On sequences of convex records in the plane

Claude Godrèche, Jean-Marc Luck

On sequences of convex records in the plane

Claude Godrèche and Jean-Marc Luck

Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique,
91191 Gif-sur-Yvette, France

Abstract. Convex records have an appealing purely geometric definition. In a sequence of d-dimensional data points, the n-th point is a convex record if it lies outside the convex hull of all preceding points. We specifically focus on the bivariate (i.e., two-dimensional) setting. For iid (independent and identically distributed) points, we establish an identity relating the mean number $\langle R_n \rangle$ of convex records up to time n to the mean number $\langle N_n \rangle$ of vertices in the convex hull of the first n points. By combining this identity with extensive numerical simulations, we provide a comprehensive overview of the statistics of convex records for various examples of iid data points in the plane: uniform points in the square and in the disk, Gaussian points and points with an isotropic power-law distribution. In all these cases, the mean values and variances of N_n and R_n grow proportionally to each other, resulting in finite limit Fano factors F_N and F_R. We also consider planar random walks, i.e., sequences of points with iid increments. For both the Pearson walk in the continuum and the Pólya walk on a lattice, we characterise the growth of the mean number $\langle R_n \rangle$ of convex records and demonstrate that the ratio $R_n/\langle R_n \rangle$ keeps fluctuating with a universal limit distribution.

E-mail: claude.godreche@ipht.fr,jean-marc.luck@ipht.fr

1. Introduction

The statistics of rare and extreme events are of great importance across various scientific disciplines. In particular, the study of the statistics of records in discrete time series has widespread relevance in fields such as climate studies, finance and economics, hydrology, sports, and complex physical systems‡. Most of these studies deal with *univariate* records. The broad scope of the present work is to revisit the subject of *multivariate* records.

Consider an infinite sequence of data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \ldots$ in d-dimensional space. The label n will be referred to as a discrete time. Loosely speaking, $\boldsymbol{x}_n$ is a record whenever it is, in some sense, larger than all previous data points. We then say that there is a record-breaking event, or a record for short, at time n.

In the one-dimensional or univariate setting, the data x_n consists of real numbers. There is a natural ordering of points on the line, and therefore a natural definition of (upper) records. There is a record at time n if

$$x_n > \max(x_1, \ldots, x_{n-1}). \tag{1.1}$$

This canonical definition of univariate records is invariant under the action of a large group of reparametrisation transformations. Records are indeed left unchanged if the data x_n are transformed into $y_n = y(x_n)$, where $y(x)$ is any continuous increasing function. The theory of univariate records, initiated by Chandler [2] and Rényi [3, 4], has become a mature subject [5, 6, 7, 8, 9]. Most classic results concern the case where the x_n are iid (independent and identically distributed) continuous random variables. The key property of records for iid variables is that there is a record at time n with probability

$$Q_n = \frac{1}{n}, \tag{1.2}$$

independently of other occurrences of records. This result holds irrespective of the underlying distribution of the random variables x_n, provided the latter is continuous. The resulting distribution of the number of records R_n up to time n has been long known (see Appendix A). Another well-studied case is when the data points are the successive positions of a one-dimensional random walker. Now, the data points are no longer iid, but their increments are. The basic knowledge on records for random walks can be found in [10]. This problem has later been revisited in the physics literature (for a review, see [1] and references therein).

In the multivariate setting, the data points $\boldsymbol{x}_n$ are d-dimensional vectors. At variance with the one-dimensional situation, there is no natural total ordering in d-dimensional space. This observation was made long ago [11, 12] and led to a variety of definitions of multivariate records [13, 14] (see also [6, 15, 16, 17, 18, 19]). One of these definitions stands out for its minimalistic beauty—the notion of convex records, defined as follows. There is a record at time n if $\boldsymbol{x}_n$ does not belong to the convex hull $\boldsymbol{C}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1})$ of all previous data points, i.e., the smallest closed convex set containing these points. Convex records have attracted very little attention so far. The phrase 'convex records' with this meaning seems to appear only once in the literature [20]. Introductions to the convex geometry of random sets can be found in [21, 22]. For a review on convex hulls in the physics literature, see [23].

‡ An extensive list of references on applications of records to the fields mentioned above can be found in [1].

The purpose of the present work is to investigate the statistics of these convex records. Henceforth we focus our attention on the planar case, where data points $\boldsymbol{x}_n$ are two-dimensional vectors representing points in the plane, and the convex hull of the first n points is a convex polygon. We consider two different settings, namely iid data points in section 2 and random walks in section 3. The results to be described below make no claim to mathematical rigour. They are based on a combination of analytical reasoning, numerical simulations and heuristic scaling analysis.

One of the key geometric properties of the convex hull of a set of points is that it is invariant under the action of the affine group. In the bivariate setting of the present work, if the data points $\boldsymbol{x}_n$ are changed into $\boldsymbol{y}_n = \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}$, where $\boldsymbol{A}$ is a constant invertible matrix (i.e., $\det \boldsymbol{A} \neq 0$) and $\boldsymbol{b}$ a constant vector, their convex hull is changed by the same affine transformation. This symmetry is less constraining than the reparametrisation invariance of the univariate case. It nevertheless gives rise to interesting consequences. For instance, the statistics of convex records for uniform iid data points in the unit square (resp. in the unit disk) is identical to that of uniform points in any parallelogram (resp. in any ellipse).

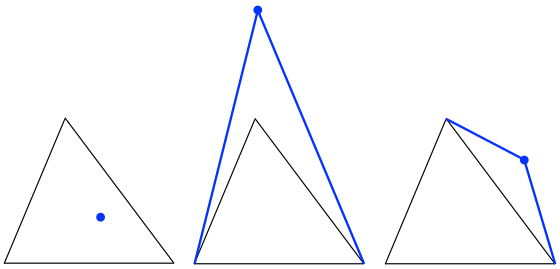

Figure 1. The three possible cases of convex records for $n = 4$ (see (1.4)). The fourth data point and the attached edges of the convex hull, if any, are shown in colour. Left to right: ($N_4 = 3$, $R_4 = 3$), ($N_4 = 3$, $R_4 = 4$), ($N_4 = 4$, $R_4 = 4$).

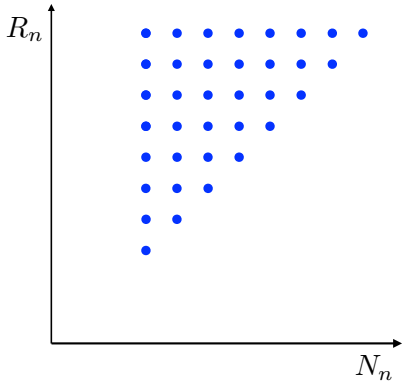

Figure 2. The 36 possible values taken by the couple (N_n, R_n) for $n = 10$.

Let us introduce a few notations. For a given sequence of two-dimensional data points $\boldsymbol{x}_n$, we denote by N_n the number of vertices of the convex hull of the first n

points, and by R_n the number of convex records up to time n. In the generic situation where the first three points are not aligned, each of them is a convex record. We have therefore

$$N_1 = R_1 = 1, \quad N_2 = R_2 = 2, \quad N_3 = R_3 = 3. \tag{1.3}$$

The construction begins to be non-trivial for $n = 4$, resulting in three possibilities (see figure 1):

$$(N_4 = 3,\, R_4 = 3), \quad (N_4 = 3,\, R_4 = 4), \quad (N_4 = 4,\, R_4 = 4). \tag{1.4}$$

More generally, N_n and R_n may take any values in the range (see figure 2)

$$3 \leq N_n \leq R_n \leq n. \tag{1.5}$$

There are therefore $(n-1)(n-2)/2$ possible couples (N_n, R_n).

The setup of this paper is as follows. Section 2 is devoted to sequences of iid data points. We consider four characteristic examples, namely uniform points in the square (section 2.2), uniform points in the disk (section 2.3), Gaussian points (section 2.4), and points with an isotropic power-law distribution (section 2.5). Section 2.6 presents a summary on the mean values and variances of N_n and R_n, whereas the extremal probabilities that N_n and R_n take their smallest or largest values are considered in section 2.7. In section 3 we investigate convex records of planar random walks. The Pearson walk in the continuum and the Pólya walk on three lattices are dealt with in parallel. Section 4 contains a brief discussion. We recall the classical theory of univariate records in an appendix.

2. Sequences of iid data points

This section is devoted to the situation where the data points $\boldsymbol{x}_n$ are iid and drawn from some arbitrary distribution in the plane, assumed to be continuous, to prevent any ties.

2.1. General results

Let us denote by I_n the characteristic function of the event that there is a convex record at time n, and by $Q_n = \langle I_n \rangle$ the corresponding record-breaking probability. The random number R_n of records up to time n reads

$$R_n = \sum_{m=1}^{n} I_m. \tag{2.1}$$

We have therefore

$$\langle R_n \rangle = \sum_{m=1}^{n} Q_m, \tag{2.2}$$

and so

$$Q_n = \langle R_n \rangle - \langle R_{n-1} \rangle. \tag{2.3}$$

Furthermore, in the iid setting, data points are exchangeable. Hence Q_n is the probability that any point $\boldsymbol{x}_m$ chosen among the first n ones does not belong to the convex hull of all the other ones. The product nQ_n is therefore the mean number of

points among the first n ones which are not inside the convex hull of all the other ones. These points are precisely the vertices of $\boldsymbol{C}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. This translates to

$$nQ_n = \langle N_n \rangle. \tag{2.4}$$

The univariate result (1.2) is recovered by setting $\langle N_n \rangle = 1$ in the above formula. Eliminating the record-breaking probability Q_n between (2.3) and (2.4), we obtain the following identity:

$$\langle N_n \rangle = n\big(\langle R_n \rangle - \langle R_{n-1} \rangle\big). \tag{2.5}$$

It is worth emphasising that the formulas (2.2) and (2.3) hold in full generality, whereas (2.4) and (2.5) are specific to the case of iid data points. Moreover, (2.5) only relates the mean value of R_n to that of N_n.

The full statistics of the number R_n of records up to time n is by no means simply related to that of the number N_n of vertices at the single time n. For instance, the second moment of R_n reads

$$\langle R_n^2 \rangle = \sum_{l=1}^{n} \sum_{m=1}^{n} P_{lm}, \tag{2.6}$$

where

$$P_{lm} = \langle I_l I_m \rangle \tag{2.7}$$

is the joint probability that there are records at times l and m. We have therefore

$$\operatorname{Var} R_n = \sum_{l=1}^{n} \sum_{m=1}^{n} (P_{lm} - P_l P_m) = \sum_{l=1}^{n} \sum_{m=1}^{n} \langle I_l I_m \rangle_c, \tag{2.8}$$

where the subscript $\langle \rangle_c$ stands for 'connected'. In the case of univariate records, where the events I_m are mutually independent, (2.8) boils down to (A.8). In the case of multivariate convex records, the events I_m are not statistically independent. Numerical results however demonstrate that $\langle R_n \rangle$ and $\operatorname{Var} R_n$ grow proportionally to each other in all examples of iid data points we have considered (see Section 2.6).

