Asymptotic behavior of the length of the longest increasing subsequences of random walks

J. Ricardo G. Mendonça [1,2,*] Hendrik Schawe [3,4,†] and Alexander K. Hartmann [4,‡]

[1]*Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 03828-000 São Paulo, Brazil*
[2]*LPTMS, CNRS UMR 8626, Université Paris–Sud, Université Paris–Saclay, 91405 Orsay CEDEX, France*
[3]*Laboratoire de Physique Théorique et Modélisation, CNRS UMR 8089, CY Cergy Paris Université, 95000 Cergy, France*
[4]*Institut für Physik, Universität Oldenburg, 26111 Oldenburg, Germany*

We numerically estimate the leading asymptotic behavior of the length L_n of the longest increasing subsequence of random walks with step increments following Student's t-distribution with parameters in the range $1/2 \leqslant \nu \leqslant 5$. We find that the expected value $\mathbb{E}(L_n) \sim n^\theta \ln n$, with θ decreasing from $\theta(\nu = 1/2) \approx 0.70$ to $\theta(\nu \geqslant 5/2) \approx 0.50$. For random walks with a distribution of step increments of finite variance ($\nu > 2$), this confirms previous observation of $\mathbb{E}(L_n) \sim \sqrt{n} \ln n$ to leading order. We note that this asymptotic behavior (including the subleading term) resembles that of the largest part of random integer partitions under the uniform measure and that, curiously, both random variables seem to follow Gumbel statistics. We also provide more refined estimates for the asymptotic behavior of $\mathbb{E}(L_n)$ for random walks with step increments of finite variance.

I. INTRODUCTION

The longest increasing subsequence (LIS) problem is to find an increasing subsequence of maximum length of a given finite sequence of n elements taken from a partially ordered set. Let $\mathcal{X}_n = (X_1, \ldots, X_n)$ be such a sequence, say, of real numbers. The longest (weakly) increasing subsequence of $\mathcal{X}_n$ is the longest subsequence $X_{i_1} \leqslant X_{i_2} \leqslant \cdots \leqslant X_{i_L}$ of $\mathcal{X}_n$ such that $1 \leqslant i_1 < i_2 < \cdots < i_L \leqslant n$, with L the length of the LIS. There may be more than one "longest" increasing subsequence for a given $\mathcal{X}_n$, with different elements but of the same maximum length. Algorithmically, it takes $O(n)$ space and $O(n \log \log n)$ time to find one LIS of a given sequence of n elements [1].

Despite a superficial similitude, the LIS and the sequence of records of a time series, a rich traditional subject frequently studied in statistical mechanics, are unrelated and should not be confused. The LIS is the maximum-length increasing subsequence, while the sequence of records is the subsequence of increasing maxima. For example, in the sequence (2,1,6,4,3,5) the sequence of records is (2,6), while the LISs are (1,3,5), (1,4,5), (2,3,5), and (2,4,5). Note also that the LIS of a time series is a much more intricate quantity than its set of records, since it depends on the entire series, not just on past events; it is a global property of the series. Algorithmically, the sequence of records can be obtained by one linear sweep through the sequence and can also be determined online (on the go) as the series progress, while more elaborate approaches are necessary to determine the LIS [1]. As far as we currently understand, results on records

are of no avail in the treatment of the LIS problem. The reader interested in the modern theory of records should consult [2,3]. Examples of the application of the LIS can be found in computer science, where the sortdedness of a data stream or list of items can be used, e.g., to decide whether the data have to be sorted again [4], and also in the design of privacy preserving algorithms for publicly shared information through sequential data streams [5].

Initially, the LIS problem was considered in the mathematical literature for random permutations. The problem seems to have been first considered by Ulam in the early 1960s [6]. The resolution of the LIS problem for random permutations culminated with the exact characterization of L_n as a random variable distributed like $L_n \sim 2\sqrt{n} + \sqrt[6]{n}\chi$ with $\mathbb{P}(\chi \leqslant s) = F_2(s)$, the Tracy-Widom distribution for the fluctuations of the largest eigenvalue of a Gaussian unitary random matrix ensemble about its expected value [7,8]. Comprehensive expositions on the LIS problem for random permutations appear in [9,10].

Recently, another version of the LIS problem was posed [11,12]: What is the behavior of the LIS of a random walk? Let $\mathcal{X}_n = (X_1, \ldots, X_n)$ be the sequence of terms of a random walk given by

$$X_0 = 0, \quad X_t = X_{t-1} + \xi_t, \quad t = 1, \ldots, n, \tag{1}$$

with the ξ_t, $t = 1, \ldots, n$, independent random variables identically distributed according to some zero-mean symmetric probability distribution $\phi(\xi)$. The sequence $\mathcal{X}_n$ constitutes a time series of correlated random variables: if the expectation $\mathbb{E}(\xi^2) \neq 0$, then $\mathbb{E}(X_t X_s) \neq 0$. In [11], the authors showed that when $\phi(\xi)$ has finite positive variance, then for all $\epsilon > 0$ and large enough n the length L_n of the LIS of $\mathcal{X}_n$ observes

