J. Phys. A: Math. Gen. 27 (1994) L907–L911. Printed in the UK

LETTER TO THE EDITOR

Extremal segments in random sequences

Yacov Kantor† and Deniz Ertaş‡

†School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69 978, Israel
‡Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 30 August 1994

Abstract. We investigate the probability for the largest segment with total displacement Q in an N-step random walk to have length L. Using analytical, exact enumeration, and Monte Carlo methods, we reveal the complex structure of the probability distribution in the large-N limit. In particular, the size of the longest loop has a distribution with a square-root singularity at $\ell \equiv L/N = 1$, an essential singularity at $\ell = 0$, and a discontinuous derivative at $\ell = \frac{1}{2}$.

Investigation of the ground states of randomly charged polymers [1] suggests that in order to take maximal advantage of condensation energy and to diminish the effects of long-range repulsion of the excess charges, the polymer will select a necklace-like configuration, consisting of a few large, almost neutral globules, connected by narrow chains. In general this presents a complicated energy minimization problem. Some aspects of the solution can be determined by asking a simpler question: for a given random sequence (RS) of N charges, what is the length L of the *longest* segment with total charge Q? Alternatively, one can think of a one-dimensional random walk (RW) in which the longest segment with an end-to-end distance Q is to be found. The problem resembles certain classical RW problems [2], such as the problem of first and last arrival at a given point, or the special case of the last return to the starting point of the RW. However, the search for the longest segment of the RW, among all possible starting points, creates a more complicated problem. We combine Monte Carlo (MC) and exact enumeration studies, with some exact analytical results in certain simple limits, to demonstrate some remarkable properties of the distribution of the maximal-length segments.

A RS is defined as a sequence of N charges $\{q_i\}$ ($i = 1, \dots, N; q_i = \pm 1$), which is picked from the set of all such sequences with equal probability. (Since there are 2^N such sequences, the probability of picking a particular sequence is 2^{-N}.) Figure 1 depicts an example of the accumulated charge $S_i = \sum_{j=1}^{i} q_j$ for a RS ($S_0 \equiv 0$). Every segment of the sequence between, say, steps i and j, has a certain charge $Q = S_j - S_i$. Every such segment will be called a Q-segment. For a particular RS, consider the set of all Q-segments for a fixed value of Q. Our task is to find the length L of the largest segments among these. Figure 1 shows the longest 0-segments and the longest 4-segments, in a RS with $N = 24$. Clearly, the longest Q-segment does not have to be unique, and if there is at least one Q-segment in the sequence, $0 \leqslant L \leqslant N$. Let $P_N(L, Q)$ denote the probability that the longest Q-segment in a randomly chosen sequence of N charges has length L. Note that for $|Q| > 0$, the set of Q-segments in a given sequence may be empty: For example, the sequence shown in figure 1 has no 8-segments. Thus, $\sum_{L=0}^{N} P_N(L, Q) < 1$ for $|Q| > 0$.

0305-4470/94/240907+05$19.50 © 1994 IOP Publishing Ltd

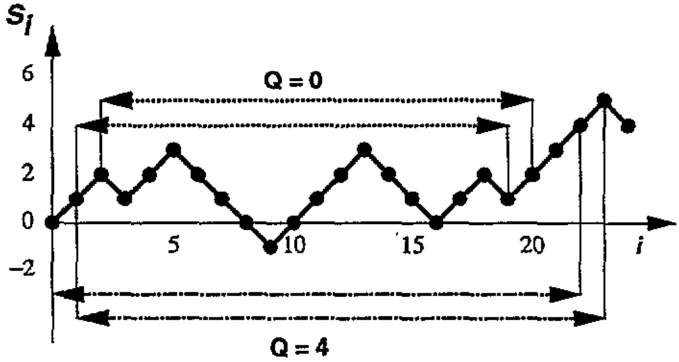

Figure 1. Example of a RS. In this case, the longest 0-segments have lengths $L = 18$ (dotted lines), while the longest 4-segments (chain lines) have lengths $L = 22$. There are no 8-segments.