The first instance where the identity (2.5) is non-trivial is $n = 4$. This situation corresponds to Sylvester's famous four point problem [24] (see [25] for a historical account). Sylvester was interested in the probability P that a random planar quadrilateral is reentrant, i.e., non-convex. The exchangeability of the data points implies that the Sylvester probability reads

$$P = 4q, \tag{2.9}$$

where q is the probability that a random point $\boldsymbol{x}$ is inside the triangle $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$. This reads formally

$$q = \langle A(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \rangle, \tag{2.10}$$

where $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ is the 'probability content' of the triangle $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$, i.e., the probability for a data point $\boldsymbol{x}$ to be inside that triangle, and the average is taken over the three iid points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$. The probability q depends on the underlying distribution of data points. It has been known since the 19th century for points uniformly distributed in various domains:

$$q(\text{triangle}) = \frac{1}{12} = 0.083333\ldots, \quad q(\text{square}) = \frac{11}{144} = 0.076388\ldots,$$
$$q(\text{disk}) = \frac{35}{48\pi^2} = 0.073880\ldots \tag{2.11}$$

The triangle and the disk respectively yield upper and lower bounds of q for points uniformly distributed in a convex domain of the plane [26]. Larger values of q may however be reached. We have indeed

$$q = 1 - \frac{3}{2\pi} \arccos\left(-\frac{1}{3}\right) = 0.087739\ldots \tag{2.12}$$

for Gaussian points (see (2.14), (2.36)).

Coming back to convex records, the exchangeability of the data points implies that the three cases listed in (1.4) and shown in figure 1 occur with respective probabilities

$$q_1 = q, \qquad q_2 = 3q, \qquad q_3 = 1 - 4q = 1 - P. \tag{2.13}$$

We have therefore

$$\langle N_4 \rangle = 4(1-q) = 4 - P, \qquad \langle R_4 \rangle = 4 - q. \tag{2.14}$$

These mean values obey the identity (2.5), as should be.

In the asymptotic regime of most interest where n is very large, it is legitimate to use a continuum approximation, so that (2.5) becomes§

$$\langle N_n \rangle \approx n \, \frac{\mathrm{d}\langle R_n \rangle}{\mathrm{d}n} \approx \frac{\mathrm{d}\langle R_n \rangle}{\mathrm{d}\ln n}. \tag{2.15}$$

The statistics of the number N_n of vertices of the convex hull of n iid points in d-dimensional space, and especially in the plane, has been the subject of a rather abundant mathematical literature since the pioneering works by Rényi and Sulanke [27] and by Efron [28]. Most available rigorous results concern the mean number $\langle N_n \rangle$ of vertices. As it turns out, the behaviour of this quantity at large n strongly depends on the underlying distribution of the data points. The identity (2.5) enables us to predict in each case the behaviour of the mean number $\langle R_n \rangle$ of records.

Hereafter we consider four characteristic examples of iid data points in the plane, namely uniform points in the square (section 2.2), uniform points in the disk (section 2.3), Gaussian points (section 2.4), and points with an isotropic power-law distribution (section 2.5).

2.2. Uniform points in the square

We begin with a reminder on the more general situation of uniform points in a convex polygon with r sides and vertices $\boldsymbol{R}_i$. The number N_n of vertices of the convex hull of n points has been shown to obey a central limit theorem at large n, i.e., to have an asymptotic normal or Gaussian distribution, whose mean and variance grow logarithmically with n [29].

More precisely, the mean value of N_n reads asymptotically [27]

$$\langle N_n \rangle \approx \frac{2r}{3} \left(\ln n + \frac{1}{r} \sum_{i=1}^{r} \ln \frac{\mathcal{A}_i}{\mathcal{A}} + \gamma \right), \tag{2.16}$$

where γ is Euler's constant, $\mathcal{A}$ is the total area of the polygon, and $\mathcal{A}_i$ is the area of the triangle $\boldsymbol{R}_{i-1}\boldsymbol{R}_i\boldsymbol{R}_{i+1}$ formed by three consecutive vertices. The area ratios entering

§ Here and throughout the following, $x \approx y$ means that x and y are asymptotically equivalent in the appropriate regime (here, $n \gg 1$), in the strong sense that y/x converges to unity, whereas the weaker form $x \sim y$ means that y/x has much slower variations than x or y taken separately.

the above formula are affine invariants. For the regular r-gon of the Euclidean plane, this reads

$$\langle N_n \rangle \approx \frac{2r}{3} \left(\ln \left(\frac{4n}{r} \sin^2 \frac{\pi}{r} \right) + \gamma \right). \tag{2.17}$$

When the number r of sides becomes itself large, the above result simplifies to

$$\langle N_n \rangle \approx \frac{2r}{3} \left(\ln \frac{4\pi^2 n}{r^3} + \gamma \right). \tag{2.18}$$

This formula exhibits a crossover for $n \sim r^3$, with $\langle N_n \rangle$ scaling as (2.18) for $1 \ll r^3 \ll n$, and as (2.26) for $1 \ll n \ll r^3$.

The variance of N_n grows as [29]

$$\mathrm{Var}\, N_n \approx \frac{10r}{27} \ln n, \tag{2.19}$$

up to an additive constant whose exact expression is not known.

For uniform points in a polygonal domain, (2.16) and (2.19) show that $\langle N_n \rangle$ and $\mathrm{Var}\, N_n$ share the same logarithmic growth law in n. The ratio

$$F_{N_n} = \frac{\mathrm{Var}\, N_n}{\langle N_n \rangle}, \tag{2.20}$$

known as the Fano factor of the distribution of N_n, goes to the limit

$$F_N = \frac{5}{9}, \tag{2.21}$$

irrespective of the number r of sides of the polygon. The Fano factor [30] is used to characterise distributions of integers counting detected particles and similar discrete events (see e.g. [31]). Poisson distributions have a Fano factor $F = 1$. Distributions with F less than unity (resp. larger than unity) are referred to as sub-Poissonian (resp. super-Poissonian).‖ Higher Fano factors can be defined similarly, in terms of higher cumulants.

We now turn to the specific case of uniform points in the unit square. Combinatorial methods give access to some results for N_n for finite n [33, 34]. When n is large, the asymptotic results (2.17) and (2.19) read

$$\langle N_n \rangle \approx A_1 \ln n, \quad \mathrm{Var}\, N_n \approx A_2 \ln n, \qquad A_1 = \frac{8}{3}, \quad A_2 = \frac{40}{27}. \tag{2.22}$$

The identity (2.5) predicts that the mean value of R_n grows as

$$\langle R_n \rangle \approx B_1 (\ln n)^2, \qquad B_1 = \frac{A_1}{2} = \frac{4}{3}. \tag{2.23}$$

We have performed extensive numerical simulations on the statistics of convex records. For that purpose, we have developed a recursive algorithm constructing the convex hull of a set of points in the plane that are added one by one. As a first illustration, we show in figure 3 the outcome of a simulation of 200 iid data points in the unit square, such that $N = 11$ and $R = 38$. Figure 4 illustrates the evolution of N_n (lower tracks) and R_n (upper tracks) against time n for three typical histories of 1,000 data points. Each history is shown by a colour. The numbers N_n of vertices exhibit non-monotonic fluctuations as a function of n. The numbers R_n of records increase faster than N_n, and monotonically in time, as should be.

‖ The Mandel parameter $Q = F - 1$ is used in other areas of physics, including quantum optics [32].

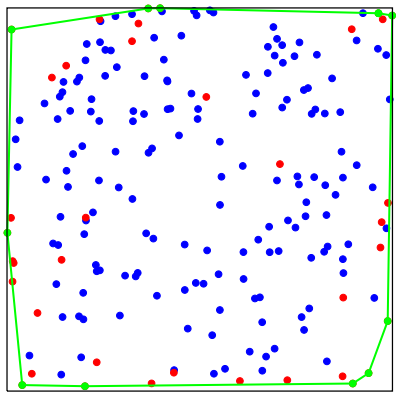

Figure 3. A sample of 200 uniform iid points in the unit square such that $N = 11$ and $R = 38$. Green polygon: convex hull of the dataset. Green symbols: the 11 vertices of the convex hull. Red symbols: the 27 other convex records. Blue symbols: the 162 data points that are not convex records.

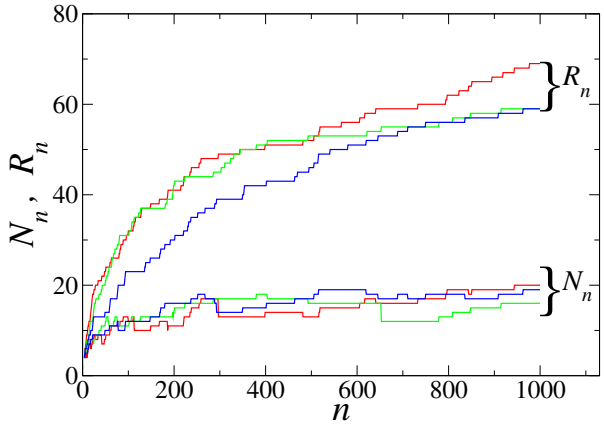

Figure 4. Numbers N_n (lower tracks) and R_n (upper tracks) plotted against time n for three histories of 1,000 uniform data points in the unit square. Each history is shown by a colour.

To come to a quantitative study, we focus our attention on the mean values and variances of N_n and R_n. Figure 5 shows plots of $\langle N_n \rangle$ and $\mathrm{Var}\, N_n$ against $\ln n$. Here and throughout the following, numerical data are gathered over 10^5 independent histories of 10^6 data points each, and only data for $n > 100$ (sometimes $n > 1,000$) are included in the analysis of their asymptotic behaviour. Dashed lines show linear fits to these data, whose respective slopes 2.67 and 1.47 are to be compared with $A_1 = 8/3 = 2.666666\ldots$ and $A_2 = 40/27 = 1.481481\ldots$ (see (2.22)). This very good agreement provides a solid validation of our numerical approach. Figure 6 shows plots of $\langle R_n \rangle$ and $\mathrm{Var}\, R_n$ against $\ln n$. Dashed curves show quadratic fits. The coefficient of $(\ln n)^2$ for $\langle R_n \rangle$ reads 1.33, again in good agreement with the analytical prediction

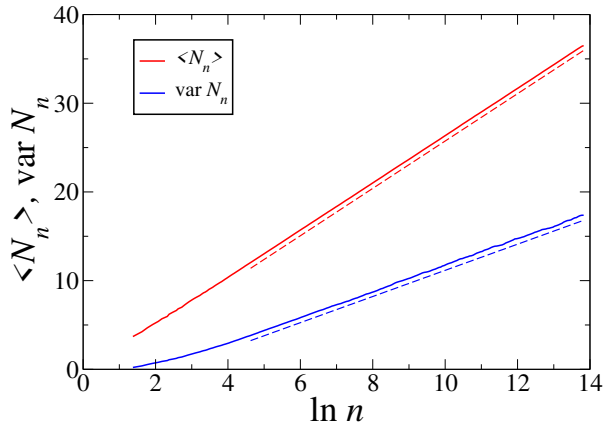

Figure 5. $\langle N_n \rangle$ and $\operatorname{Var} N_n$ plotted against $\ln n$ for uniform points in the unit square. Dashed lines show linear fits, slightly offset for greater clarity.

