf{B}_q$	$\Theta(\ln n)$	$\Theta\left(\ln n + \sqrt{\delta}n^{1/4}\right)$		$\Theta(\sqrt{n})$	$\Theta\left(n^{1/4}\right)$
		$\delta \leq \frac{1}{\sqrt{n}}$	$\frac{1}{\sqrt{n}} \leq \delta \leq n^{1/26}$	$n^{1/26} \leq \delta \leq \sqrt{n}$	$\sqrt{n} \leq \delta$	
(v)	$\mathbf{B}_1 + \delta\mathbf{B}_2$	$\Theta(\sqrt{n})$	$\Theta\left(\frac{n^{2/7}}{\delta^{3/7}}\right)$	$\Theta\left(\sqrt{\delta}n^{1/4}\right)$	$\Theta(\sqrt{n})$	$\Theta\left(n^{2/7}\right)$

Interpretation of the table:

1. When $p = q = \infty$, M_n has exactly the same distribution as if S_n were chosen from $\mathbf{B}_\infty$, so row (i) is an uninteresting case, only included for completeness.
2. When δ is small enough ($\leq 1/\sqrt{n}$), $\mathbf{E}[M_n]$ behaves almost as if S_n were chosen from $\mathbf{B}_p$ and when δ is large enough ($\geq \sqrt{n}$) $\mathbf{E}[M_n]$ behaves almost as if S_n were chosen from $\mathbf{B}_q$. This is reflected in columns (b) and (e).
3. Lemma 8 states that M_n has the same distribution for S_n chosen from both $\mathbf{B}_p + \delta\mathbf{B}_q$ and $\mathbf{B}_q + \frac{1}{\delta}\mathbf{B}_p$. Thus row (iv) gives the behavior for $\mathbf{B}_q + \delta\mathbf{B}_\infty$ for any $q \geq 1$ and row (v) the behavior for $\mathbf{B}_2 + \delta\mathbf{B}_1$.
4. When $p = q \in \{1, 2\}$, $\mathbf{E}[M_n]$ starts at $\Theta(\sqrt{n})$, smoothly decreases until reaching $\delta = 1$ and then increases again until reaching $\Theta(\sqrt{n})$. The behavior at $\delta = 1$ is different for $p = q = 1$ and $p = q = 2$. In both cases there is symmetry between δ and $1/\delta$ (from Lemma 8).
5. When $p = 1$, $q = 2$ there is no symmetry. The behavior starts at $\Theta(\sqrt{n})$, decreases to $\Theta\left(n^{7/26}\right)$ at $\delta = n^{1/26}$ and then increases again *at a different rate* to $\Theta(\sqrt{n})$.
6. When $p = \infty$, the behavior is asymptotically equivalent for all $q \in [1, \infty)$, not just $q = 1, 2$. The only difference is in the value of the constant hidden by the Θ. The behavior starts at $\Theta(\ln n)$, stays there for a short while and then smoothly increases to $\Theta(\sqrt{n})$.