$$c\sqrt{n} \leqslant \mathbb{E}(L_n) \leqslant n^{1/2+\epsilon} \tag{2}$$

*jricardo@usp.br
†hendrik.schawe@u-cergy.fr
‡a.hartmann@uni-oldenburg.de
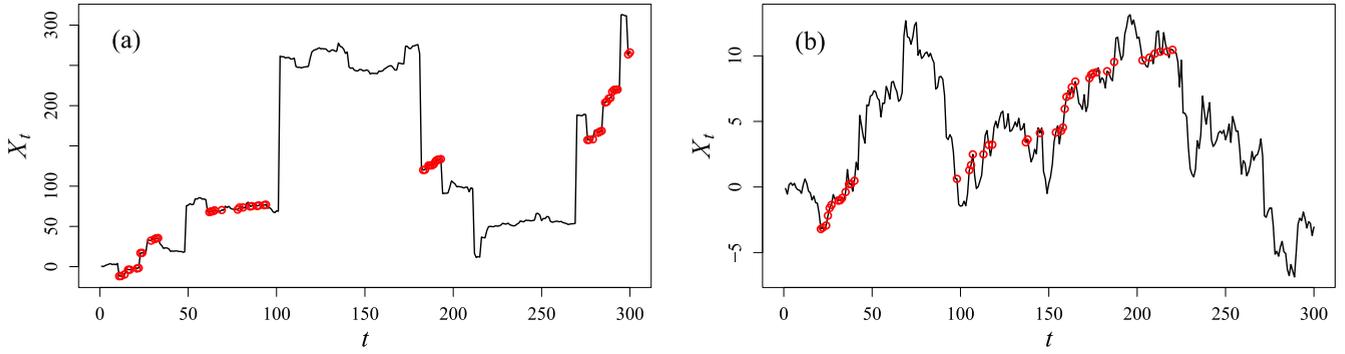
FIG. 1. Student's t random walks of 300 steps each with parameters (a) $\nu = 1$ and (b) $\nu = 4$ together with one of their LISs (small circles). Note the different vertical scales.

for some positive constant c. The upper bound in Eq. (2) does not rule out a logarithmic term and can actually be read as

$$\mathbb{E}(L_n) \leqslant \sqrt{n}(\ln n)^a \qquad (3)$$

for some $a \geqslant 0$. In [12] the authors further proved that the expected length of the LIS of a particular random walk with step lengths of ultraheavy distribution without any finite (integer or fractional) moment scales with the length of the walk as

$$n^{\beta_0 - o(1)} \leqslant \mathbb{E}(L_n) \leqslant n^{\beta_1 + o(1)}, \qquad (4)$$

with nonsharp $\beta_0 \simeq 0.690$ and $\beta_1 \simeq 0.815$. Besides the bounds (2)–(4), little is known rigorously about the LIS of random walks.

Figure 1 displays two random walks of 300 steps distributed according to a Student t-distribution (see Sec. II), one with parameter $\nu = 1$ (the same as the Cauchy distribution) and the other with parameter $\nu = 4$, together with one of their LIS each.

In order to improve our knowledge about the LIS of random walks, we ran Monte Carlo simulations [13,14] to estimate the scaling of L_n for several different distributions of step lengths [15,16]. The simulations showed that

$$\mathbb{E}(L_n) \sim n^\theta \qquad (5)$$

with a nonuniversal scaling exponent $0.60 \lesssim \theta \lesssim 0.69$ for the heavy-tailed distributions of step lengths examined, with θ increasing as the distribution of step lengths becomes more heavy tailed. For distributions of finite variance, one of us conjectured [15] that the asymptotic behavior of $\mathbb{E}(L_n)$ in these cases is given by

$$\mathbb{E}(L_n) \sim \frac{1}{e}\sqrt{n} \ln n + \frac{1}{2}\sqrt{n} \qquad (6)$$

plus lower-order terms, although the constants are presumed based on least-squares adjustments. Moreover, it has been found that the empirical distribution of L_n seems to be of the form

$$f(L_n) = \frac{1}{\mathbb{E}(L_n)} g\left(\frac{L_n}{\mathbb{E}(L_n)}\right), \qquad (7)$$

with $\mathbb{E}(L_n)$ given by Eq. (5) or (6) depending on whether $\phi(\xi)$ has a finite second moment or not. Accordingly, when the

step lengths are of finite variance $g(z)$ should be universal. Plots of $g(z)$ for different distributions of step lengths appear in Ref. [15] (see also Fig. 4). The form (7) was further tested in Ref. [16], which probed the distribution of L_n into regions of very small probabilities for random walks with uniform increments $\xi \sim U(-1, 1)$. The authors found very good agreement between Eqs. (6) and (7) and their data. They also estimated that the large-deviation rate function $\Phi(L)$ associated with the distribution of L_n by

$$f(L_n) \asymp \exp[-n\Phi(L_n)] \qquad (8)$$

behaves asymptotically, in the limit of large $n \to \infty$, like $\Phi(L) \sim L^{-1.6}$ in the left tail and $\Phi(L) \sim L^{2.9}$ in the right tail. Despite this characterization, the distribution $g(z)$ remains unknown. It is tempting to conjecture that the actual exponents in $\Phi(L)$ are, respectively, $L^{-3/2}$ and L^3, in which case they would be the same as those of the large-deviation rate function of the Tracy-Widom F_2 distribution with the sides (left $\leftrightarrow$ right) and signs flipped [17]. Note, however, that unlike in Eq. (8), the large-deviation rate function for the Tracy-Widom distribution in the right tail is defined by a relationship of the form $\exp[-\sqrt{n}\Phi(L)]$, i.e., with an unusual $\sqrt{n}$ scaling.

Interestingly, while the LIS of random permutations played a fundamental role in the development of a wealth of mathematical physics in the past few decades, the LIS of random walks had not receive any attention until very recently. Since the LIS finds applications in the analysis of data streams and time series, it is necessary to understand its fundamental behavior. In this paper we provide an updated account on the LIS problem for random walks. While in previous studies of the problem only specific distributions $\phi(\xi)$ of step increments were considered, in Sec. II we employ a parametrized distribution, namely, Student's t-distribution, considerably enlarging the class of random walks considered. This approach allows us to investigate the dependence of the scaling exponent θ in (5) with the heavy tail index of $\phi(\xi)$. In Sec. III we further verify the proposed scaling form (6) and (7) with independent data, this time taking the full distribution of the data into account. We also remark on the resemblance between the statistics of the LIS problem for random walks of finite variance and the random partition problem under the uniform measure. Section IV summarizes the paper and provides perspectives for further study.