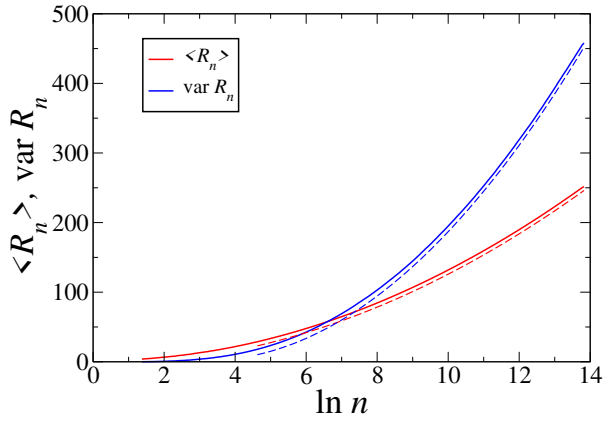

Figure 6. $\langle R_n \rangle$ and $\operatorname{Var} R_n$ plotted against $\ln n$ for uniform points in the unit square. Dashed curves show quadratic fits, slightly offset vertically for greater clarity.

$B_1 = 4/3 = 1.333\ldots$ (see (2.23)). The coefficient of $(\ln n)^2$ for $\operatorname{Var} R_n$ reads 3.96, yielding the growth law

$$\operatorname{Var} R_n \approx B_2 (\ln n)^2, \qquad B_2 \approx 3.96. \tag{2.24}$$

This is the first amplitude for which there is no analytical prediction. Equations (2.23) and (2.24) yield the finite limit Fano factor

$$F_R = \frac{B_2}{B_1} = \frac{3B_2}{4} \approx 2.97. \tag{2.25}$$

2.3. Uniform points in the disk

We now consider uniform iid points in the unit disk. Some combinatorial results on the number N_n of vertices of the convex hull of n points in the disk are available for

finite n [35]. For large n, N_n obeys a central limit theorem, i.e., it has an asymptotic Gaussian distribution [29], whose mean and variance grow as

$$\langle N_n \rangle \approx A_1\, n^{1/3}, \qquad \mathrm{Var}\, N_n \approx A_2\, n^{1/3}. \tag{2.26}$$

These power laws are common to all finite convex domains of the plane with a smooth boundary. In the case of the disk, the prefactors A_1 and A_2 are known exactly. They read [27]

$$A_1 = \left(\frac{2^7\pi^2}{3^4}\right)^{1/3} \Gamma(2/3) = 3.383228\ldots \tag{2.27}$$

and [36]

$$A_2 = \left(\frac{2^7\pi^2}{3^{13}}\right)^{1/3} \left(\frac{16\pi^2}{\Gamma(2/3)^2} - 57\,\Gamma(2/3)\right) = 0.826885\ldots \tag{2.28}$$

The corresponding limit Fano factor is therefore

$$F_N = \frac{A_2}{A_1} = \frac{16\pi^2}{27\,\Gamma(2/3)^3} - \frac{19}{9} = 0.244407\ldots \tag{2.29}$$

The identity (2.5) predicts that the mean value of R_n grows as

$$\langle R_n \rangle \approx B_1\, n^{1/3}, \qquad B_1 = 3A_1 = 10.149686\ldots \tag{2.30}$$

We have again measured the mean values and variances of N_n and R_n by extensive numerical simulations. Our data concerning N_n (not shown) are in very good agreement with (2.27), (2.28). Figure 7 shows plots of $\langle R_n \rangle$ and $\mathrm{Var}\, R_n$ against $n^{1/3}$. Dashed lines show linear fits to the data. The slope for $\langle R_n \rangle$ reads 10.15, in excellent agreement with the analytical prediction (2.30). The slope for $\mathrm{Var}\, R_n$ reads 8.73, implying the growth law

$$\mathrm{Var}\, R_n \approx B_2\, n^{1/3}, \qquad B_2 \approx 8.7. \tag{2.31}$$

Equations (2.30) and (2.31) yield the limit Fano factor

$$F_R = \frac{B_2}{B_1} \approx 0.86. \tag{2.32}$$

2.4. Gaussian points

This section is devoted to iid Gaussian (or normal) points. The invariance under the affine group can be used to map an arbitrary Gaussian distribution on a centred isotropic one such that $\langle |\boldsymbol{x}|^2 \rangle = 1$.

Many papers in the mathematical literature deal with the convex hull of two-dimensional Gaussian points [27, 28, 37, 38, 39]. Exact formulas for the mean values of several quantities for finite n have been derived in [28]. The expression for the mean number N_n of vertices reads

$$\langle N_n \rangle = \sqrt{4\pi}\, J_n, \tag{2.33}$$

with

$$J_n = n(n-1) \int_{-\infty}^{\infty} F(x)^{n-2} f(x)^2\, \mathrm{d}x \qquad (n \geq 3), \tag{2.34}$$

where

$$f(x) = \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}, \qquad F(x) = \frac{1}{2}\left(1 + \mathrm{erf}\,\frac{x}{\sqrt{2}}\right) \tag{2.35}$$

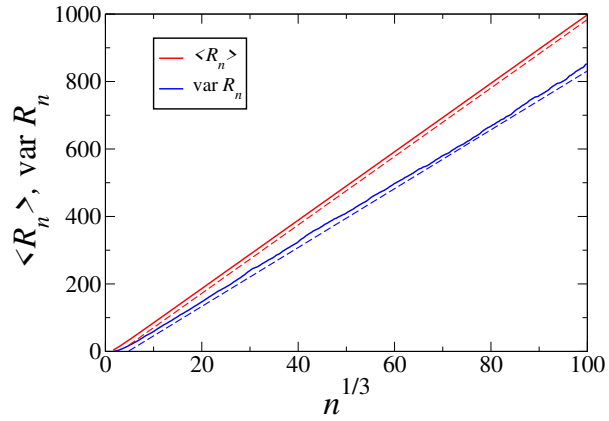

Figure 7. $\langle R_n \rangle$ and $\mathrm{Var}\, R_n$ plotted against $n^{1/3}$ for uniform points in the unit disk. Dashed lines show linear fits, slightly offset for greater clarity.

are the density and the distribution function of a Gaussian variable such that $\langle x^2 \rangle = 1$. The integrals J_n have been investigated in detail in [40], where they are denoted by A_n. Besides (1.3), the expressions

$$\langle N_4 \rangle = \frac{6}{\pi} \arccos\left(-\frac{1}{3}\right) = 3.649040\ldots, \tag{2.36}$$

$$\langle N_5 \rangle = \frac{5}{2\pi} \arccos\left(-\frac{23}{27}\right) = 4.122601\ldots \tag{2.37}$$

seem to exhaust the list of available closed-form results. The asymptotic behaviour of (2.33) reads

$$\langle N_n \rangle = (8\pi \ln n)^{1/2} \left(1 + \frac{\mu}{2\ln n} - \frac{\mu^2 + 2\mu + 2 + \pi^2/6}{8(\ln n)^2} + \cdots\right), \tag{2.38}$$

with

$$\mu = \gamma - \frac{1}{2} \ln(4\pi \ln n), \tag{2.39}$$

where γ is Euler's constant.

The $(\ln n)^{1/2}$ growth law (2.38) is rather ubiquitous among data points with isotropic distributions in the plane whose complementary radial distribution function

$$\mathbb{P}(|\boldsymbol{x}| > r) = F(r) = \exp(-\Phi(r)) \tag{2.40}$$

decays rapidly as $r \to \infty$. This situation has been investigated long ago by Carnal [37], resulting in the following parametric representation of the mean number $\langle N_n \rangle$ of vertices:

$$\ln n \approx \Phi(r), \qquad \langle N_n \rangle \approx \left(4\pi r \Phi'(r)\right)^{1/2}, \tag{2.41}$$

where the accent denotes a derivative. The parameter r has a simple interpretation: in line with the theory of extreme-value statistics, r provides an estimate for the largest radius of the first n data points. For distributions falling off as a stretched or compressed exponential of the form

$$F(r) \sim \exp(-Ar^\alpha), \tag{2.42}$$

(2.41) predicts

$$\langle N_n \rangle \approx (4\pi\alpha \ln n)^{1/2}, \tag{2.43}$$

where the exponent α only enters the prefactor of the $(\ln n)^{1/2}$ growth law. In the limit situation of distributions decaying as a power law of the form

$$F(r) \approx \frac{c}{r^\theta}, \tag{2.44}$$

(2.41) predicts that the mean number of vertices saturates to the finite value

$$\langle N_n \rangle \approx \sqrt{4\pi\theta}. \tag{2.45}$$

This is indeed the correct leading-order result for power-law data points with a large exponent θ (see (2.57)).

Coming back to Gaussian data points, the random number N_n of vertices obeys a central limit theorem at large n, i.e., it has an asymptotic Gaussian distribution [38, 39], whose mean and variance grow as

$$\langle N_n \rangle \approx A_1 (\ln n)^{1/2}, \qquad \mathrm{Var}\, N_n \approx A_2 (\ln n)^{1/2}. \tag{2.46}$$

The prefactor (see (2.38))

$$A_1 = \sqrt{8\pi} = 5.013256\ldots \tag{2.47}$$

has been long known [27]. No formula for A_2 seems to be known to date (see [38]).

The identity (2.5) predicts that the mean value of R_n grows as

$$\langle R_n \rangle \approx B_1 (\ln n)^{3/2}, \qquad B_1 = \frac{2A_1}{3} = \frac{4\sqrt{2\pi}}{3} = 3.342171\ldots \tag{2.48}$$

In analogy with previous cases, we anticipate that the variance of R_n also scales as

$$\mathrm{Var}\, R_n \approx B_2 (\ln n)^{3/2}. \tag{2.49}$$

We have measured the mean values and variances of N_n and R_n by extensive numerical simulations. Figure 8 shows plots of $\langle N_n \rangle$ and $\mathrm{Var}\, N_n$ against $\ln n$. Dashed curves show non-linear fits of the form $y = a(\ln n + b)^{1/2}$. The complexity of the sequence of subleading corrections entering (2.38) has deterred us from using more sophisticated fits. In the case of $\langle N_n \rangle$, the fit parameter $a = 4.911$, to be identified with A_1, is in good agreement with the prediction (2.47) (2 percent relative difference), especially in view of the above. In the case of $\mathrm{Var}\, N_n$, the fit parameter $a = 1.865$ yields

$$A_2 \approx 1.86, \tag{2.50}$$

with an expected relative accuracy in the range of a few percent. Figure 9 shows plots of $\langle R_n \rangle$ and $\mathrm{Var}\, R_n$ against $\ln n$. Dashed curves show non-linear fits of the form $y = a(\ln n + b)^{3/2}$. In the case of $\langle R_n \rangle$, the parameter $a = 3.17$, to be identified with B_1, is in reasonably good agreement with the prediction (2.48) (5 percent relative difference). In the case of $\mathrm{Var}\, R_n$, the parameter $a = 7.03$ yields

$$B_2 \approx 7.0, \tag{2.51}$$

again with an expected relative accuracy of a few percent.