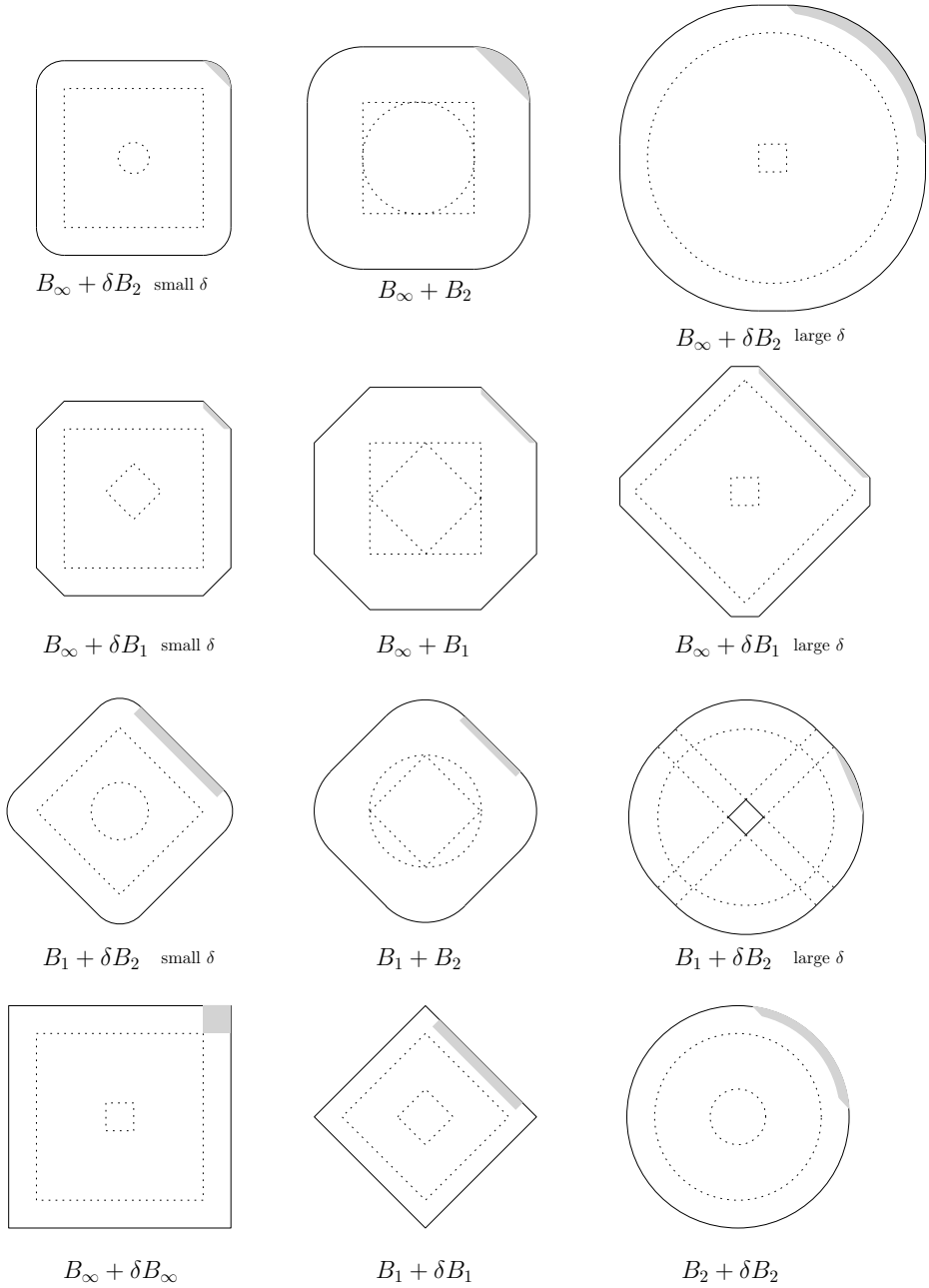

Figure 2 Illustrations of the supports of some of the different distributions in the form $\mathbf{B}_p + \delta\mathbf{B}_q$ examined in Theorem 1. The dotted lines denote the B_p and δB_q balls centred at 0. Note that in all cases the density is uniform near the centre of the support but then decreases to 0 as the boundary is approached. The grey areas denote, approximately, where the maxima of S_n are concentrated.

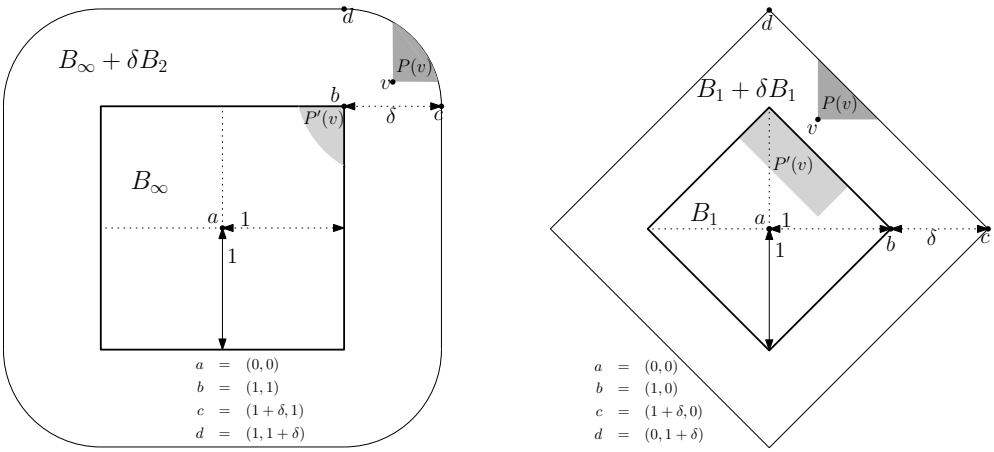

Figure 3 Illustration of definitions of $P(v)$ and $P'(v)$ for $B_p + \delta B_q$. Left side is is $B_\infty + \delta B_2$; right is $B_1 + \delta B_1$. In both diagrams the interior ball (heavy boundary) is the B_p ball centered at the origin a. $P(v)$ is the set of points in $B_p + \delta B_q$ that dominate v and $P'(v)$ is the preimage of v in B_p.

3 Basic Lemmas

The following collection of Lemmas comprise the basic toolkit used to derive Theorem 1.

Recall: Let $\mathbf{D}$ be a distribution over $\Re^2$, $x \in \Re^2$ and $A \subset \Re^2$ a measurable region. Then $f_{\mathbf{D}}(x)$ will denote the *density function* of $\mathbf{D}$, and $\mu_{\mathbf{D}}(A) = \int_A f_{\mathbf{D}}(x)dx$ will denote the *measure* of A under distribution $\mathbf{D}$. If $\mathbf{D}$ is understood, we often simply write $f(x)$ and $\mu(A)$.

▶ **Definition 2.** *(See Fig. 3)*
Let $D \subseteq \Re^2$, $v \in D$ and $A \subseteq D$.
Define: $P(v) = \{u \in D : u \text{ dominates } v\} \cup \{v\}$, and $P(A) = \bigcup_{v \in A} P(v)$.
Say that A is dominant in D *or a* dominant region in D, *if $P(A) = A$.*

Note that, by definition, $\forall v \in D$, $P(v)$ is a dominant region in D. It is straightforward to see that

▶ **Lemma 1.** *Let v and S_n be chosen from $\mathbf{D}$ and $A \subseteq D$. Then*

$$\begin{array}{lrcl} (a) & \Pr(v \in A) & = & \mu(A). \\ (b) & \mathbf{E}\left[|A \cap S_n|\right] & = & n\mu(A). \\ (c) & \Pr(A \cap S_n = \emptyset) & = & (1-\mu(A))^n. \end{array}$$

The following observation will be used to prove most of our lower bounds.