Most properties of RSs have simple continuum limits. For example, the probability that an N element RS has overall charge Q_0 (for even $N + Q_0$) is

$$W_N(Q_0) = 2^{-N}\frac{N!}{[(N-Q_0)/2]![(N+Q_0)/2]!} \underset{N\to\infty}{=} \sqrt{\frac{2}{\pi N}}\exp(-Q_0^2/2N)\,. \tag{1}$$

Similarly, we expect $P_N(L, Q)$ to approach a simple form when $N, L, Q \to \infty$, while the reduced length $\ell \equiv L/N$ and the reduced charge $q \equiv Q/\sqrt{N}$ are kept constant. In this (continuum) limit it is more convenient to work with the *probability density* $p(\ell, q) = \frac{1}{2}N\,[P_N(L, Q) + P_N(L+1, Q)]$. (At most one of the two probabilities is non-zero, since $P_N = 0$ for odd $L + Q$.) In certain cases, P_N can be calculated exactly, especially for very small values of L and $N - L$, and for arbitrary Q. For example, $P_N(L = N, Q) = W_N(Q)$ follows from the definitions of P and W (the longest segment is the whole sequence itself), and a detailed analysis of all cases [3] gives $P_N(L = N-2, Q) = \frac{1}{4}\left[W_{N-4}(Q+2) + 2W_{N-4}(Q) + W_{N-4}(Q-2)\right]$. Similarly, one can find expressions for very small L [3]. However, we were unable to find a general expression for arbitrary N, L and Q. We performed exact enumeration studies of $P_N(L, 0)$ for $N \leqslant 36$. Results for few values of N are shown in figure 2(*a*). The results converge extremely quickly to the continuum distribution $p(\ell, 0)$. The full curve in the same figure depicts the results of a MC evaluation of the probability density from 10^8 randomly selected sequences of length $N = 1000$.

The probability density $p(\ell, 0)$ shown in figure 2(*a*) has several remarkable properties: (i) MC results show that p at $\ell = \frac{1}{2}$ is very close to unity (1.004 ± 0.006). At that point the slope of the curve changes by an order of magnitude. (ii) For $\ell \to 0$, the function exhibits an essential singularity of the form $\sim \ell^{-2}\exp(-B/\ell)$, where $B \approx 1.7$. (iii) For $\ell \to 1$, the function diverges as $(1-\ell)^{-1/2}$. Qualitatively, this behaviour can be understood as follows. The size of the longest 0-segment strongly depends on the charge $Q_0 = S_N$ *of the entire chain: when* $Q_0 \approx 0$, *the longest* 0-*segment typically has* $L \approx N$, while for very large Q_0, the longest 0-segment must be short. The definition of $p(\ell, 0)$ involves averaging over *all* RSs, and thus averaging over all Q_0 with their proper (Gaussian) weights. For simplicity, let us assume that the length L of the longest 0-segment depends only on Q_0. Then, for $Q_0 \ll \sqrt{N}$ we can relate $\ell = 1 - aQ_0^2/N$, where a is of order unity. On the other hand, for $Q_0 \gg \sqrt{N}$, the length of the longest 0-segment will be of order of a scale at which the random excursion of the RW becomes comparable to the drift produced by Q_0, i.e. when $L^{1/2} \approx LQ_0/N$, and therefore $\ell \approx N/Q_0^2$. By applying the relation