The above results translate to the limit Fano factors

$$F_N = \frac{A_2}{A_1} \approx 0.37, \qquad F_R = \frac{B_2}{B_1} \approx 2.10. \tag{2.52}$$

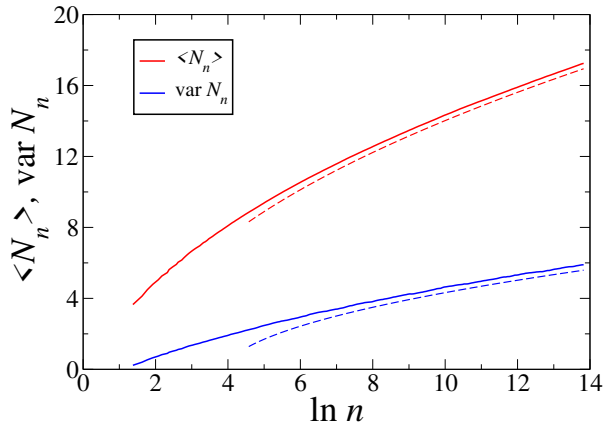

Figure 8. $\langle N_n \rangle$ and $\operatorname{Var} N_n$ plotted against $\ln n$ for Gaussian points. Dashed curves show the non-linear fits described in the text, slightly offset vertically for greater clarity.

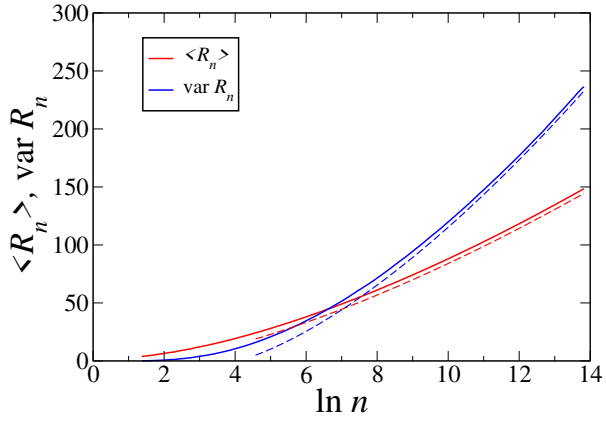

Figure 9. $\langle R_n \rangle$ and $\operatorname{Var} R_n$ plotted against $\ln n$ for Gaussian points. Dashed curves show the non-linear fits described in the text, slightly offset vertically for greater clarity.

2.5. Points with isotropic power-law distributions

Our last example concerns iid data points $\boldsymbol{x}_n$ with an isotropic distribution whose complementary radial distribution function (see (2.40)) falls off as a power law at large distances, namely

$$F(r) \approx \frac{c}{r^\theta}, \tag{2.53}$$

with an arbitrary exponent $\theta > 0$.

A few works in the mathematical literature deal with the convex hull of such iid random points [37, 39, 41]. At variance with previous examples, the distribution of the number N_n of its vertices now reaches a finite limit,

$$p_k(\theta) = \lim_{n\to\infty} \mathbb{P}(N_n = k) \qquad (k \geq 3), \tag{2.54}$$

as the number n of points becomes infinitely large. The above limit distribution is universal, in the sense that it only depends on the exponent θ. For generic values of θ, only the first moment

$$A_1(\theta) = \lim_{n\to\infty} \langle N_n \rangle = \sum_{k\geq 3} k p_k(\theta) \tag{2.55}$$

is known explicitly and reads

$$A_1(\theta) = 4\sqrt{\pi}\, \frac{\Gamma^2(\frac{1}{2}\theta+1)\,\Gamma(\theta+\frac{1}{2})}{\Gamma^2(\frac{1}{2}\theta+\frac{1}{2})\,\Gamma(\theta+1)}. \tag{2.56}$$

This expression starts from $A_1(0) = 4$ (see (2.61)), and grows at large θ as

$$A_1(\theta) = (4\pi\theta)^{1/2}\left(1 + \frac{3}{8\theta} + \frac{9}{128\theta^2} + \cdots\right). \tag{2.57}$$

We introduce for further reference the notation for the corresponding variance:

$$A_2(\theta) = \lim_{n\to\infty} \operatorname{Var} N_n = \sum_{k\geq 3} k^2 p_k(\theta) - A_1(\theta)^2. \tag{2.58}$$

We mention for completeness that the problem simplifies in the $\theta \to 0$ limit [41], where the full distribution $p_k(0)$ has been obtained explicitly:

$$p_k(0) = 2^{k-3}\left(2\,\frac{(\ln 2)^{k-2}}{(k-2)!} - 2 + \sum_{j=0}^{k-3} \frac{(\ln 2)^j}{j!}\right). \tag{2.59}$$

The corresponding generating function,

$$G(z) = \lim_{n\to\infty} \langle z^{N_n} \rangle = \sum_{k\geq 3} p_k(0) z^k = \frac{z^2}{1-2z}\,((1-z)2^{2z} - 1), \tag{2.60}$$

yields in particular

$$A_1(0) = 4, \qquad A_2(0) = 16\ln 2 - 10 = 1.090354\ldots \tag{2.61}$$

The identity (2.5) predicts that the mean value of R_n grows as

$$\langle R_n \rangle \approx B_1(\theta)\, \ln n, \qquad B_1(\theta) = A_1(\theta). \tag{2.62}$$

In analogy with previous cases, we anticipate that the variance of R_n scales as

$$\operatorname{Var} R_n \approx B_2(\theta)\, \ln n. \tag{2.63}$$

We have run extensive numerical simulations for exponents θ ranging from $1/2$ to 10. Data points with isotropic power-law distributions were generated by using the non-linear mapping

$$\boldsymbol{x} = \frac{\boldsymbol{y}}{(1-|\boldsymbol{y}|)^{1/\theta}}. \tag{2.64}$$

If $\boldsymbol{y}$ is uniformly distributed in the unit disk, the distribution of $\boldsymbol{x}$ is isotropic and obeys (2.53) with $c = 2$ and an arbitrary exponent $\theta > 0$. We have again measured the mean values and variances of N_n and R_n. Our data corroborate the above picture. The measured $\langle N_n \rangle$ and $\operatorname{Var} N_n$ go to well-defined limits $A_1(\theta)$ and $A_2(\theta)$, shown in figure 10. The observed values of $A_1(\theta)$ are in very good agreement with the analytical result (2.56). The measured $\langle R_n \rangle$ and $\operatorname{Var} R_n$ are found to follow the logarithmic growth laws (2.62) and (2.63). The numerical values of $B_1(\theta)$ are in very good agreement with the prediction (2.62). The values of $B_2(\theta)$ are also shown

in figure 10. This figure strongly suggests that the three plotted quantities share the same square-root law at large θ (see (2.57)). This observation is corroborated by figure 11, showing the Fano factors

$$F_N(\theta) = \frac{A_2(\theta)}{A_1(\theta)}, \qquad F_R(\theta) = \frac{B_2(\theta)}{B_1(\theta)}. \tag{2.65}$$

These quantities are found to converge to the limits

$$F_N(\infty) \approx 0.40, \qquad F_R(\infty) \approx 2.20. \tag{2.66}$$

These values are rather close to those corresponding to Gaussian points (see (2.52)).

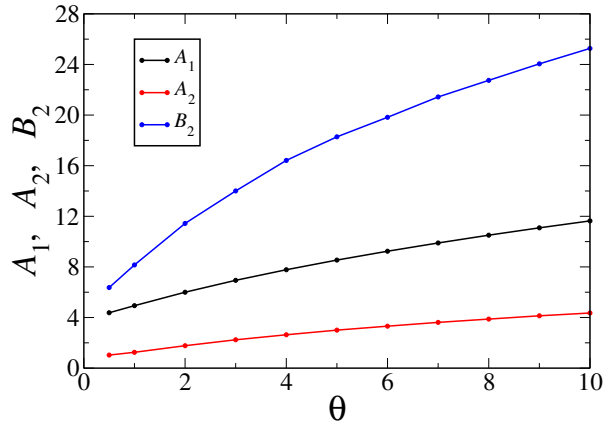

Figure 10. Amplitudes $A_1(\theta)$, $A_2(\theta)$ and $B_2(\theta)$ characterising convex hulls and convex records of data points with isotropic power-law distributions, plotted against the exponent θ. Black: $A_1(\theta)$ entering (2.55) and (2.62), and given by (2.56). Red: numerical values of $A_2(\theta)$ entering (2.58). Blue: numerical values of $B_2(\theta)$ entering (2.63).

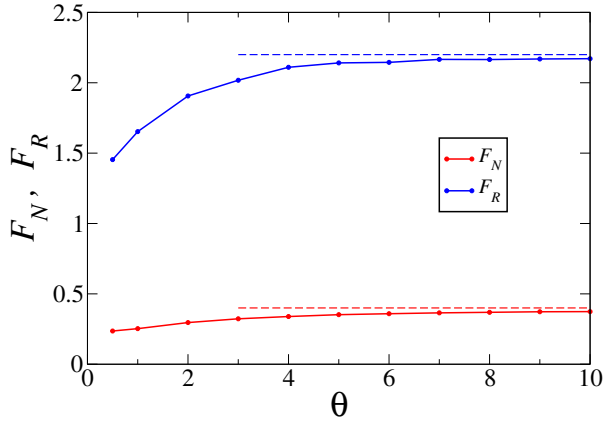

Figure 11. Fano factors defined in (2.65), plotted against the exponent θ. Red: $F_N(\theta)$. Blue: $F_R(\theta)$. Horizontal dashed lines: estimated limits (2.66).

Distribution	section	$\langle N_n \rangle$	$\mathrm{Var}\, N_n$	$\langle R_n \rangle$	$\mathrm{Var}\, R_n$
uniform (square)	(2.2)	$A_1^\star \ln n$	$A_2^\star \ln n$	$B_1^\star (\ln n)^2$	$B_2 (\ln n)^2$
uniform (disk)	(2.3)	$A_1^\star\, n^{1/3}$	$A_2^\star\, n^{1/3}$	$B_1^\star\, n^{1/3}$	$B_2\, n^{1/3}$
Gaussian	(2.4)	$A_1^\star\, (\ln n)^{1/2}$	$A_2\, (\ln n)^{1/2}$	$B_1^\star\, (\ln n)^{3/2}$	$B_2\, (\ln n)^{3/2}$
power-law	(2.5)	$A_1^\star(\theta)$	$A_2(\theta)$	$B_1^\star(\theta) \ln n$	$B_2(\theta) \ln n$

Table 1. Asymptotic behaviour of the mean values and variances of N_n and R_n for the four examples of iid data points studied in this work. Amplitudes with a star in superscript are known analytically. All other amplitudes are determined numerically.

