Tr. J. of Physics
23 (1999) , 105 – 112.

Studies on Extremal Segments in Random Sequences

Deniz ERTAŞ *
Department of Physics, Harvard University, Cambridge, Cambridge, MA 02138, U.S.A.
Yacov KANTOR
School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69 978, ISRAEL

Received ...

Abstract

We review our main findings on the size distribution of the largest neutral segments in a sequence of N randomly charged monomers. Upon mapping to one–dimensional random walks (RWs), this corresponds to finding the probability distribution for the size L of the largest segment that returns to its starting position in an N–step RW. 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 neutral segment 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 $\ell = 1/2$.

1. Introduction

It has recently been shown that ground state conformations of polyampholytes, a particular type of heteropolymers built with a random mixture of positively and negatively charged groups along their backbone, are extremely sensitive to the their total (excess) charge Q. A detailed study of the Q–dependence of the radius of gyration R_g[1, 2] determined that a reasonable compromise between stretching (which minimizes the electrostatic energy) and remaining compact (which gains in condensation energy) is for the polyampholyte to form a *necklace* of weakly charged blobs connected with highly charged "necks", by taking advantage of the charge fluctuations along the chain. The results of Monte Carlo[1] and exact enumeration[2] studies qualitatively support such a picture.

While the exact treatment of electrostatic interactions is not possible, we can pose a simplified problem which, we hope, captures some essential features of this necklace model. For example, we may ask what the typical size of the largest neutral (or weakly charged) segment in a random sequence of N charges will be, since this segment is likely to collapse into a compact segment and significantly influence the final configuration of

*Present Address: Exxon Research & Engineering, Rte 22 East, Annandale, New Jersey 08801

the polyampholyte. In order to answer this question, we investigated the size distribution of the largest neutral segments in polyampholytes with N monomers ($N-$mers). This problem can be mapped to a one-dimensional random walk (RW): the sequence of charges $\{q_i\}$ $(i = 1, \ldots, N;\ q_i = \pm 1)$ corresponds to an N–step walk $\omega \equiv \{q_1, \cdots, q_N\}$ with the same sequence of unit steps in the positive or negative directions along an axis, where the probability of going up or down is equal to 1/2 at each step. Fig. 1 depicts an example of such a sequence and the corresponding path, where $S_i(\omega) = \sum_{j=1}^{i} q_j$ is the position of the path at index i. ($S_0(\omega) \equiv 0$.) A segment of L monomers with zero total charge thus corresponds to an L–step loop inside the RW. In this paper, we review our findings on the properties of the probability $P_N(L)$ that the *largest* loop in an N–step RW has length L, or, equivalently, the probability $Z_N(L) = \sum_{L'=0}^{L-1} P_N(L')$ that all loops in an N–step RW are shorter than L. A complete account of the results about a generalized version of this and other related problems can be found in Refs. [3, 4, 5].

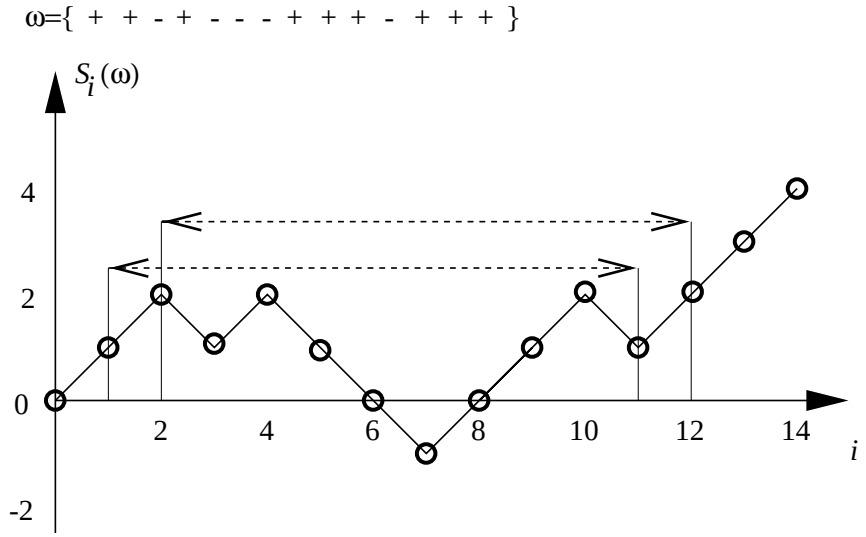

Figure 1. Example of a sequence ω with $N = 14$ charges, and the corresponding walk depicted by $S_i(\omega)$. In this case, the longest loops have lengths $L = 10$ (dotted lines).