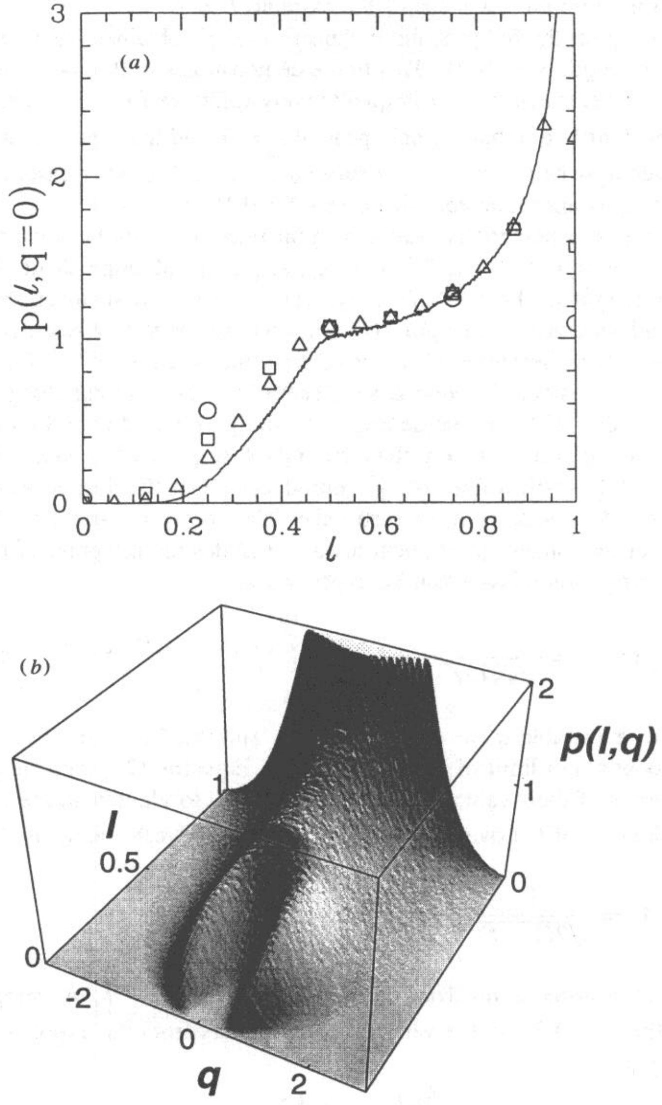

Figure 2. (*a*) Probability density of 0-segments as a function of reduced length ℓ. Circles, squares, and triangles depict the exact enumeration results for $N = 8$, 16 and 32, respectively. The full curve shows results of MC simulations (see text). (*b*) Probability density of Q-segments as a function of reduced charge q and reduced length ℓ. The results have been obtained from MC simulations (see text).

$p(\ell, 0) = (N/2)W_N(Q_0)|\mathrm{d}Q_0/\mathrm{d}\ell|$ in both limits, we correctly reproduce the square-root divergence for $\ell \to 1$, and the exp(constant/ℓ) singularity for $\ell \to 0$. (The leading pre-exponential power is not reproduced correctly in the latter case. A more involved argument [3] also reproduces this power correctly.) It is interesting to note that, by matching the asymptotic form of $p(\ell, 0)$ near $\ell = 1$ with $P_N(L, 0)$ for $L = N - 2$, we reproduce almost the exact value of the prefactor, i.e. the discrete distribution approaches its asymptotic (continuum) form within a few steps of the extreme $L = N$.

Figure 2(*b*) depicts the full probability density $p(\ell, q)$, obtained from a MC evaluation of 10^7 sequences of length $N = 1024$. This figure demonstrates further peculiarities of $p(\ell, q)$: for fixed ℓ, the q-dependence of p is qualitatively different for $\ell > \frac{1}{2}$ and $\ell < \frac{1}{2}$. In the former case, the distribution has a single peak at $q = 0$, and the areas $A_\ell \equiv \int_{-\infty}^{+\infty} \mathrm{d}q\, p(\ell, q)$ under fixed-ℓ sections have the form constant/$\sqrt{1-\ell}$. In the latter case, however, we see two peaks, and A_ℓ is approximately linear in ℓ for $0.15 < \ell < 0.5$.