II. THE LIS OF HEAVY-TAILED RANDOM WALKS

We investigate the behavior of the scaling exponent appearing in the relation $L_n \sim n^\theta$ for random walks with a heavy-tail distribution of step increments as a function of their characteristic index α, defined by

$$\phi_\alpha(|\xi| \gg 1) \sim |\xi|^{-1-\alpha}. \tag{9}$$

We want to check whether there exists a well defined relationship between θ and α. In order to access a range of values of $\alpha \leqslant 2$ [such that $\mathbb{E}(\xi^2) = \infty$], we employ Student's t-distribution [18]

$$\phi_\nu(\xi) = \frac{\Gamma[\frac{1}{2}(\nu+1)]}{\sqrt{\nu\pi}\,\Gamma(\frac{1}{2}\nu)}\left(1 + \frac{\xi^2}{\nu}\right)^{-(\nu+1)/2}, \tag{10}$$

where $\Gamma(z)$ is the usual gamma function and ν is a parameter. This distribution appears in inference problems about unknown parameters (mean or variance or both) of a normal population. In statistical applications $\nu \geqslant 1$ is a natural number, but for modeling purposes ν can be taken a real positive number. When $\nu < \infty$, Student's t-distribution displays a heavy tail $\phi_\nu(|\xi| \gg 1) \sim |\xi|^{-1-\nu}$, with infinite variance if $\nu \leqslant 2$ and finite variance $\nu/(\nu-2)$ for $\nu > 2$. We see that ν plays the role of the tail index α in Eq. (9). Student's t-distribution becomes the Gaussian distribution in the limit $\nu \to \infty$.

For each parameter ν and walk length n, we generate 10^4 realizations of $\mathcal{X}_n$, compute their L_n, and estimate the empirical average $\langle L_n \rangle$ and variance $\langle L_n^2 \rangle - \langle L_n \rangle^2$. In our simulations $1/2 \leqslant \nu \leqslant 5$ and $10^4 \leqslant n \leqslant 10^7$ (and up to 2×10^8 for some values around $\nu = 2$). Student's t random deviates can be efficiently and reliably generated by the polar method [19]. Whenever $\mathbb{E}(\xi^2)$ is finite, i.e., $\nu > 2$, we use normalized random variables $\xi/\sqrt{\mathbb{E}(\xi^2)}$ for the step increments. To empirically obtain θ as a function of ν for all values of ν, and led by the previous results embodied in Eqs. (5) and (6), we fitted the data of the empirical averages $\langle L_n \rangle$ for all values of ν to

$$\langle L_n \rangle \sim n^\theta (1 + c\ln n) \tag{11a}$$

and

$$\langle L_n^2 \rangle - \langle L_n \rangle^2 \sim n^\gamma (1 + d\ln n) \tag{11b}$$

over the range of sizes investigated. Figure 2 displays log-log plots of $\langle L_n \rangle$ and $\langle L_n^2 \rangle - \langle L_n \rangle^2$ for $\nu = 2/3$ for illustration. The least-squares-fit estimates obtained for θ and γ are shown in Table I. For $\nu < 3/2$, the size of the logarithmic correction is very small; thus we set it to zero to obtain the final value of θ (which reduced the error bars for θ here). Note that for $\nu \geqslant 2$ all values of θ obtained are roughly in the range $[0.49, 0.51]$, i.e., $\theta = 1/2$ with a logarithmic correction seems to hold, confirming previous predictions [15,16]. In Sec. III B, however, we revisit the estimation of the constants appearing in (6). For $\nu \leqslant 3/2$, we obtained $\gamma \simeq 2\theta$ to a very good precision, in agreement with the picture provided by Fig. 2, suggesting that the probability density function (PDF) of L_n indeed follows the form (7), since then the kth moment
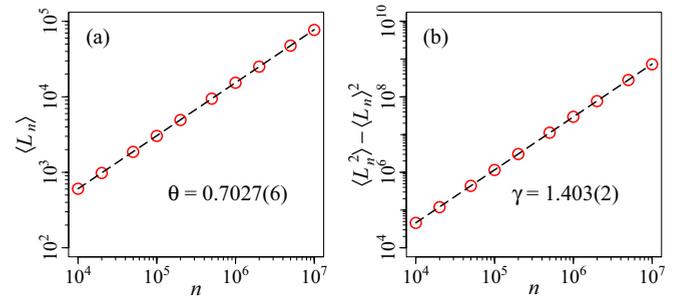
FIG. 2. Log-log plot of (a) the empirical mean and (b) the empirical variance of L_n for Student's t random walk with parameter $\nu = 2/3$ together with the least-squares fits (dashed lines). For this value of ν, the expression (11a) with $c = 0$ yields the best fit. The fact that the curves have virtually the same slope in the different vertical scales on the graphs suggests that $\gamma \simeq 2\theta$. Each point was obtained from an average over 10^4 sample random walks.

of L_n becomes

$$\mathbb{E}(L_n^k) = \int L_n^k f(L_n)dL_n = \mathbb{E}(L_n)^k \int z_n^k g(z_n)dz_n = c_{n,k}\mathbb{E}(L_n)^k, \tag{12}$$

with $z_n = L_n/\mathbb{E}(L_n)$ and $c_{n,k}$ the kth moment of the distribution $g(z_n) = \mathbb{E}(L_n)f(\mathbb{E}(L_n)z_n)$. We see that, with $f(L_n)$ like in Eq. (7), all moments $\mathbb{E}(L_n^k) \propto \mathbb{E}(L_n)^k$, as our data for $k = 1$ and 2 do for small enough values of ν. Note that, as a consequence,

$$\text{Var}(L_n) = \mathbb{E}(L_n^2) - \mathbb{E}(L_n)^2 = (c_{n,2} - c_{n,1}^2)\mathbb{E}(L_n)^2, \tag{13}$$

and the random variable L_n cannot possibly be self-averaging unless $(c_{n,2} - c_{n,1}^2) \xrightarrow{n} 0$, i.e., unless $g(z_n)$ becomes increasingly more concentrated with n,

$$g(z_n) \xrightarrow{n} \delta(z - c_1). \tag{14}$$

Our data indicate, however, that $g(z_n)$ remains broad irrespective of how large n gets.