▶ **Lemma 2** (Lower Bound). *Let S_n be chosen from $\mathbf{D}$. Further let $A_1, A_2, \ldots, A_m$ be a collection of pairwise disjoint dominant regions in D with $\mu(A_i) = \Omega(1/n)$ for all i. Then*

$$\mathbf{E}\left[M_n\right] \geq \mathbf{E}\left[\left|\mathrm{MAX}\left(S_n \cap \bigcup_{i=1}^{m} A_i\right)\right|\right] = \Omega(m).$$

Proof. From Lemma 1, $\Pr(S_n \cap A_i = \emptyset) = (1-\mu(A_i))^n$. Thus $\mu(A_i) = \Omega(1/n)$ implies

$$\Pr(S_n \cap A_i \neq \emptyset) = 1 - \Pr(S_n \cap A_i = \emptyset) = \Omega(1).$$

If region A is dominant then points in A can only be dominated by other points in A then $A \cap \mathrm{MAX}(S_n) = \mathrm{MAX}(S_n \cap A)$. Since each A_i is dominant, this implies

$$\mathbf{E}\left[|\mathrm{MAX}(S_n) \cap A_i|\right] \geq \Pr(S_n \cap A_i \neq \emptyset) = \Omega(1).$$

Since the A_i are pairwise disjoint,

$$\mathbf{E}\left[|MAX(S_n)|\right] \geq \mathbf{E}\left[\left|MAX(S_n) \cap \left(\bigcup_i A_i\right)\right|\right] \geq \sum_{i=1}^{m} \Omega(1) = \Omega(m).$$ ◀

▶ **Definition 3.** *(See Fig. 3)*
Let $D = B_p + \delta B_q$. For $v \in D$ define the preimage *of v in B_p as*

$$P'(v) = B_q(v, \delta) \cap B_p = (v + \delta B_q) \cap B_p.$$

▶ **Lemma 3.** *Fix $p, q \in [1, \infty]$. Let $\mathbf{D} = \mathbf{B}_p + \delta \mathbf{B}_q$ and let v be a point chosen from $\mathbf{D}$. Let $A \subseteq \Re^2$. Then*

$$f(v) = \frac{1}{a_p a_q} \frac{\mathrm{Area}(\{u \in B_p \,:\, v - u \in \delta B_q\})}{\delta^2} = \frac{1}{a_p a_q} \frac{\mathrm{Area}(P'v)}{\delta^2} \tag{1}$$

$$\mu(A) = \frac{1}{a_p a_q} \int_{u \in B_p} \frac{\mathrm{Area}((u + \delta B_q) \cap A)}{\delta^2} du. \tag{2}$$

Proof. Note that for $u \in B_p$, $f_{\mathbf{B}_p}(u) = \frac{1}{a_p}$ and for $u' \in \delta B_q$, $f_{\delta \mathbf{B}_q}(u') = \frac{1}{a_q \delta^2}$. To see Eq. 2,

$$\mu(A) = \int_{u \in \mathbf{B}_p} \left(\int_{\substack{w \in \delta B_q \\ u + w \in A}} f_{\delta \mathbf{B}_q}(w) dw \right) f_{\mathbf{B}_p}(u) du = \frac{1}{a_p a_q} \int_{u \in B_p} \frac{\mathrm{Area}\left((u + \delta B_q) \cap A\right)}{\delta^2} du.$$

For Eq. 1, use a change of variables $v = u + w$,

$$\begin{aligned} \mu(A) = & \frac{1}{a_p a_q \delta^2} \int_{u \in B_p} \left(\int_{\substack{w \in \delta B_q \\ u + w \in A}} dw \right) du \\ = & \frac{1}{a_p a_q \delta^2} \int_{u \in B_p} \left(\int_{\substack{v \in u + \delta B_q \\ v \in A}} dv \right) du = \frac{1}{a_p a_q} \int_{v \in A} \frac{\mathrm{Area}\left\{u \in B_p \,:\, v - u \in \delta B_q\right\}}{\delta^2} dv. \end{aligned}$$

Differentiating around v yields Eq. 1. ◀

▶ **Lemma 4.** *Fix $p, q \in [1, \infty]$. Let $\mathbf{D} = \mathbf{B}_p + \delta \mathbf{B}_q$ and $\kappa > 0$ be any constant. Then*

(a)	$v \in D$	$\Rightarrow$	$f(v) = O(1)$.
(b)	$v \in B_p$ and $\delta \leq \kappa$	$\Rightarrow$	$f(v) = \Theta(1)$.
(c)	$A \subseteq D$	$\Rightarrow$	$\mu(A) = O(\mathrm{Area}(A))$.
(d)	$A \subseteq B_p$ and $\delta \leq \kappa$	$\Rightarrow$	$\mu(A) = \Theta(\mathrm{Area}(A))$.

The constants implicit in the $O()$ in (a) and (c) are only dependent upon p, q, while the constants implicit in the $\Theta()$ in (b) and (d) are only dependent upon p, q, κ.

Proof.
(a) Use the fact that, for $\forall u \in B_p$,

$$\mathrm{Area}(B_p \cap (u + \delta B_q))) \leq \mathrm{Area}(u + \delta B_q) = a_q \delta^2,$$

so from Eq. 1, $f(v) = O(1)$.