2.6. A summary on mean values and variances

Our investigation of convex records for four characteristic examples of iid points in the plane (uniform points in the square and in the disk, isotropic Gaussian and power-law points) allows us to draw the following conclusions. Consider first the number N_n of vertices of the convex hull of the first n points. In the first three examples recalled above, $\langle N_n \rangle$ and $\mathrm{Var}\, N_n$ grow proportionally to each other, resulting in a finite limit Fano factor F_N. For these examples, and many other cases studied in the mathematical literature, N_n is known to obey a central limit theorem, i.e., to have an asymptotic normal or Gaussian distribution, with mean values and variances growing at the same rate. Note however that a counterexample 'whose support is quite a complicated geometric object' has been constructed [42], for which the relative fluctuations of N_n around $\langle N_n \rangle$ do not shrink to zero as n is very large. Extensive numerical simulations have demonstrated that the number R_n of convex records behaves quite similarly, in that $\langle R_n \rangle$ and $\mathrm{Var}\, R_n$ also grow proportionally to each other, also resulting in a finite limit Fano factor F_R. Our numerical results make it very plausible that R_n asymptotically obeys a central limit theorem, and that higher cumulants of N_n and R_n grow at the same rate as their mean values, resulting in non-trivial higher limit Fano factors $F_{N,k} = \langle N_n^k \rangle_c / \langle N_n \rangle$ and $F_{R,k} = \langle R_n^k \rangle_c / \langle R_n \rangle$.

Table 1 summarises the asymptotic behaviour of the mean values and variances of N_n and R_n. Amplitudes whose analytical expression was known exactly are marked by stars. Figure 12 shows a scatter plot of all Fano factors thus obtained. Among these numbers, only the first two values of F_N are known exactly (see (2.21), (2.29)). In all cases, F_N is smaller than unity, so that the distribution of N_n is asymptotically sub-Poissonian. In all cases but the disk, F_R is larger than unity, so that the distribution of R_n is super-Poissonian. The case of uniform points in the disk is somehow an outlier in two respects: $\langle N_n \rangle$ and $\langle R_n \rangle$ grow as a power law, and the distribution of R_n is sub-Poissonian. We recall that the distribution of R_n is asymptotically Poissonian in the classical case of univariate records (see Appendix A).

2.7. Extremal probabilities

The full distributions of N_n and R_n have many features of potential interest, besides their mean values and variances investigated so far. Hereafter we focus our attention on the extremal values of these random numbers, namely 3 and n (see (1.5) and figure 2).

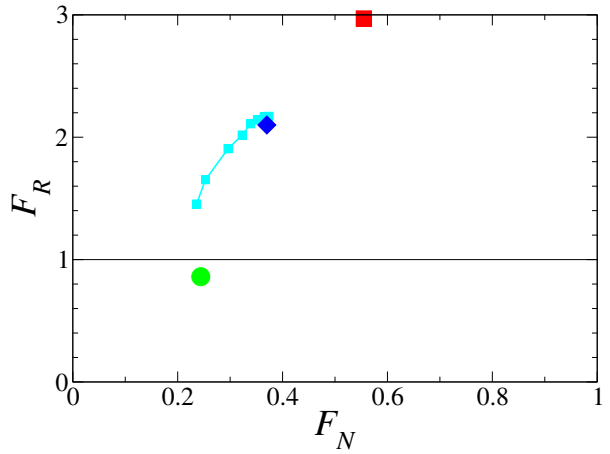

Figure 12. Scatter plot of the limit Fano factors F_N and F_R obtained for the four examples of iid points in the plane considered in this work. Red square: uniform points in the square (section 2.2). Green disk: uniform points in the disk (section 2.3). Blue diamond: Gaussian points (section 2.4). Cyan symbols joined by a line: Isotropic power-law points (section 2.5).

The probability that the number of records takes its minimal value $R_n = 3$ reads [20]

$$\mathbb{P}(R_n = 3) = \langle A(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)^{n-3} \rangle \qquad (n > 3), \tag{2.67}$$

with the notation used in (2.10). The meaning of this general result is clear: $R_n = 3$ holds for datasets of n points where all subsequent $n - 3$ points fall inside the triangle formed by the first three ones. Along this line of thought, using again the exchangeability of the data points, $N_n = 3$ corresponds to datasets of n points where the remaining $n-3$ points fall inside the triangle formed by any three different points. This observation translates to the identity

$$\mathbb{P}(N_n = 3) = \binom{n}{3} \mathbb{P}(R_n = 3). \tag{2.68}$$

The event $R_n = 3$ corresponds to the bottom left-hand corner of figure 2, whereas $N_n = 3$ corresponds to the entire leftmost column. The corresponding extremal probabilities only differ by a simple combinatorial factor. Equations (2.5) and (2.68) are the only two general identities we have found for convex records of bivariate iid data.

The asymptotic behaviour of the extremal probability $\mathbb{P}(N_n = 3)$ depends on the underlying distribution of points. For isotropic power-law points, $\mathbb{P}(N_n = 3)$ goes to the universal limit $p_3(\theta)$. For uniform points inside a convex domain, this probability generically falls off exponentially fast, as

$$\mathbb{P}(N_n = 3) \sim A_\star^n, \tag{2.69}$$

where $A_\star$ is the 'probability content' of the largest triangle inscribed in the domain, i.e., the fraction of the total area enclosed by this largest triangle. We have

$$A_\star(\text{triangle}) = 1, \qquad A_\star(\text{square}) = \frac{1}{2},$$

$$A_\star(\text{disk}) = \frac{3\sqrt{3}}{4\pi} = 0.413496\ldots \tag{2.70}$$

In the case of a triangular domain, $A_\star = 1$ prevents the exponential decay of (2.69). The extremal probabilities obey the power law $\mathbb{P}(N_n = 3) \approx 8/n^3$ [43, 44], and so $\mathbb{P}(R_n = 3) \approx 48/n^6$.

The probability that the number of vertices takes its maximal value $N_n = n$ has been investigated for uniform points in several domains. Exact combinatorial results are available for the triangle and the square [45, 46], whereas a full asymptotic analysis has been performed in the case of the disk [47]. This extremal probability is found to fall off super-exponentially, as

$$\mathbb{P}(N_n = n) \sim \frac{B^n}{(n!)^2}, \tag{2.71}$$

with

$$\begin{aligned} &B(\text{triangle}) = \frac{27}{2} = 13.5, \qquad B(\text{square}) = 16, \\ &B(\text{disk}) = 2\pi^2 = 19.739208\ldots \end{aligned} \tag{2.72}$$

A general expression for B for an arbitrary convex domain is known [48] (see also [49]). It is highly likely that the exponential law (2.69) and the $1/(n!)^2$ law (2.71) hold for a much larger class of distributions of iid data points.

The probability $\mathbb{P}(R_n = n)$ that the number of records takes its maximal value $R_n = n$ is expected to fall off less rapidly than $\mathbb{P}(N_n = n)$. The event $N_n = n$ indeed corresponds to the top right-hand corner of figure 2, whereas $R_n = n$ corresponds to the entire top row. We have run numerical simulations to measure $\mathbb{P}(R_n = n)$ for the above four characteristic examples of iid points in the plane. Figure 13 shows logarithmic plots of the product $n!\,\mathbb{P}(R_n = n)$ against time n. The fits to the data (see caption) convincingly suggest the behaviour

$$\mathbb{P}(R_n = n) \sim \frac{C^n}{n!}, \tag{2.73}$$

where the constant C has a weak dependence on the underlying distribution of points:

$$\begin{aligned} &C(\text{square}) \approx 9.7, \qquad C(\text{disk}) \approx 10.2, \\ &C(\text{Gaussian}) \approx 8.5, \qquad C(\text{power-law}, \theta = 2) \approx 7.5. \end{aligned} \tag{2.74}$$

A rationale for the $1/n!$ decay of $\mathbb{P}(R_n = n)$ is that the event $R_n = n$ corresponds to histories where every new data point is outside the convex hull of all previous ones. Loosely speaking, data point are further and further away from the origin. In this regard, they resemble univariate records, for which $\mathbb{P}(R_n = n) = 1/n!$ exactly (see (A.10)).

3. Random walks

We now consider convex records for data points generated by planar random walks. The points $\boldsymbol{x}_n$ are the successive positions of a random walker launched at the origin:

$$\boldsymbol{x}_n = \boldsymbol{x}_{n-1} + \boldsymbol{\delta}_n \qquad (\boldsymbol{x}_0 = \boldsymbol{0}). \tag{3.1}$$

The increments $\boldsymbol{\delta}_n$ are iid two-dimensional random vectors, drawn from some fixed distribution. After n time steps, the dataset consists of $n+1$ points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$.

In this work we consider in parallel four examples of planar random walks: the Pearson walk [50], where the increments $\boldsymbol{\delta}_n$ are uniformly distributed over the unit

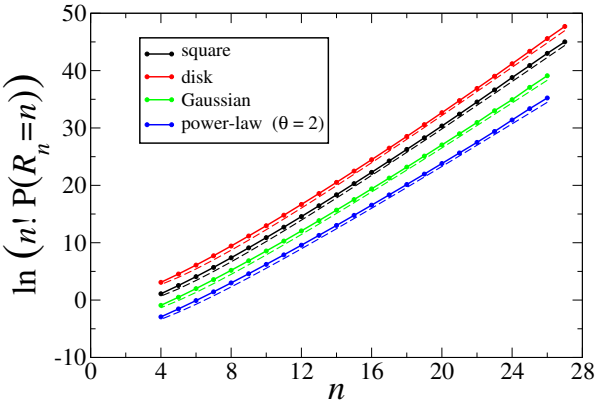

Figure 13. Logarithmic plot of the extremal probability $\mathbb{P}(R_n = n)$, multiplied by $n!$, against n for the above four characteristic examples of iid points in the plane. Black: uniform points in the square. Red: uniform points in the disk. Green: Gaussian points. Blue: power-law points with $\theta = 2$. Dashed curves: fits $y = an + b \ln n + c$, so that $C = \mathrm{e}^a$. Curves are slightly offset vertically from one another for greater clarity.

circle, and the Pólya walk [51] on three lattices (triangular, square, hexagonal), where the $\boldsymbol{\delta}_n$ take a finite number z of discrete values, equal to the coordination number of the lattice (respectively 6, 4 and 3). In all these models we have $|\boldsymbol{\delta}_n| = 1$, and so

$$\langle |\boldsymbol{x}_n|^2 \rangle = n. \tag{3.2}$$

Many rigorous results on the convex hulls of planar random walks and Brownian curves have been derived, concerning in particular the mean value of their area, perimeter length and number of vertices [52, 53, 54] (see [23] for a synthetic review). More recent developments on the combinatorics of random walks in two and higher dimensions also address properties of their convex hulls [55, 56].

Various types of records may be attached to planar random walks. Three classes of such records, namely diagonal, simultaneous and radial ones, have been investigated recently [57]. The mean numbers of these records grow as universal powers of time, with respective exponents 1/4, 1/3 and 1/2. Their full asymptotic distributions have also been determined.