An interesting and potentially useful integral relation exists between the probabilities $P_N(L, Q)$. By definition, there are $2^N P_N(L, Q)$ sequences of length N in which the longest Q-segment has length L. For $L \geqslant N/2$, we can construct all such sequences as follows. First, we take all sequences of length $2(N - L)$ whose longest Q'-segments are of length $N - L$, i.e., exactly half their total length. By definition, there are $2^{2(N-L)} P_{2(N-L)}(N - L, Q')$ of these. Next, we consider all sequences of length $2L - N$ and total charge $Q - Q'$. There are $2^{2L-N} W_{2L-N}(Q - Q')$ such sequences. It is straightforward to show that inserting any chain from the second group into any chain from the first group *at its midpoint*, and repeating this process for all possible values of Q', reproduces all of the desired sequences, without constructing the same sequence more than once. The necessary condition, however, is that $L \geqslant N/2$, so that the longest Q-segment always includes the midpoint of the sequence. In the continuum limit, this relation can be expressed as

$$p(\ell, q) = \frac{1}{\sqrt{4\pi(2\ell - 1)(1 - \ell)}} \int_{-\infty}^{+\infty} \mathrm{d}q'\, \mathrm{e}^{-(q - q'\sqrt{2(1-\ell)})^2/2(2\ell-1)}\, p\left(\tfrac{1}{2}, q'\right) \qquad (2)$$

where the reduced variable $q' = Q'/\sqrt{2(N - L)}$, and the Gaussian term in the integrand represents the continuum limit of $W_{2L-N}(Q - Q')$. Equation (2) gives the probabilities for any $\ell \geqslant \frac{1}{2}$ in terms of their value at $\ell = \frac{1}{2}$, and reduces to identity in the $\ell \to \frac{1}{2}$ limit. By integrating both sides of (2) over q, we find a relation between the areas A_ℓ, for $\ell > \frac{1}{2}$:

$$A_\ell = \frac{1}{\sqrt{2(1 - \ell)}} \int_{-\infty}^{+\infty} \mathrm{d}q\, p\left(\tfrac{1}{2}, q\right) \qquad (3)$$

which confirms the observation from the MC data that for $\ell > \frac{1}{2}$, A_ℓ simply increases as $1/\sqrt{1-\ell}$. In the $\ell \to 1$ limit, the variable q' disappears from the exponent in (2), and the relation reduces to

$$p(\ell \to 1, q) = \frac{A_{1/2}}{\sqrt{4\pi(1 - \ell)}}\, \mathrm{e}^{-q^2/2} \,. \qquad (4)$$

This relation both confirms our contention that $p(\ell, 0)$ has a square-root divergence $A/\sqrt{\pi(1-\ell)}$ with $A \equiv \frac{1}{2} A_{1/2}$, and demonstrates that the fixed-ℓ sections of the surface in figure 2(*b*) approach a pure Gaussian shape when $\ell \to 1$ as expected, since p behaves like W in this limit. In addition to the MC study, we performed an exact enumeration study to determine A for sequences with $N \leqslant 30$, and found that it extrapolates to the value 1.011 ± 0.001, in perfect agreement with the MC result: definitely larger than unity, but surprisingly close to it.

We did not find analogous integral relations for $\ell < \frac{1}{2}$. Here, the situation is complicated by the fact that, in a given sequence, there may be several longest Q-segments that are disjoint. The q-dependence of $p(\ell, q)$ for small values of ℓ has a minimum at $q = 0$. The minimum disappears as ℓ increases, at $\ell \approx \frac{1}{2}$. Further analysis is necessary to understand the behaviour of $p(\ell, q)$ in this region.

In conclusion, we have demonstrated that the probability density $p(\ell, q)$ has some peculiar and unexpected properties and very rich behaviour, despite the apparent simplicity of its formulation, and its similarity to classical RW problems. More analysis is needed to fully understand various properties of the extremal segments in a RS.

We thank M Kardar for helpful discussions. This work was supported by the US–Israel BSF grant no 92–00026, and by the NSF through grant nos DMR–87–19217 (at MIT's CMSE) and DMR 91–15491 (at Harvard).

References

[1] Kantor Y and Kardar M 1994 Excess charge in polyampholytes *Europhys. Lett.* at press; 1994 Instabilities of charged polyampholytes *Phys. Rev.* E submitted

[2] Chandrasekhar S 1943 *Rev. Mod. Phys.* **15** 1

[3] Ertaş D and Kantor Y to be published