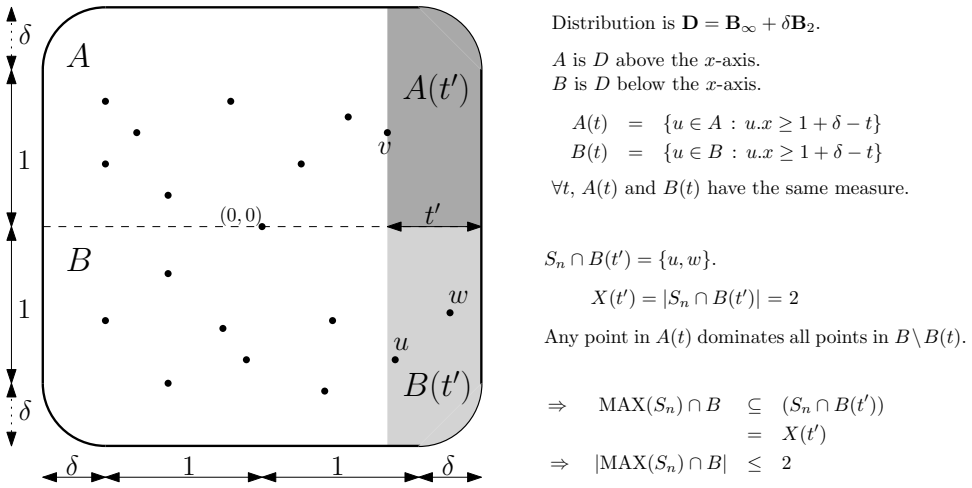

Figure 4 Illustration of Lemmas 5 and 6. The regions A and B are each swept by parameter t and it is required that $\mu(B(t)) = O(\mu(A(t))$. In the case above, by the symmetry of distribution $\mathbf{D}$, $\mu(B(t)) = \mu(A(t))$ trivially. t' is the first time a point in $A(t)$ is found. Since every point in $A(t)$ dominates all points in $B \setminus B(t)$, all maxima in $S_n \cap B$ must be in $B(t')$. The definition of t' intuitively implies that $\mu(A(t')) \sim \frac{1}{n}$ so, also intuitively, the expectation of $|S_n \cap B_n|$ should be $n\mu(B(t')) \sim 1$. This is proven formally in the text.

(b) If $u \in B_p$ then

$$\mathrm{Area}(B_p \cap (u + \delta B_q))) \geq c\mathrm{Area}(u + \delta B_q) = ca_q\delta^2,$$

where c is only dependent upon p, q, κ. Thus, from Eq. 1, $f(v) = \Theta(1)$.

The proofs for (c) and (d) follow from plugging (a) and (b) into Eq. 2. ◀

▶ **Lemma 5.** *(See Fig. 4)*
Let $\mathbf{D}$ be any distribution with a continuous density function $f(u)$ and S_n a set of points chosen from $\mathbf{D}$. Let A, B be two disjoint regions in the support D that are parameterized by $t \in [0, T]$ and satisfy:

- $\mu(A(0)) = \emptyset$.
- $A(T) = A$; $B(T) = B$.
- *(Monotonicity in t)* $\forall t_1 < t_2$, $A(t_1) \subseteq A(t_2)$ *and* $B(t_1) \subseteq B(t_2)$.
- $\mu(B(t)), \mu(A(t))$ *are both continuous in* t.
- *(Asymptotic dominance in measure)* $\forall t$, $\mu(B(t)) = O(\mu(A(t))$.

Define the random variables

$$X = |S_n \cap B(t')|, \qquad t' = \begin{cases} \min\{t : A(t) \cap S_n \neq \emptyset\} & \text{if } A \cap S_n \neq \emptyset, \\ T & \text{if } A \cap S_n = \emptyset. \end{cases}$$

Then, $\mathbf{E}[X] = O(1)$. (3)

Proof. W.l.o.g. rescale t so that $\mu(A(t)) = t$, and $T = \mu(A)$.

The proof's intuition is that since the "first" point in A appears at t', then $\mu(A(t')) \sim \frac{1}{n}$. As B is asymptotically dominated by A, $\mu(B(t'))$=$O(1/n)$ and $\mathbf{E}[X(t')] = n\mu(B(t')) = O(1)$.

Formally, by the continuity of the measure, $\Pr(|S_n \cap A(t')| = 1) = 1$. So we may assume that $|D \setminus A(t')| = n - 1$.

Conditioned on known t', the remaining $n-1$ points in S_n are chosen from $D \setminus A(t')$ with the associated conditional distribution. If u is one of those $n-1$ points,

$$\Pr\left(u \in B(t') \;\middle|\; t'\right) = \frac{\mu(B(t'))}{\mu(D \setminus A(t'))} = \frac{\mu(B(t'))}{1 - \mu(A(t'))}.$$

Thus, conditioning on t', and applying Lemma 1(b)

$$\mathbf{E}\left[X \;\middle|\; t'\right] = (n-1)\frac{\mu(B(t'))}{1 - \mu(A(t'))},$$

$$\text{therefore } \mathbf{E}[X] = \mathbf{E}\left[\mathbf{E}\left[X \;\middle|\; t'\right]\right] = \mathbf{E}\left[(n-1)\frac{\mu(B(t'))}{1 - \mu(A(t'))}\right].$$

From the definition of t' and Lemma 1 (c), $\mu(A(t')) > 1/2$ with exponentially low probability. Therefore, recalling that $\mu(A(t)) = t$,

$$\mathbf{E}[X] = (n-1)\mathbf{E}[O(\mu(B(t')))] = (n-1)\mathbf{E}[O(\mu(A(t')))] = (n-1)O(\mathbf{E}[t']).$$

$$\text{Using Lemma 1 (c)}: \mathbf{E}[t'] = \int_{\alpha=0}^{T} \Pr(t' \geq \alpha)d\alpha = O\left(\tfrac{1}{n-1}\right).$$ ◀

▶ **Lemma 6** (Sweep). *(See Fig. 4)*
Let $\mathbf{D}$ *be any distribution with a continuous density function* $f(u)$*, and let* S_n *be a set of points chosen from* $\mathbf{D}$*.*