Figure 3 displays a log-log plot of θ against ν. The plot does not suggest any clear functional relationship between θ and ν; we were hoping for something like $\theta \sim (\nu_c - \nu)^z$ in the interval $\nu \leqslant 2$. Otherwise, θ saturates at $\theta \simeq 0.5$ for

TABLE I. Exponents θ and γ fitted according to (11a) and (11b) for selected values of tail index ν. The ratio $\gamma/\theta \simeq 2$ suggests form (7) [see also Eq. (12)] for the PDF of L_n. The numbers in parentheses indicate the uncertainty in the last digit(s) of the data.

ν	θ	γ	γ/θ
1/2	0.7050(6)	1.419(2)	2.01
1	0.6850(6)	1.372(3)	2.00
3/2	0.588(7)	1.355(13)	2.30
2	0.516(3)	1.294(7)	2.51
5/2	0.500(4)	1.259(9)	2.52
3	0.496(3)	1.246(8)	2.51
7/2	0.501(5)	1.260(16)	2.51
4	0.504(4)	1.259(10)	2.40
5	0.497(3)	1.257(10)	2.52
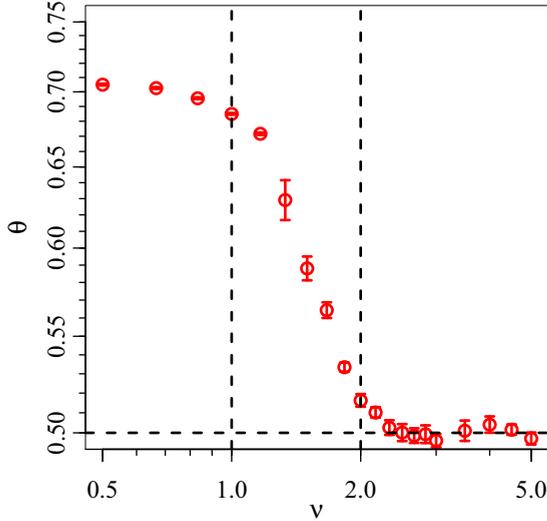
FIG. 3. Log-log plot of the scaling exponent θ against the tail index ν. The dashed vertical lines delimit the intervals in which $\phi_\nu(\xi)$ [Eq. (10)] does not have finite integer moments ($\nu \leqslant 1$), has only finite mean ($1 < \nu \leqslant 2$), and has finite mean and variance ($\nu > 2$). The dashed horizontal line marks the $\theta = 1/2$ line. For some data points the error bars are smaller than the symbols shown.

distributions of step lengths of finite variance ($\nu > 2$), with a transient behavior in the interval $2 < \nu \leqslant 5/2$ that we attribute to the finite length of the random walks ($n \leqslant 10^7$ steps).

Figure 4 displays data collapses for our LIS data for some selected ν employing expression (5) or (6), depending on whether $\nu < 2$ or $\nu \geqslant 2$, respectively. The curve resulting from the data collapse corresponds to the empirical distribution $g(z)$ in Eq. (7). We see very good data collapses, all virtually with the same form for $g(z_n)$. We also see that it is definitely not the case that $g(z_n) \xrightarrow{n} \delta(z)$ [cf. Eqs. (12)–(14)].

III. THE LIS OF RANDOM WALKS OF FINITE VARIANCE

A. A wishful (but unlikely) connection with integer partitions

Recall that a partition of a natural number n is a sequence of integers $\lambda_1 \geqslant \cdots \geqslant \lambda_k > 0$ such that $\lambda_1 + \cdots + \lambda_k = n$. For example, (5,4,3) and (4,4,2,1,1) are two partitions of $n = 12$. No closed-form expression for the number $p(n)$ of partitions of n is known. Asymptotically, for $n \to \infty$ we have the Hardy-Ramanujan formula $p(n) \sim \exp(\pi\sqrt{2n/3})/4\sqrt{3}n$ [20].