In the continuum ($N \to \infty$) limit, it is more convenient to work with the *probability*

density

$$p(\ell) \equiv \frac{N}{2}\left[P_N(L) + P_N(L+1)\right] \tag{1}$$

and

$$z(\ell) = \int_0^\ell d\ell' p(\ell'), \tag{2}$$

where $\ell = L/N$ is the appropriate scaling variable for this problem.

There is an apparent simplicity of the formulation of the problem, i.e. it is similar (and related) to the classical RW problems[6], such as the problem of first passage times or the problem of last return to the starting point, for which probability distributions can be computed exactly by using the method of reflections[7], and obey the same scaling in the continuum limit. However, the search for the *longest* loop of the RW, among all possible starting points, creates a more complicated problem. In its essence, the problem is more related to the statistics of self–avoiding, rather than regular, random walks. This relation becomes more transparent in the $\ell \to 1$ and $\ell \to 0$ limits. The "self–interacting nature" of the problem can be seen even more clearly in its generalizations to arbitrary space dimension d, where many analogies between this problem and the self–avoiding walks exist[4].

Our initial investigations[3] revealed remarkable properties of the probability density $p(\ell)$: It diverges as $p(\ell) \sim 1/\sqrt{1-\ell}$ for $\ell \to 1$, which we have been able to confirm analytically, and has a discontinuous derivative at $\ell = 1/2$. Furthermore, it has an essential singularity at $\ell = 0$ of the form $p(\ell) \sim (BC/\ell^2)\exp(-B/\ell)$. An analytical solution in this limit still remains elusive, but in a followup work[5] we have developed an improved Monte Carlo algorithm that is capable of probing significantly smaller values of ℓ numerically. Combined with strict analytical bounds on $z(\ell)$ [cf. Eq.(2)], the results strongly favor the existence of this singularity, and the proper form of the $\ell \to 0$ limit can be determined with high precision. Unfortunately, the behavior near $\ell = 1/2$ still remains a mystery.

It should be noted that similar behavior is exhibited by extremal properties of a number of random processes, such as a one-dimensional random cutting process[8] (which can be generalized to higher dimensions[9]) and return times in a random walk[9]. These models exhibit singularities at $\ell = 1/k$, which become progressively weaker as the integer k is increased, leading to an essential singularity at $\ell = 0$. Although it was claimed that our problem falls into the same category and therefore should exhibit singularities at $\ell = 1/2, 1/3, 1/4, \cdots$[9], we believe that it differs from these models in a way that undermines the reasoning for this claim. In particular, we have numerically verified that the suggested singularity at $\ell = 1/3$ does not exist, unless it has a very small prefactor[5].

2. Numerical Findings

We have initially examined the behavior of $P_N(L)$ using numerical (exact enumeration and Monte Carlo) methods. Monte Carlo results obtained for a variety of large Ns up to $N = 10^4$ were virtually indistinguishable from each other when plotted in the properly scaled variables. The results for $N = 1000$ are depicted as a solid curve in each one of the graphs in Fig. 2. For that particular value of N we evaluated the probability density from 10^8 randomly selected sequences. For short chains (up to $N = 36$) it was possible to perform a complete enumeration and get the exact results for $P_N(L)$. When these exact results are plotted in the scaled form, as presented in Fig. 2, we can see that even for such modest values of N, there is an extremely fast convergence to the continuum distribution $p(\ell)$, depicted by the solid curve (especially for $\ell > 0.5$).