Let A, B *be two disjoint regions in the support* D *that are parameterized by* $t \in [0, T]$*, satisfy conditions 1-3 of Lemma 5 and, in addition satisfy that*

$$\forall t \in [0, T], \quad \textit{if } u \in A(t) \textit{ and } v \in B \setminus B(t) \textit{ then } u \textit{ dominates } v.$$

In such a case we say that A continuously dominates B*. Then*

$$\mathbf{E}[|\mathrm{MAX}(S_n) \cap B|] = O(1). \tag{4}$$

Proof. By the definition of t', $|A(t') \cap S_n| \geq 1$. Since all points in $B \setminus B(t')$ are dominated by all points in $A(t')$, $\mathrm{MAX}(S_n) \cap (B \setminus B(t')) = \emptyset$. Thus from Lemma 5,

$$\mathbf{E}[|\mathrm{MAX}(S_n) \cap B|] = \mathbf{E}[|\mathrm{MAX}(S_n) \cap B(t')|] \leq \mathbf{E}[|S_n \cap B(t')|] = O(1).$$ ◀

▶ **Corollary 7.** *Fix* $p, q \in [1, \infty]$ *and choose* S_n *from* $\mathbf{D} = \mathbf{B}_p + \delta\mathbf{B}_q$*. Let* Q_1 *be the positive (upper-right) quadrant of the plane and* O_1 *the first octant , i.e.,* $Q_1 = \{u \in \Re^2 : 0 \leq u.x, 0 \leq u.y\}$ *and* $O_1 = \{u \in \Re^2 : 0 \leq u.y \leq u.x\}$*. Then*

$$\mathbf{E}[M_n] = \mathbf{E}[|\mathrm{MAX}(S_n)|] = \mathbf{E}[|Q_1 \cap \mathrm{MAX}(S_n)|] + O(1) \tag{5}$$

$$= \Theta\Big(\mathbf{E}[|O_1 \cap \mathrm{MAX}(S_n)|]\Big). \tag{6}$$

Proof. Restrict $t \in [0, 2 + 2\delta]$ and set

$$A = D \cap \{u \in \Re^2 : u.y \geq 0\}, \qquad A(t) = \{u \in A : u.x \geq 1 + \delta - t\},$$
$$B = D \cap \{u \in \Re^2 : u.y < 0\}, \qquad B(t) = \{u \in B : u.x \geq 1 + \delta - t\}.$$

Conditions (1) and (2) of Lemma 5 trivially hold. Condition (3) holds because, by x-axis symmetry, $\mu(B(t)) = \mu(A(t))$. The additional condition of Lemma 6 holds because every point in $B \setminus B(t)$ is below and to the left of every point in $A(t)$. Thus the expected number of maximal points in S_n *below* the x-axis is $O(1)$. Note that this is *independent* of n.

Similarly, the expected number of maximal points to the left of the y-axis is $O(1)$. This proves Eq. 5.

To prove Eq. 6 define the second octant to be $O_2 = \{u \in \Re^2 : 0 \le u.x \le u.y\}$. By the symmetry between the x and y coordinates in the distribution,

$$\mathbf{E}\left[|O_1 \cap \mathrm{MAX}(S_n)|\right] = \mathbf{E}\left[|O_2 \cap \mathrm{MAX}(S_n)|\right].$$

Futhermore, since O_1 and O_2 partition Q_1,

$$\mathbf{E}\left[|Q_1 \cap \mathrm{MAX}(S_n)|\right] = \mathbf{E}\left[|O_1 \cap \mathrm{MAX}(S_n)|\right] + \mathbf{E}\left[|O_2 \cap \mathrm{MAX}(S_n)|\right] = 2\mathbf{E}\left[|O_1 \cap \mathrm{MAX}(S_n)|\right].$$

Thus

$$\mathbf{E}\left[M_n\right] = \mathbf{E}\left[Q_1 \cap |\mathrm{MAX}(S_n)|\right] + O(1) = \Theta\left(\mathbf{E}\left[|O_1 \cap \mathrm{MAX}(S_n)|\right]\right). \quad \blacktriangleleft$$

The fact that for $\delta > 0$, u dominates v if and only if δu dominates δv implies the following result which is used very often in this work,

▶ **Lemma 8** (Scaling). *Fix $p, q \in [1, \infty]$, $\mathbf{D} = \mathbf{B}_p + \delta \mathbf{B}_q$ and $\mathbf{D}' = \mathbf{B}_q + \frac{1}{\delta}\mathbf{B}_p$. Let S_n be n points chosen from $\mathbf{D}$ and let S'_n be n points chosen from $\mathbf{D}'$. Then $|\mathrm{MAX}(S_n)|$ and $|\mathrm{MAX}(S'_n)|$ have exactly the same distribution. In particular, $\mathbf{E}\left[|\mathrm{MAX}(S_n)|\right] = \mathbf{E}\left[|\mathrm{MAX}(S'_n)|\right]$.*

Proof. Let $S_n = \{u_1, \ldots, u_n\}$ be chosen from $\mathbf{D}$. Recall that the process of choosing point u from $\mathbf{D}$ is to choose w from $\mathbf{B}_p$, v from $\mathbf{B}_q$ and return $u = w + \delta v$. Choosing a point u' from $\mathbf{D}'$ is the same except that it returns $u' = v + \frac{1}{\delta}w = \frac{1}{\delta}u$. Thus the distribution of choosing $S_n = \{u_1, \ldots, u_n\}$ from $\mathbf{D}$ is exactly the same as choosing $S_n = \{\frac{1}{\delta}u_1, \ldots, \frac{1}{\delta}u_n\}$ from $\mathbf{D}'$.