As is well known, integer partitions play an important role in the solution of the LIS problem for random permutations [8–10]. In this case, the partitions carry the Plancherel measure given by $\mathbb{P}(\lambda) = (\dim \lambda)^2/n!$, where $\dim \lambda$ is the dimension of the irreducible representation of the symmetric group $\mathfrak{S}_n$ indexed by λ or, equivalently, the number of Young tableaux of shape λ. The correspondence between permutations and integer partitions (via the Robinson-Schensted-Knuth correspondence between permutations and Young tableaux) then allows one to identify the largest part of the partition λ with the length of the LIS of the original permutation.
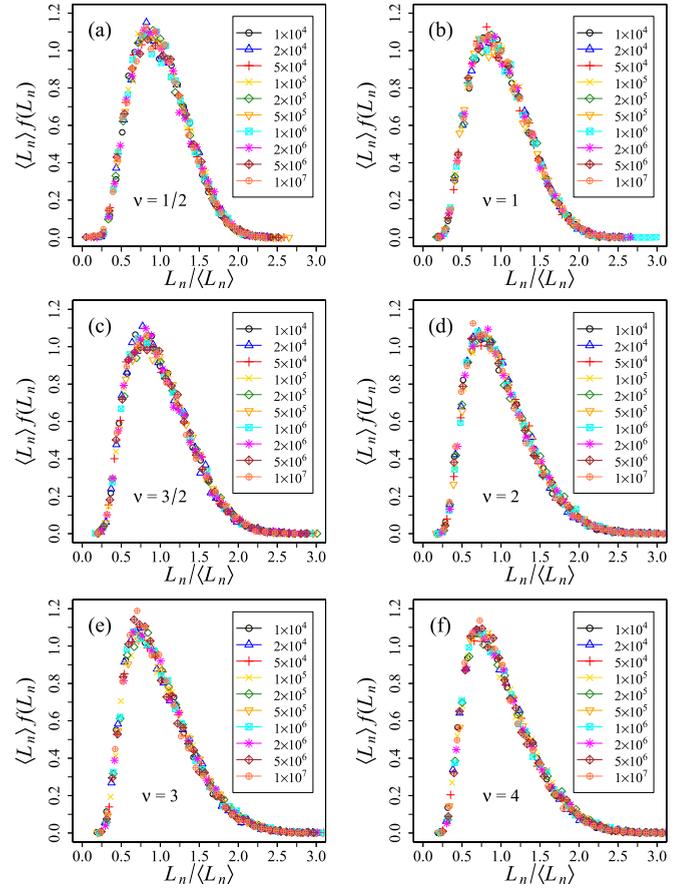
FIG. 4. Data collapses for some LIS data according to expressions (5) and (6). The upper panels (a) and (b) display data for the LIS of random walks with step lengths with both infinite mean and variance ($\nu \leqslant 1$), the middle panels (c) and (d) display data when the step increments have finite mean but infinite variance ($1 < \nu \leqslant 2$), and the bottom panels (e) and (f) display data obtained from random walks with step increments with both mean and variance finite ($\nu > 2$).

Clearly, other probability measures for random integer partitions have also been considered. Of particular interest to us is the uniform measure given by $\mathbb{P}(\lambda) = 1/p(n)$. This is so because the expected size of the largest part λ_1 of a partition of a large integer n drawn from the set of all partitions of n uniformly at random is given asymptotically by [21–26]

$$\mathbb{E}(\lambda_1) = \sqrt{\frac{n}{4\zeta(2)}} \Big[\ln n + 2\gamma_E - \ln \zeta(2) \Big] + O(\ln n), \quad (15)$$

where $\zeta(2) = \pi^2/6$ and $\gamma_E = 0.577\,215\ldots$ is Euler's constant. Equation (15) has, to leading and subleading order, the same functional form as the conjectured expression (6) for the asymptotic behavior of the length of the LIS of random walks with step lengths of finite variance. Moreover, the constant $1/\sqrt{4\zeta(2)} = 0.389\,848\ldots$ accompanying the leading term of Eq. (15) is close to the conjectured $1/e = 0.367\,879\ldots$ in (6). In the random partition model, however, the largest part λ_1 fluctuates, asymptotically for large n, like a Gumbel random
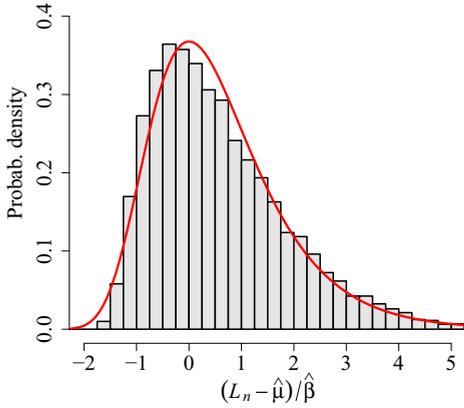
FIG. 5. Histogram of L_n obtained from 10^4 Gaussian random walks of 10^8 steps each, together with the standard Gumbel PDF (solid line). The LIS data are centered and normalized by the adjusted parameters $\hat{\mu}$ and $\hat{\beta}$, respectively [see Eqs. (17)–(19)].

variable with distribution

$$\mathbb{P}(\lambda_1 \leqslant \lambda) = F_G\left(\frac{\lambda - \sqrt{n/4\zeta(2)} \ln\left[n/4\zeta(2)\right]}{2\sqrt{n/4\zeta(2)}} \right), \quad (16)$$

where $F_G(z) = \exp[-\exp(-z)]$ [22–25]. We can thus check whether Eq. (15) makes sense in the context of the LIS problem for random walks beyond mere coincidence by checking whether our LIS data follow a Gumbel distribution.

The mean and variance of a Gumbel random variable with the PDF

$$f_G(x; \mu, \beta) = \frac{1}{\beta} \exp\left[-\left(\frac{x-\mu}{\beta}\right) - \exp\left(-\frac{x-\mu}{\beta}\right) \right] \quad (17)$$

are given by

$$\mathbb{E}(x) = \mu + \gamma_E \beta, \quad \text{Var}(x) = \frac{\pi^2}{6}\beta^2. \quad (18)$$