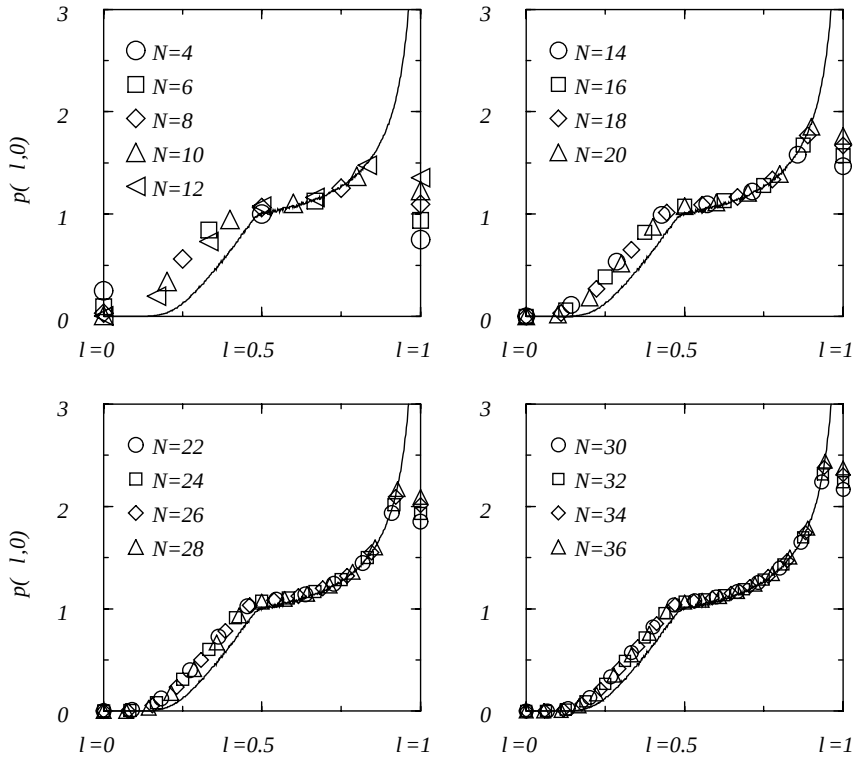

Figure 2. Probability density of largest neutral segments as a function of reduced length $\ell = L/N$. Symbols depict exact enumeration results for N up to 36. In each graph, the solid line shows the MC evaluation of $p(\ell, 0)$ from 10^8 randomly selected sequences of length $N = 1000$.

The probability density $p(\ell)$ shown in Fig. 2 has several remarkable properties:
(a) 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. While it is impossible to ascertain from the numerical results that there is actually a discontinuity in the first derivative of $p(\ell)$ with respect to ℓ, both the MC results and analytical arguments indicate that $\ell = \frac{1}{2}$ is a special point of the curve.
(b) For $\ell \to 0$, the function exhibits an essential singularity of the form

$$p(\ell) \sim \frac{BC}{\ell^2} \exp(-B/\ell), \qquad \ell \ll 1, \tag{3}$$

where $B = 1.73 \pm 0.02$ and $C = 4.57 \pm 0.01$. These precise estimates have been obtained by a finite size scaling analysis of data from a special recursive MC algorithm, designed to significantly enhance efficiency for small values of ℓ[5]. Figure 3 depicts a sample plot obtained from this improved MC procedure, which can measure probabilities as small as 10^{-15}, and confirms the asymptotic form (3). The existence of the singularity can be understood from the fact that for small ℓ, the absence of large loops in the entire chain can be thought of as a requirement that such loops are absent in many separate and independent segments of the sequence. In Section 3, this argument will be discussed in more detail.
(c) For $\ell \to 1$, $p(\ell, 0)$ diverges as $A/\sqrt{\pi(1-\ell)}$, with $A = 1.011 \pm 0.001$. In Ref. [4] we proved the existence of the square–root singularity from an integral relation. The proof, however, does not provide a value for the prefactor A. This particular estimate of the constant A has been obtained from a finite size scaling analysis of the numerical results for a different quantity that is related to A through an exact relation[4]. Two additional independent numerical estimates of A are consistent with this result, with somewhat larger error bars.