Finally, note that dominance is invariant under multiplication by a scalar, i.e., p_i dominates p_j if and only if $\frac{1}{\delta}p_i$ dominates $\frac{1}{\delta}p_j$.

Thus $|\mathrm{MAX}(S_n)|$ and $|\mathrm{MAX}(S'_n)|$ have the same distribution, so $\mathbf{E}\left[|\mathrm{MAX}(S_n)|\right] = \mathbf{E}\left[|\mathrm{MAX}(S'_n)|\right]$. ◀

The next lemma formalizes the intuition that for small values of δ, the value of $\mathbf{E}\left[M_n\right]$ for $\mathbf{B}_p + \delta\mathbf{B}_q$ is the same as the value for $\mathbf{B}_p$.

▶ **Lemma 9** (Limiting Behavior). *Let $p \in [1, \infty]$, $q \in [1, \infty)$, $\delta = O(1/\sqrt{n})$ and S_n chosen from $\mathbf{D} = \mathbf{B}_p + \delta\mathbf{B}_q$. Then*

$$\mathbf{E}\left[M_n\right] = \begin{cases} \Theta(\ln n) & \text{if } p = \infty, \\ \Theta(\sqrt{n}) & \text{if } p \neq \infty. \end{cases}$$

4 General approach to proving Theorem 1

Note that if u is chosen from $\mathbf{B}_\infty$, then $u.x$ and $u.y$ are independent random variables. Thus, for any $\delta > 0$ if v is chosen from $\mathbf{D} = \mathbf{B}_\infty + \delta\mathbf{B}_\infty$, $v.x$ and $v.y$ are independent random variables. As noted in the introduction, this means that if S_n is chosen from $\mathbf{D}$, $\mathbf{E}\left[M_n\right]$ is exactly the same as if S_n was chosen from $\mathbf{B}_\infty$, i.e., $\mathbf{E}\left[M_n\right] = \Theta(\ln n)$, proving row (i).

Lemma 9 combined with Lemma 8 imply the limiting behavior in columns (b) and (e) of the table in Theorem 1. Note too that for rows (ii) and (iii), column (d) follows directly from applying Lemma 8 to column (c).

Thus, proving Theorem 1 reduces to proving cells (ii) c, (iii) c, (iv) c, d and (v) c, d.

Proving Theorem 1 will require case-by-case analyses of $\mathbf{D} = \mathbf{B}_p + \delta\mathbf{B}_q$ for the different pairs p, q. The analysis for each pair will all follow the same 4 step pattern:

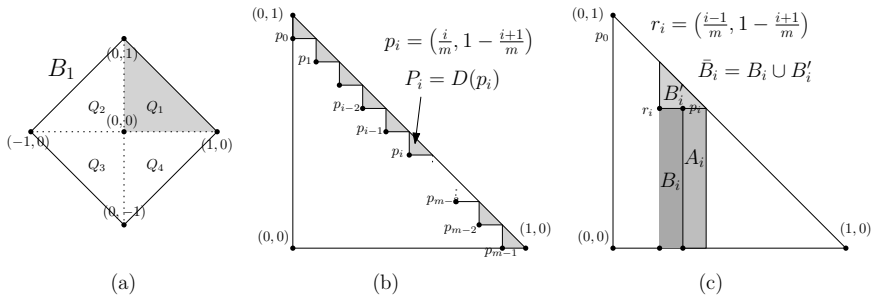

Figure 5 Illustration of proof $\mathbf{E}[M_n] = \Theta(\sqrt{n})$ when S_n is chosen from $\mathbf{B}_1$ All but $O(1)$ maxima will be in quadrant Q_1; (b) and (c) illustrate Q_1. (b) illustrates the lower bound and (c) the upper.

4.1 A Simple Example: $\mathbf{D} = \mathbf{B}_1$

Before sketching our results it is instructive to see how the Lemmas in the previous section can be used to re-derive that fact that, if $\mathbf{D} = \mathbf{B}_1$ then $\mathbf{E}[M_n] = \Theta(\sqrt{n})$. See Fig. 5.

Even though the behavior for $\mathbf{D} = \mathbf{B}_1$ is already well known we provide this to illustrate the generic steps for deriving $\mathbf{E}[M_n]$. These are exactly the same steps that are needed when $\mathbf{D} = \mathbf{B}_p + \delta\mathbf{B}_q$ and this example permits identifying where the complications can arise in those more general cases. Set $m = \lfloor\sqrt{n}\rfloor$ and let p_i, r_i be the points defined in the figure with $P_i = P(p_i)$ and $B'_i = P(r_i)$. Also set

$$B_i = \left\{(x,y) \,:\, \frac{i-1}{m} \le x \le \frac{i}{m},\ 0 \le y \le 1 - \frac{i+1}{m}\right\}, \quad A_i = \left(\frac{1}{m}, 0\right) + B_i$$

and $\bar{B}_i = B_i \cup B'_i$. Finally, for $0 \le t \le (1+i)/m$ set $B_i(t) = B_i \cap \{(x,y) \,:\, y \le (1+i)/m - t\}$ and $A_i(t) = \left(\frac{1}{m}, 0\right) + B_i(t)$. The steps in the derivation are.

Step 1: *Restricting to first Quadrant:*
Corollary 7 states that $\mathbf{E}[M_n] = \mathbf{E}[|Q_1 \cap \mathrm{MAX}(S_n)|] + O(1)$.

Step 2: *Calculating Density and Measure:*
Because $\mathbf{D}$ has a uniform density, $\mu(A) = \Theta(\mathrm{Area}(A))$ for all regions $A \subseteq D$.