If we substitute the sample mean $\langle L_n \rangle = 69\,946$ for $\mathbb{E}(x)$ and the sample variance $\langle L_n^2 \rangle - \langle L_n \rangle^2 \simeq 8.399 \times 10^8$ for $\text{Var}(x)$ of a LIS data set obtained from 10^4 Gaussian random walks of $n = 10^8$ steps each, we obtain the following simple estimation for the parameters in (17),

$$\hat{\mu} \simeq 56\,903, \quad \hat{\beta} \simeq 22\,597. \quad (19)$$

We do not care about the uncertainties in $\hat{\mu}$ or $\hat{\beta}$ because they are relatively small and because the estimation procedure itself (the method of moments) is only approximate. While $\hat{\mu}$ (as well as $\langle L_n \rangle$) is not very far from the respective factor in Eq. (16) with $n = 10^8$, to wit, $\mu = \sqrt{n/4\zeta(2)} \ln[n/4\zeta(2)] = 64\,468$, the value of $\hat{\beta}$ differs significantly from $\beta = 2\sqrt{n/4\zeta(2)} = 7797$. Figure 5 displays the histogram of the LIS data together with a plot of $f_G(z) = F'_G(z) = \exp[-z - \exp(-z)]$. The fit looks good, but not excellent. In fact, the Gumbel distribution has a skewness of $12\sqrt{6}\zeta(3)/\pi^3 = 1.139\ldots$, while the data distribution has skewness $\simeq 0.976$ (irrespective of linear scaling). Whether this discrepancy is a finite-size effect is not clear at this moment. It should be remarked, however, that LIS data obtained from a uniform $U(-1, 1)$ distribution of step increments with other

values of walk length n provide the same overall picture as in Fig. 5 and seem to be nearly independent of n.

We leave the quantification of the "Gumbel hypothesis" to a future study employing more sophisticated density estimation techniques and hypothesis testing to tame uncontrolled wishful thinking [27]. In Sec. III B, however, we will discover that Eq. (15) cannot be easily discarded as a possible scaling form for the length of the LIS of random walks with step lengths of finite variance.

B. Constraining the range of the parameters in the asymptotic formula for $\mathbb{E}(L_n)$

The constants $1/e$ and $1/2$ appearing in the conjectured asymptotic formula (6) for the $\mathbb{E}(L_n)$ of random walks with distribution of step increments of finite variance were originally guessed in Ref. [15] based on linear least-squares fits to the data. Equation (15), with the same functional form, displays, respectively, the constants $1/\sqrt{4\zeta(2)} = 0.389\,848\ldots$ and $[2\gamma_E - \ln\zeta(2)]/\sqrt{4\zeta(2)} = 0.256\,025\ldots$. The proximity between the two sets of constants led us to reassess the numerical values of the constants appearing in (6) by a more refined approach.

Assuming that Eq. (7) holds for the distribution of L_n and that $\mathbb{E}(L_n)$ behaves asymptotically like [cf. Eq. (6) or (11a)]

$$\mathbb{E}(L_n) \sim a\sqrt{n} \ln n + b\sqrt{n}, \quad (20)$$

we can use the entire measured distributions to estimate the parameters a and b and give confidence intervals on their possible values. For a given pair (a, b) we scale the data points according to Eq. (20) and estimate the quality $S_{a,b}^{(c)}$ [superscript (c) for collapse, discussed below] of the resulting data collapse by a method first introduced in [28] and refined in [29] in the context of the finite-size scaling analysis of phase transitions. The method works by estimating the best master curve on which the data points for different sizes n should collapse. The quality $S_{a,b}^{(c)}$ is defined as the mean-square distance of the data points to the master curve in units of the standard error, similar to a χ^2 test. If the data points are on average one standard error away from the estimated master curve, the data collapse will have a quality of $S_{a,b}^{(c)} = 1$. Values $S_{a,b}^{(c)} \ll 1$ indicate that the standard errors are overestimated; values $S_{a,b}^{(c)} \gg 1$ indicate that the data points do not collapse within error bars, i.e., that the quality of the data collapse is bad. A data collapse of bad quality might be due to, besides the inevitable errors in the estimation of the master curve, also finite-size effects in the data and corrections to the functional form (20) itself.

In our case, care should be exercised in the application of the method because the scaling function $g(z)$ [see Eq. (7)] is insensitive to the multiplication of L_n, and thus $\mathbb{E}(L_n)$, by a nonzero factor, i.e., to any rescaling $(a, b) \to (ra, rb)$ by some (real) $r \neq 0$. This leaves the determination of the optimal (a, b) ill-defined. To fix this we compare the average value of each data set with the values predicted by Eq. (20) for every pair (a, b) tested. We then apply the same method as before to compute the quality figure $S_{a,b}^{(m)}$ [superscript (m) for mean] using Eq. (20) as the master curve with an added generous uncertainty of $\pm 0.05 \times \mathbb{E}(L_n)$ to account for
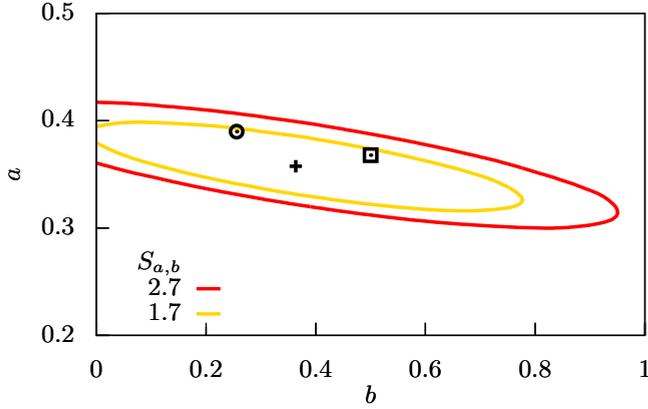
FIG. 6. Quality landscape $S_{a,b}$ obtained for random walks of $n = 2^{16}$, 2^{17}, 2^{18}, and 2^{19} uniformly, $U(-1, 1)$ distributed steps; 10^6 random walks were generated for each n. The square indicates the point $(1/e, 1/2)$ proposed in [15], the circle indicates the point $(0.389\,848\ldots, 0.256\,025\ldots)$ corresponding to Eq. (15), and the plus indicates the point $(0.36, 0.36)$ at which $S_{a,b}$ attains its minimum.

FIG. 7. Data collapse of the distributions with the optimal pair $(a, b) = (0.36, 0.36)$ that minimizes $S_{a,b}$. The inset shows the empirical averages $\langle L_n \rangle$ and their prediction according to Eq. (20) with the optimal constants $(a, b) = (0.36, 0.36)$.

finite-size effects and possible lower-order terms. Note that the minimum of $S_{a,b}^{(m)}$ corresponds to a standard least-squares fit.