Hereafter we consider convex records, denoting by N_n the number of vertices of the convex hull of the walk at time n, and by R_n the number of convex records up to time n. We recall that $\boldsymbol{x}_n$ is not counted as a convex record if it falls exactly on the boundary of the convex hull $\boldsymbol{C}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1})$. This event occurs with non-zero probability in the case of lattice walks. Furthermore, to keep our numerical algorithm unchanged, we have imposed the constraint that the first two steps $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ of lattice walks are not parallel to each other.

The most notable difference with respect to the case of iid data points investigated in section 2 is that the position of the random walker spreads away from the origin, according to the diffusion law (3.2). As a consequence, records are progressively buried deeper and deeper inside the current convex hull which expands at the same diffusive scale as the walk. In particular, the number R_n of records is expected to grow much faster than N_n. Stated otherwise, data points are by far not exchangeable, so that the

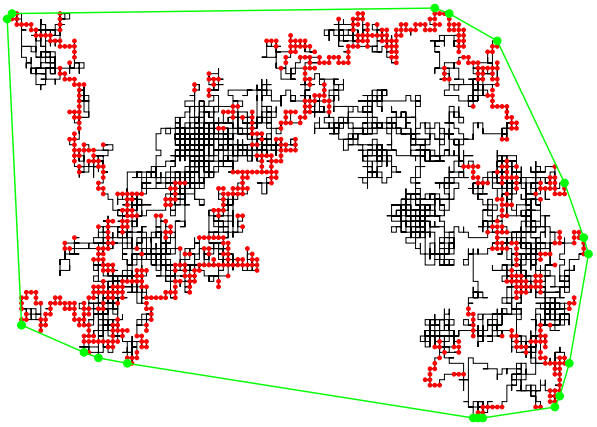

Figure 14. A Pólya walk of 10,000 steps on the square lattice with $N = 18$ and $R = 820$. Black line: trajectory of the random walker. Green polygon: convex hull of the walk. Green symbols: the 18 vertices of the convex hull. Red symbols: the 802 other convex records.

relation (2.5) can be expected to be violated by large amounts. This is illustrated in figure 14, showing a Pólya walk of 10,000 steps on the square lattice such that $N = 18$ and $R = 820$.

Let us now turn to a quantitative analysis, and consider first the number N_n of vertices of the convex hull at time n. For all microscopically isotropic planar random walks, including the Pearson walk, the mean number $\langle N_n \rangle$ of vertices has been long known. A combinatorial argument due to Baxter [53] yields the simple expression

$$\langle N_n \rangle = 2H_n \approx 2(\ln n + \gamma), \tag{3.3}$$

where the harmonic numbers H_n are defined in (A.9) and γ is Euler's constant. No expression seems to be known for the corresponding variance.

The logarithmic growth law (3.3) appears to be universal, including its prefactor, at least among the planar random walks we have investigated. Figure 15 shows plots of the difference $\langle N_n \rangle - 2 \ln n$ against $\ln n$ for the four examples of random walks considered in this work. The plotted data strongly support the behaviour

$$\langle N_n \rangle \approx 2 \ln n + G, \tag{3.4}$$

where the additive constant G depends on the type of walk (see table 2 below). For the Pearson walk, data for finite n converge rather fast from above to $G \approx 1.15$, in perfect agreement with the known limit $G(\text{Pearson}) = 2\gamma = 1.154431\ldots$ (see (3.3)). For the Pólya walk on three lattices, data exhibit a slower convergence from below to higher values of G.

Figure 16 shows plots of $\operatorname{Var} N_n$ against $\ln n$ for the same four random walks. The data support the universal logarithmic grow law

$$\operatorname{Var} N_n \approx A_2 \ln n, \tag{3.5}$$

where the prefactor assumes the seemingly universal value

$$A_2 \approx 1.50, \tag{3.6}$$

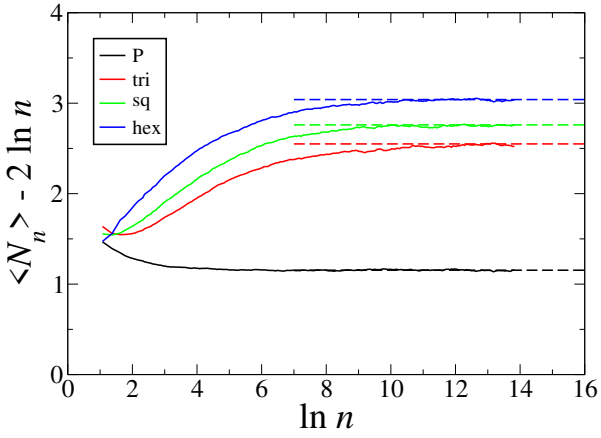

Figure 15. Difference $\langle N_n \rangle - 2 \ln n$ plotted against $\ln n$ for the four random walks considered in this work. Black: Pearson walk. Other colours: Pólya walk on three lattices. Red: triangular. Green: square. Blue: hexagonal. Horizontal dashed lines: extrapolated limits yielding the values of the additive constant G listed in table 2.

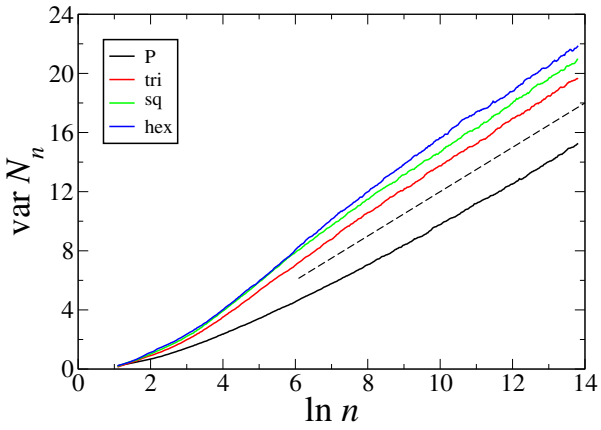

Figure 16. Variance of N_n plotted against $\ln n$ for the four random walks (see legend). Colours are as in figure 15. The dashed line has slope 1.50.

with an expected relative accuracy of a few percent. The results (3.4) and (3.5) imply that the distribution of N_n is characterised by the universal limit Fano factor

$$F_N = \frac{A_2}{2} \approx 0.75. \tag{3.7}$$

The occurrence of long transients in the data shown in figure 16 prevents us from measuring the additive constant of the logarithmic law (3.5) in an accurate way. Here, too, it is likely that N_n obeys a central limit theorem, i.e., has an asymptotic Gaussian distribution, and that its higher cumulants grow proportionally to $\ln n$, even though it is difficult to reliably confirm this hypothesis by purely numerical means.

Before we pursue, it is worth emphasising the analogy between the convex records

of planar random walks and their radial records, investigated in [57]. The position $\boldsymbol{x}_n$ of the walker at time n is a radial record if it is outside the circle whose radius¶

$$\overline{r}_n = \max(|\boldsymbol{x}_0|, \ldots, |\boldsymbol{x}_{n-1}|) \tag{3.8}$$

is the largest distance to the origin reached by the walker before time n, whereas $\boldsymbol{x}_n$ is a convex record if it is outside the convex hull $\boldsymbol{C}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1})$. This convex polygon keeps forever fluctuating, to the extent that it may assume any shape [58]. It is nevertheless to be expected that the radius $\overline{r}_n$ and the diameter of $\boldsymbol{C}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n)$ grow proportionally to the diffusive scale $\sqrt{n}$, and so that there are similarities between the statistics of radial and convex records.

We recall that the main outcomes of [57] concerning radial records of random walks are based on the asymptotic equivalence

$$R_n^{(\mathrm{rad})} \approx \frac{\overline{r}_n}{a} \tag{3.9}$$

between the number $R_n^{(\mathrm{rad})}$ of radial records and the radius $\overline{r}_n$ introduced in (3.8). In (3.9), the microscopic length scale a depends on the type of walk, whereas the radius $\overline{r}_n$ scales as

$$\overline{r}_n \approx U\sqrt{n}, \tag{3.10}$$

where the random variable U has a universal distribution that is known explicitly [59, p. 280] and recalled in [57, Eq. (4.12)].

In particular, the mean number of radial records grows as

$$\langle R_n^{(\mathrm{rad})} \rangle \approx B^{(\mathrm{rad})}\sqrt{n}, \tag{3.11}$$

where the prefactor

$$B^{(\mathrm{rad})} = \frac{\langle U \rangle}{a} \tag{3.12}$$

depends on the type of walk through a. Reference [57] gives $B^{(\mathrm{rad})}(\mathrm{Pearson}) \approx 2.35$ for the Pearson walk and $B^{(\mathrm{rad})}(\mathrm{square}) \approx 2.10$ for the Pólya walk on the square lattice. The number of radial records is asymptotically distributed according to

$$\frac{R_n^{(\mathrm{rad})}}{\langle R_n^{(\mathrm{rad})} \rangle} \to X^{(\mathrm{rad})} = \frac{U}{\langle U \rangle}. \tag{3.13}$$

The limit random variable X therefore has a universal distribution such that $\langle X^{(\mathrm{rad})} \rangle = 1$, by construction, whereas

$$\mathrm{Var}\, X^{(\mathrm{rad})} = \frac{\mathrm{Var}\, U}{\langle U \rangle^2} = 0.110751\ldots \tag{3.14}$$

Let us come back to convex records of random walks. Concerning the mean number $\langle R_n \rangle$ of these records, the numerical data shown in figure 17 strongly suggest the growth law

$$\langle R_n \rangle \approx B\sqrt{n}\, \ln n, \tag{3.15}$$

where the prefactor B has a weak dependence on the type of walk (see table 2). A plausible justification for the above scaling law is that R_n contains both a factor $\sqrt{n}$, already present in (3.11), representing the diffusive growth of the walk, and a

¶ In [57] n is denoted by t, $\overline{r}_n$ by $\overline{R}_t$, $R_n^{(\mathrm{rad})}$ by N_t, and $B^{(\mathrm{rad})}$ by A.

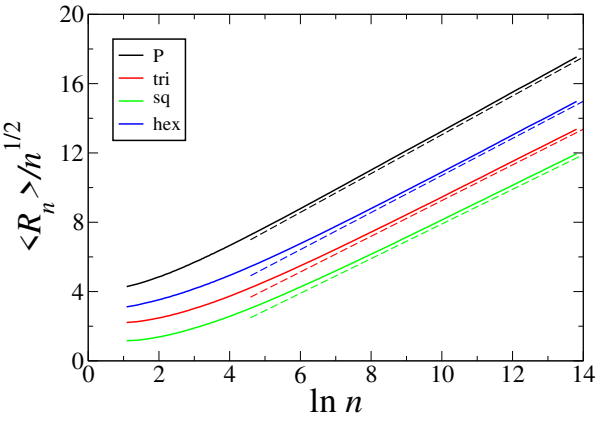

Figure 17. Ratio $\langle R_n \rangle / \sqrt{n}$ plotted against $\ln n$ for the four random walks (see legend). Colours are as in figure 15. Dashed lines have the slopes B listed in table 2.