Step 3: *Lower Bound:*
The P_i are a collection of m pairwise disjoint dominant regions with

$$\mu(P_i) = \Theta(\mathrm{Area}(P_i)) = \Theta(m^{-2}) = \Theta(1/n).$$

Thus, from Lemma 2, $\mathbf{E}[M_n] = \Omega(m) = \Omega(\sqrt{n})$.

Step 4: *Upper bound:*
Note that $Q_1 \cap D = \left(\bigcup_{i=1}^{m-1} \bar{B}_i\right) \cup B'_m$ so

$$\mathbf{E}[|\mathrm{MAX}(S_n) \cap Q_1|] = \mathbf{E}\left[\left|\mathrm{MAX}(S_n) \cap \left(\bigcup_{i=1}^{m} \bar{B}_i\right)\right|\right] + \mathbf{E}[|\mathrm{MAX}(S_n) \cap B'_m|],$$

$$\mathbf{E}\left[\left|\mathrm{MAX}(S_n) \cap \left(\bigcup_{i=1}^{m} \bar{B}_i\right)\right|\right] \le \sum_{i=1}^{m} \mathbf{E}[|\mathrm{MAX}(S_n) \cap B_i|] + \sum_{i=1}^{m} \mathbf{E}[|\mathrm{MAX}(S_n) \cap B'_i|].$$

Furthermore, $\forall i$, $\mu(B'_i) = \Theta(\mathrm{Area}(B'_i)) = \Theta(1/n)$. Thus

$$\forall i, \quad \mathbf{E}[|\mathrm{MAX}(S_n) \cap B'_i|] \le \mathbf{E}[|S_n \cap B'_i|] = O(n\mu(B'_i)) = O(1).$$

Since $m = O(\sqrt{n})$ this yields

$$\mathbf{E}\left[|\mathrm{MAX}(S_n) \cap Q_1|\right] \leq \sum_{i=1}^{m} \mathbf{E}\left[|\mathrm{MAX}(S_n) \cap B_i|\right] + O(\sqrt{n}).$$

The crucial observation is that, $\forall i$, A_i continuously dominates B_i as defined in Lemmas 5 and 6. Thus, plugging into Lemma 6 yields $\forall i, \quad \mathbf{E}\left[|\mathrm{MAX}(S_n) \cap B_i|\right] = O(1)$, leading to

$$\mathbf{E}\left[|\mathrm{MAX}(S_n) \cap Q_1|\right] = O(m) + O(\sqrt{n}) = O(\sqrt{n}).$$

Combining the $\mathbf{E}\left[|\mathrm{MAX}(S_n) \cap Q_1|\right] = \Omega(\sqrt{n})$ from step (3) with the $\mathbf{E}\left[|\mathrm{MAX}(S_n) \cap Q_1|\right] = O(\sqrt{n})$ from step (4) with step (1) gives the final result

$$\mathbf{E}\left[M_n\right] = \mathbf{E}\left[|\mathrm{MAX}(S_n) \cap Q_1|\right] + O(1) = \Theta(\sqrt{n}) + O(1) = \Theta(\sqrt{n}).$$

4.2 The general approach for $\mathrm{D} = \mathrm{B}_p + \delta \mathrm{B}_q$

For each p, q pair the proof of Theorem 1 follows the same four steps as the analysis of $\mathbf{D} = \mathbf{B}_1$ above.

Step 1: *Restricting to first Quadrant:*
Corollary 7 again states that $\mathbf{E}\left[M_n\right] = \mathbf{E}\left[|Q_1 \cap \mathrm{MAX}(S_n)|\right] + O(1)$.

Step 2: *Calculating Density $f(u)$ and Measure $\mu(A)$:*
This step is often quite technical. In the example $\mathbf{D} = \mathbf{B}_1$ case above, the density was constant. For general $\mathbf{D}$ this is no longer true. The density is constant in some region in the center of the support but decreases to zero as the boundary is approached. While Lemma 3 provides an integral formula for general $\mathbf{D}$ this, in many cases, is unusable. A substantial amount of technical work is involved in finding usable functional representations for the densities/measures in different parts of the support.

Step 3: *Lower Bounding $\mathbf{E}\left[M_n\right]$:*
For most cases this is a relatively straightforward application of Lemma 2 using the results of Step 2. In the general case, it is still necessary to identify a region that contains an asymptotically dominant number of maxima. It is then necessary to partition this region into pairwise disjoint dominant regions, all of which have measure $\Theta(1/n)$. Note that, unlike in the example $\mathbf{D} = \mathbf{B}_1$ case, these regions might no longer all have the same shape or size.

Step 4: *Upper bounding $\mathbf{E}\left[M_n\right]$:*
This is the most delicate part of the proof. It is proven using the Sweep Lemma (Lemma 6) with the major difficulties arising from how to decompose the support into regions that continuously dominate each other. This decomposition strongly depends upon *how* the measure/density is *represented* in Step 2 and can be very differently structured in different parts of the support. In particular, in the case $\mathbf{D} = \mathbf{B}_1 + \delta \mathbf{B}_2$, there are two different parts of the support that require two different decompositions and the decompositions must be designed so that the two upper bounds derived match each other.

More broadly, the density/measure representations developed for $\mathbf{D}_1 = \mathbf{B}_1 + \delta \mathbf{B}_1$ and $\mathbf{D}_2 = \mathbf{B}_2 + \delta \mathbf{B}_2$ are quite different. The analysis of $\mathbf{D}_3 = \mathbf{B}_1 + \delta \mathbf{B}_2$ which is the most delicate, combines the approaches developed for $\mathbf{D}_1$, $\mathbf{D}_2$. The analysis of $\mathbf{D}_4 = \mathbf{B}_\infty + \delta \mathbf{B}_q$ is different from the first three, but much more straightforward.