The above-mentioned analyses were performed on data obtained from random walks of $n = 2^{16}$, 2^{17}, 2^{18}, and 2^{19} steps distributed according to a uniform $U(-1, 1)$ distribution; for each value of n, 10^6 sample random walks (and thus 10^6 data points L_n) are generated. Because data collapse is a matter of the form of a curve, we perform the collapse on $\ln f(L_n)$. This means that instead of the absolute standard errors σ of each data point we use the relative $\sigma / f(L_n)$ instead.

Figure 6 displays the contour plot of the composite quality factor $S_{a,b} = \frac{1}{2}(S_{a,b}^{(c)} + S_{a,b}^{(m)})$ for our data, which are collected by a scan through the (a, b) space in discrete steps of $\Delta a = 0.003$ and $\Delta b = 0.01$ of the aforementioned procedure. The best quality $S_{a,b}^{(\min)} \approx 0.7$ was achieved for $(a, b) = (0.36, 0.36)$. The equiquality lines at $S_{a,b}^{(\min)} + 1$ and $S_{a,b}^{(\min)} + 2$ can be roughly understood as 1σ and 2σ confidence intervals around the best quality $S_{a,b}^{(\min)}$ [28,29]. Note that the precise shape of the confidence intervals depends on the weighting of the terms in the composite quality; for instance, a larger weight on $S_{a,b}^{(m)}$, say, $\hat{S}_{a,b} = \frac{1}{4}(S_{a,b}^{(c)} + 3S_{a,b}^{(m)})$, leads to a further elongation of the equiquality loci. Figure 7 shows the data collapse as well as the least-squares fit which, in combination, yield the best quality. A further data collapse using high-precision data from Ref. [16] in the far tails of the distribution shows a different picture, with $S_{a,b}^{(\min)} = 81$ at $(a, b) = (0.33, 0.68)$. This data collapse, although still including the pair $(1/e, 1/2)$ within the second interval, is clearly of bad quality ($S_{a,b}^{(\min)} \gg 1$), probably because of strong finite-size effects, since we only have data in the far tails of the distributions for small walk lengths $n \leqslant 4096$. This data collapse is therefore not shown.

We see from Fig. 6 that both pairs $(a, b) = (1/e, 1/2)$, which was proposed in [15] based on least-squares fits of mean values $\langle L_n \rangle$ for the LIS of Gaussian random walks, and $(a, b) = (0.389\,848\ldots, 0.256\,025\ldots)$ of Eq. (15) lie
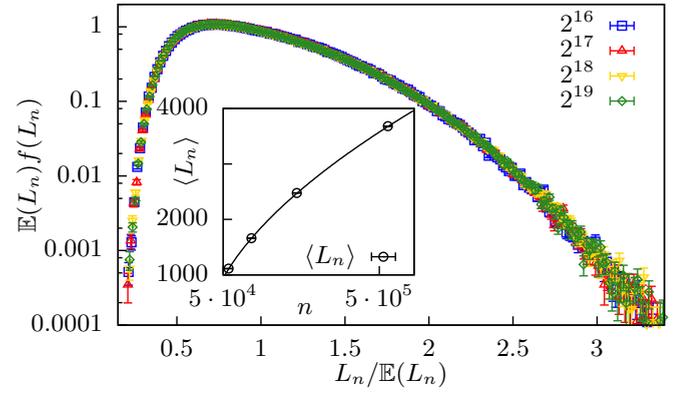
within the second confidence interval suggested by our method. Since the data sources are independent and even originate from different distributions of step increments, we see this as an argument in favor of the proposed scaling form (20).

IV. SUMMARY AND CONCLUSIONS

We have extended previous studies on the length L_n of the LIS of heavy-tailed random walks by considering Student's t-distributions with several different values of the parameter ν. We found that L_n scales like $\mathbb{E}(L_n) \sim n^\theta$ with a nonuniversal θ when $\phi(\xi)$ has infinite variance, but could not find a clear relationship $\theta = \theta(\nu)$ between these quantities besides the decreasing behavior $\theta'(\nu) < 0$. We were expecting to find something like $\theta(\nu) \sim (\nu - \nu_c)^z$ for some ν_c, possibly equal to 1 or 2, and some "critical exponent" z. Unfortunately, though, as Fig. 3 indicates (note the log-log scale), such a relationship would entail a varying exponent z. Whether this is a finite-n artifact we cannot tell. To seriously consider the hypothesis of a continuous phase transition for the behavior of the LIS between the $\nu > \nu_c$ and $\nu < \nu_c$ regions with θ as the order parameter one would probably have to simulate much longer random walks and possibly also larger sample sizes. Since we do not observe any strong change in the values of θ when increasing the sequence n by a factor of 10, the walks probably would have to be by a factor of many decades longer, much beyond the available computational resources.

When $\phi(\xi)$ is of finite variance ($\nu > 2$), we recover the asymptotic behavior given by (6), but with newly estimated constants. Our best current estimates for the constants appearing in expression (20), based on a sophisticated combined consideration of the behavior of the mean and of the scaling of the full distribution, are $a = 0.36(3)$ and $b = 0.36(30)$.

We could not obtain data for $\nu < 1/2$. The simulation of very-heavy-tailed random walks is complicated by the fact that one needs to add numbers of widely different orders of magnitude while keeping their full significance. This can be done with numerical libraries that implement arbitrary precision arithmetic, but the efficiency of the simulations suffers enormously. It would be desirable to compute the LIS

of heavy-tailed random walks in the $\nu \to 0$ limit to check how θ scales with ν in this limit and how it compares with the bounds in (4).