Type of walk	z	G	B
Pearson	∞	1.15	1.12
Pólya (triangular)	6	2.55	1.03
Pólya (square)	4	2.76	1.00
Pólya (hexagonal)	3	3.04	1.07

Table 2. Various characteristic constants of the four random walks considered in this work: z is the number of directions taken by the increments $\boldsymbol{\delta}_n$, i.e., the coordination number of the lattice, G is the additive constant of the logarithmic law (3.4) for the mean number $\langle N_n \rangle$ of vertices, B is the prefactor of the growth law (3.15) for the mean number $\langle R_n \rangle$ of convex records.

factor $\ln n$, representing the number N_n of vertices of the convex hull at the current time n (see (3.4)).

It is interesting to notice that the ratio $B(\text{Pearson})/B(\text{square}) \approx 1.12$ equals the corresponding ratio for radial records, i.e., $B^{(\text{rad})}(\text{Pearson})/B^{(\text{rad})}(\text{square}) \approx 2.35/2.10 \approx 1.12$, given the available precision. This coincidence suggests that the prefactor B, just as $B^{(\text{rad})}$, only depends on the type of walk through the microscopic length scale a introduced in [57].

Concerning the full distribution of the number of convex records, it is to be expected, in line with (3.13), that R_n keeps fluctuating and is asymptotically distributed according to

$$\frac{R_n}{\langle R_n \rangle} \to X, \tag{3.16}$$

where the limit random variable X has a universal distribution such that $\langle X \rangle = 1$. This expectation, too, is corroborated by numerical simulations. Figure 18 shows the

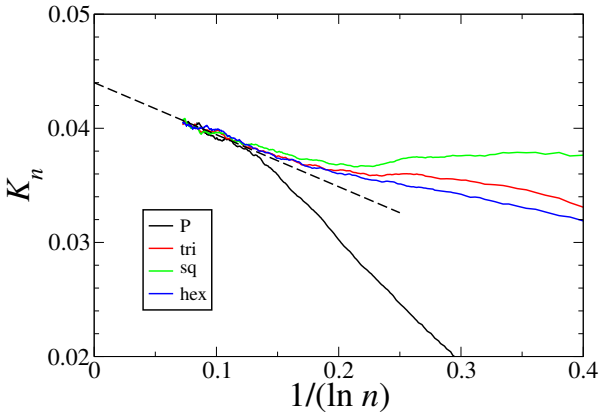

Figure 18. Reduced variance K_n of the number of convex records plotted against $1/(\ln n)$, for the four random walks (see legend). Colours are as in figure 15. The dashed line has intercept 0.044.

reduced variance of the number of convex records,

$$K_n = \frac{\operatorname{Var} R_n}{\langle R_n \rangle^2}, \tag{3.17}$$

plotted against $1/(\ln n)$ for the four random walks considered in this work. The data is observed to converge linearly to the universal limit

$$K = \operatorname{Var} X \approx 0.044. \tag{3.18}$$

The very slowly decaying corrections in $1/(\ln n)$ however come as a surprise.

Figure 19 shows the distribution of $R_n/\langle R_n \rangle$ for Pearson walk and for Pólya walk on the square lattice with $n = 10,000$ steps. Both datasets fall on the same smooth curve, representing a good approximation of the distribution of the limit variable X. The exactly known distribution of $X^{(\mathrm{rad})}$, corresponding to radial records (see (3.13)), is shown for comparison. The latter distribution is broader. The variance of $X^{(\mathrm{rad})}$ is indeed some 2.5 times larger than that of X (see (3.14), (3.18)).

4. Discussion

The primary aim of this paper is to draw the reader's attention to convex records. Among all possible definitions of multivariate records, the convex records investigated in this work stand out for the elegance of their geometric definition and for the ensuing invariance of their construction under the affine group.

The present work is focused on the bivariate (i.e., two-dimensional) case. This choice is motivated by simplicity, and chiefly by the existence of a simple algorithm to recursively build convex hulls of growing data sets. We wish to highlight that some of the statistics of higher-dimensional convex records can be sketched in the light of the results presented above. To be more specific, for iid (independent and identically distributed) data points, the identity (2.5) has been instrumental in relating the mean number $\langle R_n \rangle$ of convex records up to time n to the mean number $\langle N_n \rangle$ of vertices of the convex hull of the first n points. The above identity is in fact quite general,

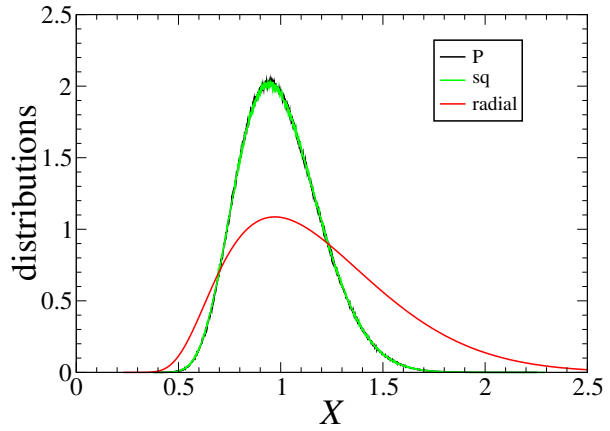

Figure 19. Distribution of the reduced variable $R_n/\langle R_n\rangle \approx X$ (unbinned data) for walks of $n = 10,000$ steps. Black: Pearson walk. Green: Pólya walk on the square lattice. Red: exactly known distribution of $X^{(\mathrm{rad})}$, corresponding to radial records (see (3.13)).

and applies to iid data points in any dimension d. Let us consider two characteristic examples, for which some results on convex hulls are available in the mathematical literature. For uniform data points in an arbitrary d-dimensional convex polytope, we have

$$\langle N_n\rangle \approx A_1(\ln n)^{d-1}, \tag{4.1}$$

where the prefactor A_1 is known [60], and so (2.5) yields

$$\langle R_n\rangle \approx B_1(\ln n)^d, \qquad B_1 = \frac{A_1}{d}. \tag{4.2}$$

For uniform data points in a d-dimensional sphere, we have

$$\langle N_n\rangle \approx A_1\, n^{(d-1)/(d+1)}, \tag{4.3}$$

where A_1 is also known exactly [61, 62], and so (2.5) yields

$$\langle R_n\rangle \approx B_1\, n^{(d-1)/(d+1)}, \qquad B_1 = \frac{d+1}{d-1}A_1. \tag{4.4}$$

Moreover, it is quite plausible that the variances of N_n and R_n generically grow proportionally to the corresponding mean values given above, resulting in finite limit Fano factors F_N and F_R, such as those shown in figure 12. On the other hand, for isotropic random walks in d dimensions, a known exact expression of $\langle N_n\rangle$ for any finite number n of steps [56] reads asymptotically

$$\langle N_n\rangle \approx \frac{2}{(d-1)!}\,(\ln n)^{d-1}. \tag{4.5}$$

We can therefore expect that the mean number of convex records scales as

$$\langle R_n\rangle \approx B\sqrt{n}\,(\ln n)^{d-1}. \tag{4.6}$$

This estimate should hold for the Pearson walk and for the Pólya walk on a lattice in any dimension d, with the prefactor B depending on the type of walk. Finally, it is

also to be expected that $R_n/\langle R_n\rangle$ keeps fluctuating and goes to a universal random variable X, whose distribution only depends on d.

We conclude with a discussion on the growth of the number of records for various kinds of records in sequences of d-dimensional iid data points. The mean number of records can be expressed in terms of the record-breaking probability Q_n (see (2.2)). For the convex records studied in this work, a d-dimensional simplex has $d+1$ vertices. As a result, we have generically $nQ_n = \langle N_n\rangle \geq d+1$ for all $n \geq d+1$, implying that the mean number of records, $\langle R_n\rangle$, grows at least as $(d+1)\ln n$. This minimal logarithmic growth should be compared with the growth rates observed for other definitions of multivariate records. Consider simultaneous records (also known as complete or concomitant records), where there is a record at time n if each component x_n^i of $\boldsymbol{x}_n$ is larger than all previous x_m^i. There, the record-breaking probability Q_n can assume any value between 0 and $1/n$ (see [15] and references therein). The upper bound coincides with the result (1.2) of the univariate case. Simple explicit examples can be built in any dimension d, for which either $Q_n = 0$ for all $n > 1$, or $Q_n = 1/n$. For data points with independent components x_n^i following continuous distributions, we have $Q_n = 1/n^d$. For Gaussian data points, we have $Q_n \sim n^{-\alpha}$, where the exponent $\alpha > 1$ depends continuously on parameters. In both examples, the total number of simultaneous records remains finite for an infinitely large dataset.

Acknowledgments

It is a pleasure to thank Philippe Naveau for the discussions that motivated this work.

Data availability statement

Data sharing not applicable to this article.

Conflict of interest

The authors declare no conflict of interest.

Orcid ids

Claude Godrèche https://orcid.org/0000-0002-1833-3490
Jean-Marc Luck https://orcid.org/0000-0003-2151-5057

Appendix A. The distribution of univariate records

This appendix is a self-contained reminder of the classical theory of the statistics of records in sequences of iid univariate random variables drawn from an arbitrary continuous distribution (see [2, 3, 4, 5, 6, 7, 8, 9]). In this setting, there is a record at time n with probability (see (1.2))

$$Q_n = \frac{1}{n}, \tag{A.1}$$

and the occurrences of records at different times are statistically independent.

The quantity of interest is the number R_n of records up to time n. The distribution of this random number,

$$p_n(k) = \mathbb{P}(R_n = k) \qquad (k = 1, \ldots, n), \tag{A.2}$$

is conveniently encoded in the generating function

$$G_n(z) = \langle z^{R_n} \rangle = \sum_{k=1}^{n} p_n(k) z^k, \tag{A.3}$$

which is a polynomial in z with degree n. The independence of the occurrences of records at different times yields the product formula

$$G_n(z) = \prod_{m=1}^{n} (1 - Q_m + zQ_m) = \prod_{m=1}^{n} \frac{m - 1 + z}{m}. \tag{A.4}$$

This can be recast as

$$G_n(z) = \frac{\Gamma(n+z)}{n!\,\Gamma(z)} = \frac{1}{n!} \sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix} z^k, \tag{A.5}$$

where the $\begin{bmatrix} n \\ k \end{bmatrix}$ are the Stirling numbers of the first kind, which are ubiquitous in combinatorics (see e.g. [63, 64]). The distribution of R_n therefore reads

$$p_n(k) = \frac{1}{n!} \begin{bmatrix} n \\ k \end{bmatrix}. \tag{A.6}$$

The integer $\begin{bmatrix} n \\ k \end{bmatrix}$ is, among many other things, the number of permutations of n objects having k cycles, so that the number R_n of records is distributed as the number of cycles in a uniform random permutation.