5 Conclusion

This paper developed a suite of tools for deriving the expected number of maximal points in a set of n points chosen IID from $\mathbf{B}_p + \delta\mathbf{B}_q$, which is the convolution of two distributions.

The results presented here seem to be the first general analysis of $\mathbf{E}[M_n]$ for non-uniform and non-Gaussian distributions. This paper is only a first step. Obvious next steps are

- The results in the paper were only proven for $p, q \in \{1, 2, \infty\}$ and $p = \infty, q \in [1, \infty.]$ The next step would be to attempt to extend the results to *all* pairs $p, q, \in [1, \infty]$.
- There is a rich literature stretching back more than fifty years on the average number of points on the *convex hull* of points chosen IID from a uniform distribution in a planar region or a Gaussian distribution, e.g., [14, 19]. It would be interesting to see how the convex hull evolves in the convoluted distributions $\mathbf{B}_p + \delta\mathbf{B}_q$.
  Such an analysis would require a much tighter understanding of how the distribution behaves "close" to the boundary of its support $B_p + \delta B_q$. One approach might be to introduce some form of measure weighting to the definition of *Macbeath-regions* [3] (which are a known technique for characterizing this boundary region).
- Finally, we note that the results on $\mathbf{E}[M_n]$ for n points chosen IID from a uniform distribution over an L_p ball have analogues in higher dimensions, i.e., $\Theta\left(\log^{d-1} n\right)$ if $p = \infty$ and $\Theta\left(n^{1-\frac{1}{d}}\right)$ if $p \in [1, \infty)$ [4, 14]. The next step would be to attempt to extend the results in this paper to higher dimensions.

References

1. Akash Agrawal, Yuan Li, Jie Xue, and Ravi Janardan. The most-likely skyline problem for stochastic points. *Proc. 29th CCCG*, pages 78–83, 2017.
2. Zhi-Dong Bai, Luc Devroye, Hsien-Kuei Hwang, and Tsung-Hsi Tsai. Maxima in hypercubes. *Random Struct. Algorithms*, 27(3):290–309, 2005.
3. I Bárány. The technique of M-regions and cap-coverings: a survey. *Rendiconti di Palermo*, 65:21–38, 2000.
4. Yuri Baryshnikov. On expected number of maximal points in polytopes. In *Discrete Mathematics and Theoretical Computer Science*, pages 247–258, 2007.
5. Stephan Börzsönyi, Donald Kossmann, and Konrad Stocker. The Skyline Operator. In *Proceedings of the 17th International I.C.D.E.,*, pages 421–430. IEEE Computer Society, 2001.
6. Christian Buchta. On the average number of maxima in a set of vectors. *Information Processing Letters*, 33:63–65, 1989.
7. Wei-Mei Chen, Hsien-Kuei Hwang, and Tsung-Hsi Tsai. Maxima-finding algorithms for multi-dimensional samples: A two-phase approach. *Comput. Geometry: Theory and Applications*, 45(1-2):33–53, 2012.
8. Valentina Damerow. *Average and smoothed complexity of geometric structures.* PhD thesis, University of Paderborn, Germany, 2006.
9. Valentina Damerow and Christian Sohler. Extreme Points Under Random Noise. In *Algorithms – ESA 2004*, pages 264–274, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.
10. Olivier Devillers, Marc Glisse, Xavier Goaoc, and Rémy Thomasse. Smoothed complexity of convex hulls by witnesses and collectors. *Journal of Computational Geometry*, 7(2):101–144, 2016.
11. Luc Devroye. *Lecture notes on bucket algorithms.* Birkhauser Boston, 1986.
12. Luc Devroye. Records, the maximal layer, and uniform distributions in monotone sets. *Computers Math. Applic.*, 25(5):19–31, 1993.
13. Josep Diaz and Mordecai Golin. Smoothed Analysis of the Expected Number of Maximal Points in Two Dimensions. *arXiv preprint*, 2018. `arXiv:1807.06845`.

14 R A Dwyer. Kinder, gentler average-case analysis for convex hulls and maximal vectors. *SIGACT News*, 21(2):64–71, 1990.

15 Marc Geilen, Twan Basten, Bart Theelen, and Ralph Otten. An algebra of Pareto points. *Fundamenta Informaticae*, 78(1):35–74, 2007.

16 V. M. Ivanin. Asymptotic estimate for the mathematical expectation of the number of elements in the Pareto set. *Cybernetics*, 11(1):108–113, 1975.

17 J.L.Bentley, H.T. Kung, M. Schkolnick, and C.D. Thompson. On the average number of maxima in a set of vectors and its applications. *Jour. ACM*, 25(4):536–543, 1978.

18 H. T. Kung, Fabrizio Luccio, and Franco P. Preparata. On Finding the Maxima of a Set of Vectors. *J. ACM*, 22(4):469–476, 1975.

19 Alfréd Rényi and Rolf Sulanke. Über die konvexe hülle von n zufällig gewählten punkten. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 2(1):75–84, 1963.

20 Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

21 Daniel A Spielman and Shang-Hua Teng. Smoothed analysis: an attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, 52(10):76–84, 2009.

22 Subhash Suri, Kevin Verbeek, and Hakan Yildiz. On the most likely convex hull of uncertain points. In *European Symposium on Algorithms*, pages 791–802. Springer, 2013.