Our results about the scaling behavior of the mean LIS length, although useful on their own, can find applications, for instance, in the descriptive analysis of time series: If the LIS length grows stronger than expected, this is an indication that hidden trends or correlations are present in the data.

While the Plancherel distribution of the largest part of an integer partition coincides with the distribution of the length of the LIS of a uniformly distributed random permutation [8–10], the similarity between (6) and Eq. (15) does not imply any obvious relationship between the LIS of random walks and random integer partitions under the uniform measure. Our best estimated constants a and b for expression (20), however, cannot rule out Eq. (15) as a good candidate scaling form for the length of the LIS of random walks with step lengths of finite variance, and whether the LIS of these random walks follows a Gumbel distribution is open to debate. In a further study we intend to apply more refined density estimation techniques in the selection of an empirical model for the data; knowledge of the tail behavior, as provided by [16], is a valuable piece of information in this regard.

The elucidation of a possible combinatorial structure behind the LIS of random walks remains a tantalizing issue.

[1] S. Bespamyatnikh and M. Segal, Enumerating longest increasing subsequences and patience sorting, Inf. Process. Lett. **76**, 7 (2000).

[2] V. B. Nevzorov, *Records: Mathematical Theory* (AMS, Providence, 2001).

[3] 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).

[4] P. Gopalan, T. Jayram, R. Krauthgamer, and R. Kumar, Estimating the sortedness of a data stream, in *SODA'07 Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (SIAM, Philadelphia, 2007), pp. 318–327.

[5] L. Bonomi and L. Xiong, On differentially private longest increasing subsequence computation in data stream, Trans. Data Priv. **9**, 73 (2016).

[6] S. M. Ulam, in *Modern Mathematics for the Engineer: Second Series*, edited by E. F. Beckenbach (McGraw-Hill, New York, 1961), pp. 261–281.

[7] C. A. Tracy and H. Widom, Level-spacing distributions and the Airy kernel, Commun. Math. Phys. **159**, 151 (1994).

[8] J. Baik, P. Deift, and K. Johansson, On the distribution of the length of the longest increasing subsequence of random permutations, J. Am. Math. Soc. **12**, 1119 (1999).

[9] D. Aldous and P. Diaconis, Longest increasing subsequences: From patience sorting to the Baik-Deift-Johansson theorem, Bull. Am. Math. Soc. **36**, 413 (1999).

[10] D. Romik, *The Surprising Mathematics of Longest Increasing Subsequences* (Cambridge University Press, New York, 2015).

[11] O. Angel, R. Balka, and Y. Peres, Increasing subsequences of random walks, Math. Proc. Cambridge **163**, 173 (2017).

[12] R. Pemantle and Y. Peres, Non-universality for longest increasing subsequence of a random walk, ALEA Lat. Am. J. Probab. Math. Stat. **14**, 327 (2017).

[13] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon, Oxford, 1999).

[14] A. K. Hartmann, *Big Practical Guide to Computer Simulations* (World Scientific, Singapore, 2015).

[15] J. R. G. Mendonça, Empirical scaling of the length of the longest increasing subsequences of random walks, J. Phys. A: Math. Theor. **50**, 08LT02 (2017).

[16] J. Börjes, H. Schawe, and A. K. Hartmann, Large deviations of the length of the longest increasing subsequence of random permutations and random walks, Phys. Rev. E **99**, 042104 (2019).

[17] S. N. Majumdar and G. Schehr, Top eigenvalue of a random matrix: large deviations and third order phase transition, J. Stat. Mech. (2014) P01012.

[18] Wikipedia, Student's t-distribution, https://en.wikipedia.org/wiki/Student's_t-distribution, accessed 15 April 2019.

[19] R. W. Bailey, Polar generation of random variates with the t-distribution, Math. Comput. **62**, 779 (1994).

[20] G. H. Hardy and S. Ramanujan, Asymptotic formulæ in combinatory analysis, Proc. Lond. Math. Soc. **17**, 75 (1918).

[21] K. Husimi, Partitio numerorum as occurring in a problem of nuclear physics, Proc. Phys.-Math. Soc. Japan (3rd Ser.) **20**, 912 (1938).

[22] P. Erdős and J. Lehner, The distribution of the number of summands in the partitions of a positive integer, Duke Math. J. **8**, 335 (1941).

[23] B. Fristedt, The structure of random partitions of large integers, Trans. Am. Math. Soc. **337**, 703 (1993).

[24] A. M. Vershik and Y. Yakubovich, Fluctuations of the maximal particle energy of the quantum ideal gas and random partitions, Commun. Math. Phys. **261**, 759 (2006).

[25] A. Comtet, P. Leboeuf, and S. N. Majumdar, Level Density of a Bose Gas and Extreme Value Statistics, Phys. Rev. Lett. **98**, 070404 (2007); A. Comtet, S. N. Majumdar, and S. Ouvry, Integer partitions and exclusion statistics, J. Phys. A: Math. Theor. **40**, 11255 (2007).

[26] P. J. Grabner, A. Knopfmacher, and S. Wagner, A general asymptotic scheme for the analysis of partition statistics, Combin. Probab. Comput. **23**, 1057 (2014).

[27] P. Diaconis, Theories of data analysis: From magical thinking through classical statistics, in *Exploring Data Tables, Trends and Shapes*, edited by D. C. Hoaglin, F. Mosteller, and J. W. Tukey (Wiley, New York, 1985), pp. 1–36.

[28] N. Kawashima and N. Ito, Critical behavior of the three-dimensional $\pm J$ model in a magnetic field, J. Phys. Soc. Jpn. **62**, 435 (1993).

[29] J. Houdayer and A. K. Hartmann, Low-temperature behavior of two-dimensional Gaussian Ising spin glasses, Phys. Rev. B **70**, 014418 (2004).