In particular, the mean value and the variance of R_n read

$$\langle R_n \rangle = \sum_{m=1}^{n} Q_m = H_n \approx \ln n + \gamma, \tag{A.7}$$

$$\operatorname{Var} R_n = \sum_{m=1}^{n} Q_m (1 - Q_m) = H_n - H_n^{(2)} \approx \ln n + \gamma - \frac{\pi^2}{6}, \tag{A.8}$$

where

$$H_n = \sum_{m=1}^{n} \frac{1}{m}, \qquad H_n^{(2)} = \sum_{m=1}^{n} \frac{1}{m^2}, \tag{A.9}$$

and γ is Euler's constant.

The number R_n of records takes its smallest and largest values with respective probabilities

$$p_n(1) = \frac{1}{n}, \qquad p_n(n) = \frac{1}{n!}. \tag{A.10}$$

The full distribution of R_n takes a simple asymptotic form at large n. A first approximation to (A.5) at large n reads

$$G_n(z) \sim \mathrm{e}^{(z-1)\ln n}, \tag{A.11}$$

and so the distribution of R_n becomes a Poisson distribution with parameter $\lambda = \ln n$ and Fano factor $F = 1$. A more refined asymptotic form of (A.5) is

$$G_n(z) \approx \frac{\mathrm{e}^{(z-1)\ln n}}{\Gamma(z)}, \tag{A.12}$$

implying that the cumulants of R_n grow as

$$\langle R_n^p \rangle_c \approx \ln n + a_p, \tag{A.13}$$

with a common logarithmic term with unit prefactor, and additive constants a_p given by

$$\sum_{p\geq 1} \frac{a_p}{p!}\, s^p = -\ln \Gamma(\mathrm{e}^s), \tag{A.14}$$

i.e.,

$$\begin{aligned} &a_1 = \gamma, \qquad a_2 = \gamma - \frac{\pi^2}{6}, \qquad a_3 = \gamma - \frac{\pi^2}{2} + 2\zeta(3), \\ &a_4 = \gamma - \frac{7\pi^2}{6} - \frac{\pi^4}{15} + 12\zeta(3), \end{aligned} \tag{A.15}$$

and so on. The first two expressions agree with (A.7) and (A.8).

References

[1] C. Godrèche, S. N. Majumdar, and G. Schehr. Record statistics of a strongly correlated time series: random walks and Lévy flights. *J. Phys. A: Math. Theor.*, 50:333001, 2017.

[2] K. N. Chandler. The distribution and frequency of record values. *J. R. Statist. Soc. B.*, 14:220–228, 1952.

[3] A. Rényi. Théorie des éléments saillants d'une suite d'observations. *Ann. Sci. Univ. Clermont-Ferrand*, 8:7–13, 1962.

[4] A. Rényi. Théorie des éléments saillants d'une suite d'observations. In *Colloquium on Combinatorial Methods in Probability Theory*, pages 104–117. Mathematical Institute of Aarhus University, Aarhus, Denmark, 1962.

[5] N. Glick. Breaking records and breaking boards. *Amer. Math. Monthly*, 85:2–26, 1978.

[6] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *Records*. Wiley, New York, 1998.

[7] V. B. Nevzorov and N. Balakrishnan. A record of records. *Handbook of Statistics*, 16:515–570, 1998.

[8] V. B. Nevzorov. *Records: Mathematical Theory (Translation of Mathematical Monographs vol* **194***)*. American Mathematical Society, Providence, RI, 2001.

[9] J. Bunge and C. M. Goldie. Record sequences and their applications. *Handbook of Statistics*, 19:277–308, 2001.

[10] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 2. Wiley, New York, 2nd edition, 1971.

[11] M. G. Kendall. Discrimination and classification, multivariate analysis. In P. R. Krishnaiah, editor, *Multivariate Analysis*, pages 165–184, New York, 1966. Academic.

[12] V. Barnett. The ordering of multivariate data. *J. R. Statist. Soc. A*, 139:318–355, 1976.

[13] C. M. Goldie and S. I. Resnick. Records in a partially ordered set. *Ann. Probab.*, 17:678–699, 1989.

[14] C. M. Goldie and S. I. Resnick. Many multivariate records. *Stoch. Proc. Appl.*, 59:185–216, 1995.

[15] A. V. Gnedin. Records from a multivariate normal sample. *Stat. Probab. Lett.*, 39:11–15, 1998.

[16] H. K. Hwang and T. H. Tsai. Multivariate records based on dominance. *Electron. J. Probab.*, 15:1863–1892, 2010.

[17] C. Dombry and M. Zott. Multivariate records and hitting scenarios. *Extremes*, 21:343–361, 2018.

[18] N. Balakrishnan, A. Stepanov, and V. B. Nevzorov. North-east bivariate records. *Metrika*, 83:961–976, 2020.

[19] S. Tat and M. R. Faridrohani. A new type of multivariate records: depth-based records. *Statistics*, 55:296–320, 2021.

[20] M. Kaluszka. Estimates of some probabilities in multidimensional convex records. *Applicationes Math.*, 23:1–11, 1995.

[21] M. G. Kendall and P. A. P. Moran. Geometrical probability. In M. G. Kendall, editor, *Griffin's Statistical Monographs and Courses*, volume 10. Hafner, New York, 1963.

[22] R. Schneider and W. Weil. *Stochastic and Integral Geometry*. Springer, Berlin, 2008.

[23] S. N. Majumdar, A. Comtet, and J. Randon-Furling. Random convex hulls and extreme value statistics. *J. Stat. Phys.*, 138:955–1009, 2010.
[24] J. J. Sylvester. Problem 1491. *The Educational Times*, April 1864.
[25] R. E. Pfeifer. The historical development of J. J. Sylvester's four point problem. *Math. Mag.*, 62:309–317, 1989.
[26] W. Blaschke. *Vorlesungen über Differentialgeometrie II, Affine Differentialgeometrie*. Springer, Berlin, 1923.
[27] A. Rényi and R. Sulanke. Über die konvexe Hülle von n zufällig gewählten Punkten. *Z. Wahr.*, 2:75–84, 1963.
[28] B. Efron. The convex hull of a random set of points. *Biometrika*, 52:331–343, 1965.
[29] P. Groeneboom. Limit theorems for convex hulls. *Probab. Th. Rel. Fields*, 79:327–368, 1988.
[30] U. Fano. Ionization yield of radiations. II. The fluctuations of the number of ions. *Phys. Rev.*, 72:26–29, 1947.
[31] J. Tworzydło, B. Trauzettel, M. Titov, A. Rycerz, and C. W. J. Beenakker. Sub-Poissonian shot noise in graphene. *Phys. Rev. Lett.*, 96:246802, 2006.
[32] L. Mandel. Sub-Poissonian photon statistics in resonance fluorescence. *Optics Lett.*, 4:205–207, 1979.
[33] C. Buchta. The exact distribution of the number of vertices of a random convex chain. *Mathematika*, 53:247–254, 2006.
[34] C. Buchta. On the number of vertices of the convex hull of random points in a square and a triangle. *Anzeiger Abt. II*, 143:3–10, 2009.
[35] J. F. Marckert. The probability that n random points in a disk are in convex position. *Braz. J. Probab. Stat.*, 31:320–337, 2017.
[36] S. Finch and I. Hueter. Random convex hulls: a variance revisited. *Adv. Appl. Probab.*, 36:981–986, 2004.
[37] H. Carnal. Die konvexe Hülle von n rotationssymmetrisch verteilten Punkten. *Z. Wahr.*, 15:168–176, 1970.
[38] I. Hueter. The convex hull of a normal sample. *Adv. Appl. Probab.*, 26:855–875, 1994.
[39] I. Hueter. Limit theorems for the convex hull of random points in higher dimensions. *Trans. Amer. Math. Soc.*, 351:4337–4363, 1999.
[40] P. L. Krapivsky and J. M. Luck. On multidimensional record patterns. *J. Stat. Mech.*, 2020:063205, 2020.
[41] D. J. Aldous, B. Fristedt, P. S. Griffin, and W. E. Pruitt. The number of extreme points in the convex hull of a random sample. *J. Appl. Probab.*, 28:287–304, 1991.
[42] B. Massé. On the LLN fo the number of vertices of a random convex hull. *Adv. Appl. Probab.*, 32:675–681, 2000.
[43] W. J. Reed. Random points in a simplex. *Pacific J. Math.*, 54:183–198, 1974.
[44] V. S. Alagar. On the distribution of a random triangle. *J. Appl. Probab.*, 14:284–297, 1977.
[45] P. Valtr. Probability that n random points are in convex position. *Discrete Comp. Geom.*, 13:637–643, 1995.
[46] P. Valtr. The probability that n random points in a triangle are in convex position. *Combinatorica*, 16:567–573, 1996.
[47] H. J. Hilhorst, P. Calka, and G. Schehr. Sylvester's question and the random acceleration process. *J. Stat. Mech.*, 2008:P10010, 2008.
[48] I. Bárány. Sylvester's question: The probability that n points are in convex position. *Ann. Probab.*, 27:2020–2034, 1999.
[49] A. Vershik and O. Zeitouni. Large deviations in the geometry of convex lattice polygons. *Israel J. Math.*, 109:13–27, 1999.
[50] K. Pearson. The problem of the random walk. *Nature*, 72:294, 1905.
[51] G. Pólya. Wahrscheinlichkeitstheoretisches über die Irrfahrt. *Mitt. der Phys. Ges. Zürich*, 19:75–86, 1919.
[52] F. Spitzer and H. Widom. The circumference of a convex polygon. *Proc. Amer. Math. Soc.*, 12:506–509, 1961.
[53] G. Baxter. A combinatorial lemma for complex numbers. *Ann. Math. Statist.*, 32:901–904, 1961.
[54] L. Takacs. Expected perimeter length. *Amer. Math. Monthly*, 87:142, 1980.
[55] J. M. Steele. The Bohnenblust-Spitzer algorithm and its applications. *J. Comp. Appl. Math.*, 142:235–249, 2002.
[56] Z. Kabluchko, V. Vysotsky, and D. Zaporozhets. Convex hulls of random walks: Expected number of faces and face probabilities. *Adv. Math.*, 320:595–629, 2017.
[57] C. Godrèche and J. M. Luck. On sequences of records generated by planar random walks. *J.*

Phys. A: Math. Theor., 54:325003, 2021.

[58] J. McRedmond and A. R. Wade. The convex hull of a planar random walk: perimeter, diameter, and shape. *Electron. J. Probab.*, 23:1–24, 2018.

[59] A. N. Borodin and P. Salminen. *Handbook of Brownian Motion - Facts and Formulae*. Birkhäuser, Basel, 1996.

[60] I. Bárány and C. Buchta. Random polytopes in a convex polytope, independence of shape, and concentration of vertices. *Math. Ann.*, 297:467–497, 1993.

[61] H. Raynaud. Sur l'enveloppe convexe des nuages de points aléatoires dans $\mathbb{R}^n$. *J. Appl. Probab.*, 7:35–48, 1970.

[62] M. Reitzner. Random polytopes and the Efron-Stein jackknife inequality. *Ann. Probab.*, 31:2136–2166, 2003.

[63] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, MA, 1989.

[64] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, 2009.