3. Analytical Findings

In this section, we outline the derivation of rigorous upper and lower bounds on the probability distribution $z(\ell)$, both of which have the same functional form, in order to provide a flavor of the typical nature of analytical arguments used throughout this study. The existence of these bounds significantly restrict possible asymptotic forms of $z(\ell)$ in the $\ell \to 0$ limit.

The main strategy is the similar for establishing both upper and lower bounds. Walks whose largest loops are much smaller than their overall length are typically very biased in one direction, and sections of the walk that are separated by more than the largest loop size are very weakly correlated. For a given (small) value of ℓ, let us divide each walk into roughly $1/\ell$ segments of similar size. There are *necessary* conditions that each segment must satisfy independently for the overall walk to contribute to $z(\ell)$. If the probability for a random segment to satisfy these conditions is p_n, then $z(\ell) < p_u^{1/\ell}$. Similarly, each segment can be designed to satisfy certain conditions that are *sufficient* to ensure that

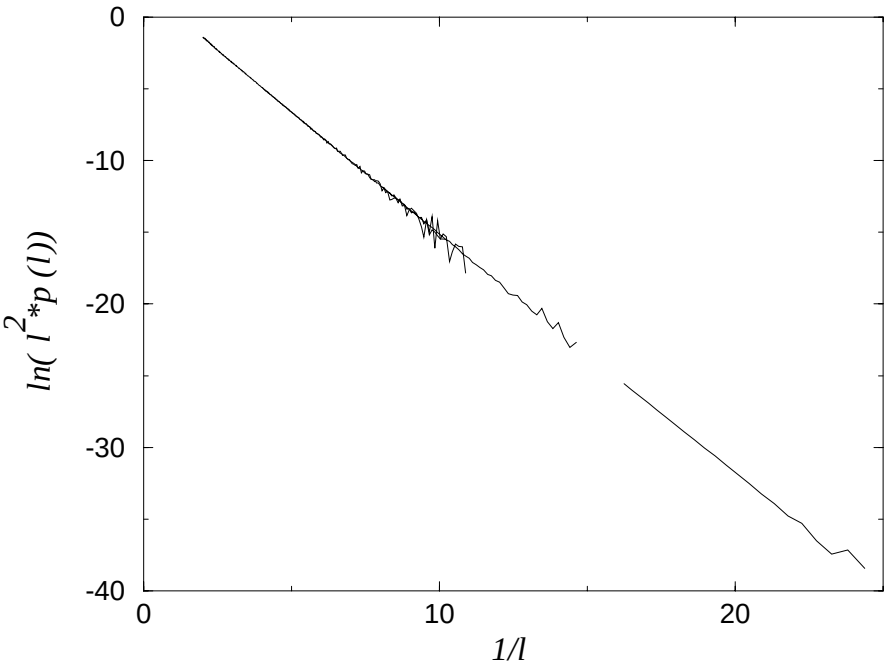

Figure 3. The probability density $p(\ell)$ for $0.04 < \ell < 1/2$ confirms the suggested form (3) down to probabilities as low as 10^{-15}. The overall walk size is $N = 2048$.

the overall walk contributes to $z(\ell)$. If the corresponding probability for these conditions is p_s, then $z(\ell) > p_s^{1/\ell}$.

Let us first investigate necessary conditions. It can easily be shown that for a sequence with positive total displacement, each of the $m \approx N/L$ segments need to have a positive displacement for that sequence to contribute to $p(\ell)$. The probability for this is just $p_n = 1/2$, and therefore $Z_N(N/m) < 2^{1-m}$ (the additional factor of 2 comes from RWs with $S_N < 0$). Consequently, $Z_N(L) < 2^{2-(N/L)}$ for any value of N and L. This establishes a strict upper bound, which is significant for small values of ℓ:

$$z(\ell) < 4\exp(-\ln 2/\ell). \tag{4}$$

It is possible to further improve this upper bound to

$$z(\ell) < 4\sqrt{2}\exp\left(-\frac{3\ln 2}{2\ell}\right), \tag{5}$$

which makes the best (so far) analytical lower bound on the exponential factor $B > 3\ln 2/2 \approx 1.03972$.

In order to find a lower bound on the probability distribution, let us again consider the sequence ω and its m pieces $\{\omega_i\}$ of length L each. We'd like to construct each ω_i independently in such a way to guarantee that the resulting walk ω does not have loops larger than L. One such specification is[5]

$$\begin{cases} -\alpha < S_i < S_L - \alpha, & 0 < i \le L/2, \\ \alpha < S_i < S_L + \alpha, & L/2 < i \le L, \end{cases} \tag{6}$$

where a near-optimal value of α is $\sqrt{L}/2$, for which $p_s \approx 0.031585$. This yields

$$z(\ell) > 2p_s \exp(-\ln p_s/\ell) \approx 0.06317e^{-3.455/\ell}. \tag{7}$$

Clearly, neither the upper nor the lower bounds we have established are very tight, and they do not rule out the possibility of a power-law prefactor. However, there is very convincing numerical evidence that there is no power law prefactor in $z(\ell)$, i.e. that $\lim_{\ell\to 0} z(\ell) = C\exp(-B/\ell)$ where C and B are constants that can be determined numerically.

4. Discussion

The problem of extremal segments originated from the desire to consider a simplified description of the ground states of randomly charged polymers. We used MC, exact enumeration and analytical techniques to analyze the problem, and our results provide convenient tools for a semi–quantitative analysis of the the ground states of PAs. In particular, we show that a "typical" RS contains very large neutral segments, i.e. it is possible to construct a ground state from a single very large blob with relatively short ends of the chain dangling outside the blob.

Besides the original motivation, the problem of extremal segments is interesting in its own right. It looks like one of the classical problems of random walks and, nevertheless, is highly non–trivial, and the results indicate a solution with very rich and unexpected structure. While several features of the problem have been established analytically, we did not find a complete analytical solution of the problem. We think that such a solution is possible and further attempts of finding it are worthwhile.

The numerical "proof" of the continuum limit in our work was limited to a particular class of RWs, in which a unit displacement appears at each step. Within that class we presented evidence of a continuum limit where the properly scaled functions become independent of N. Preliminary results within a slightly broader class of RWs, in which the size of the step has a binomial distribution, indicate that the same universal curves are attained even within this broader class of RWs. It may be possible to prove the universality of the continuum limit by attempting to perform a renormalization–group–like treatment of the problem, i.e. attempting to define the problem in the limit where the RW becomes a true Gaussian walk (walk of idealized Brownian particle). This limit,

however, is far from being trivial. In particular the definition of what is called a loop (i.e. how close two different points of the walk should be located so that the segment will be called a closed loop) presents a non–trivial problem in the continuum limit. Such short distance scale can undergo a non–trivial scaling, similarly to the excluded volume parameter in the treatment of self–avoiding walks. A different approach to the question of universality may begin from an expansion of the solution near the dimension $d = 4$, as in the treatment of self–avoiding walks.

This work was supported by the National Science Foundation, by the MRSEC program through Grant DMR-9400396 and through Grants DMR-9106237, DMR-9417047 and DMR-9416910; and by the Israel Science Foundation founded by The Israel Academy of Sciences under Grant No. 246/96.

References

[1] Y. Kantor and M. Kardar, Europhys. Lett. **27**, 643 (1994), and Phys. Rev. **E51**, 1299 (1995).

[2] Y. Kantor and M. Kardar, Phys. Rev. **E52**, 835 (1995).

[3] Y. Kantor and D. Ertaş, J. Phys. A: Math. Gen. **27**, L907 (1994).

[4] D. Ertaş and Y. Kantor, Phys. Rev. E**53**, 846 (1996).

[5] D. Ertaş and Y. Kantor, Phys. Rev. E**55**, 261 (1997).

[6] S. Chandrasekhar, Rev. Mod. Phys. **15**, 1 (1943).

[7] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd Ed., John Wiley and Sons, New York (1968).

[8] B. Derrida and H. Flyvbjerg, J. Phys. **A20**, 5273 (1987).

[9] L. Frachebourg, I. Ispilatov and P. L. Krapivsky, Phys. Rev. E**52**, R5727 (1995).