Journal of Statistical Physics, Vol. 126, No. 1, January 2007 (© 2006)
DOI: 10.1007/s10955-006-9247-z

String Matching and 1d Lattice Gases

Muhittin Mungan[1,2]

Received December 6, 2004; accepted May 3, 2006
Published Online: December 28, 2006

We calculate the probability distribution for the number of occurrences n of a given l letter word $\mathbf{x}$ inside a random string of k letters, whose letters have been generated by a known stationary stochastic process. Denoting by $p(\mathbf{x})$ the probability of occurrence of the word, it is well-known that the distribution of occurrences in the asymptotic regime $k \to \infty$ such that $kp(\mathbf{x}) \gg 1$ is Gaussian, while in the limit $k \to \infty$, and $p(\mathbf{x}) \to 0$, such that $kp(\mathbf{x})$ is finite, the distribution is Compound Poisson. It is also known that these limiting forms do not work well in the intermediate regime when $kp(\mathbf{x}) \gtrsim 1$ and k is finite. We show that the problem of calculating the probability of occurrences is equivalent to determining the configurational partition function of a 1d lattice gas of interacting particles, with the probability distribution given by the n-particle terms of the grand-partition function and the number of particles corresponding to the number of occurrences on the string. Utilizing this equivalence, we obtain the probability distribution from the equation of state of the lattice gas. Our result reproduces rather well the behavior of the distribution in the asymptotic as well as the intermediate regimes. Within the lattice gas description, the asymptotic forms of the distribution naturally emerge as certain low density approximations. Thus our approach which is based on statistical mechanics, also provides an alternative to the usual statistics based treatments employing the central limit and Chen–Stein theorems.

KEY WORDS: Pattern occurrences, combinatorics, lattice gases, theory of liquids
PACS Nos: 02.10.Ox,05.70.Ce, 2.50.-r.

1. INTRODUCTION

The problem of determining the probability of encountering (matching) a given word $\mathbf{x}$ of length l in another string of length k, whose letters have been drawn randomly from an alphabet of r letters, has a variety of applications ranging from

[1] Boğaziçi University, Department of Physics, 34342 Bebek, Istanbul, Turkey; e-mail: mmungan@boun.edu.tr.
[2] Gürsey Institute, P. O. B. 6, Çengelköy, 34680 Istanbul, Turkey.

designing fast algorithms for pattern searching,[1,2] to problems in genetics of assessing the likelihood of events such as the frequency of occurrence of DNA segments,[3,4] or that certain DNA segments align.[5,6] In each of these cases the likelihood estimates for random sequences can be used as a benchmark against which to evaluate the statistical significance of actually observed events.

The problem is non-trivial, because of the possibility of overlapping occurrences in the string, which introduce correlations that cannot be neglected. In a first and crucial step towards the solution of this problem, Guibas and Odlyzko?[7–9] derived the moment generating functions associated with the probability for not encountering a given set of words in a random string. The resulting distributions turn out to depend on a set of correlation functions that capture the overlap properties of the words with each other.

Building on the work of Guibas and Odlyzko, several authors have studied the probability distribution for the number of occurrences n of of a given l letter word in a random string of k letters, under various assumptions on the distribution of random letters[3,4,10–19]: The cases where the letters of the random string are independently and identically distributed (i.i.d.) was treated by Chrysaphinou and Papastavridis[10,11] and later by Fudos *et al.*[12] The case where the letter distribution follows the steady state distribution of a Markov process has been investigated by several authors.[13–15,17,19] The distributions are obtained in asymptotic regimes when k is large along with various assumptions on the probability of occurrences of the word, where methods of statistics such as the central limit theorem leading to Gaussian distributions,[4,12,15] the theory of large deviations,[15,19] or the Chen-Stein Method leading to (compound) Poisson distributions[11,13,14,16,18] are applicable.

The regimes of applicability can be difficult to identify. It has been noted that, even in the case of i.i.d. letters, when the length l of the word to be matched is fixed, and assuming that k is large, the approximate distribution that captures the actual distribution best (*e.g.* gaussian vs. compound Poisson), still depends on the word whose occurrence is sought.[20] As an example, take a binary random string $\mathbf{y}$ of length $k = 256$ whose letters y_i are i.i.d. distributed with uniform probability $\text{Prob}\{y_i = 0\} = \text{Prob}\{y_i = 1\} = 1/2$ and consider the distribution of occurrences n for the following words of length 4: $\mathbf{x} = 0001$, $\mathbf{x} = 1001$, $\mathbf{x} = 0101$, and $\mathbf{x} = 1111$. The resulting distributions are shown in Fig. 1. The solid circles are the exact probabilities, the dashed line corresponds to the Gaussian approximation,[21] while the dot-dashed line is the Compound Poisson approximation.[13,14,18] Note that the Gaussian and Compound-Poisson distributions approximate the true distribution better only for some strings, but less well for others, as remarked above. The solid line on the other hand is the analytical result of this article that has been obtained by observing that the probability distribution function can be regarded as the (configurational) partition function of an interacting 1d lattice gas and using techniques of statistical mechanics to obtain an approximate distribution. Note in

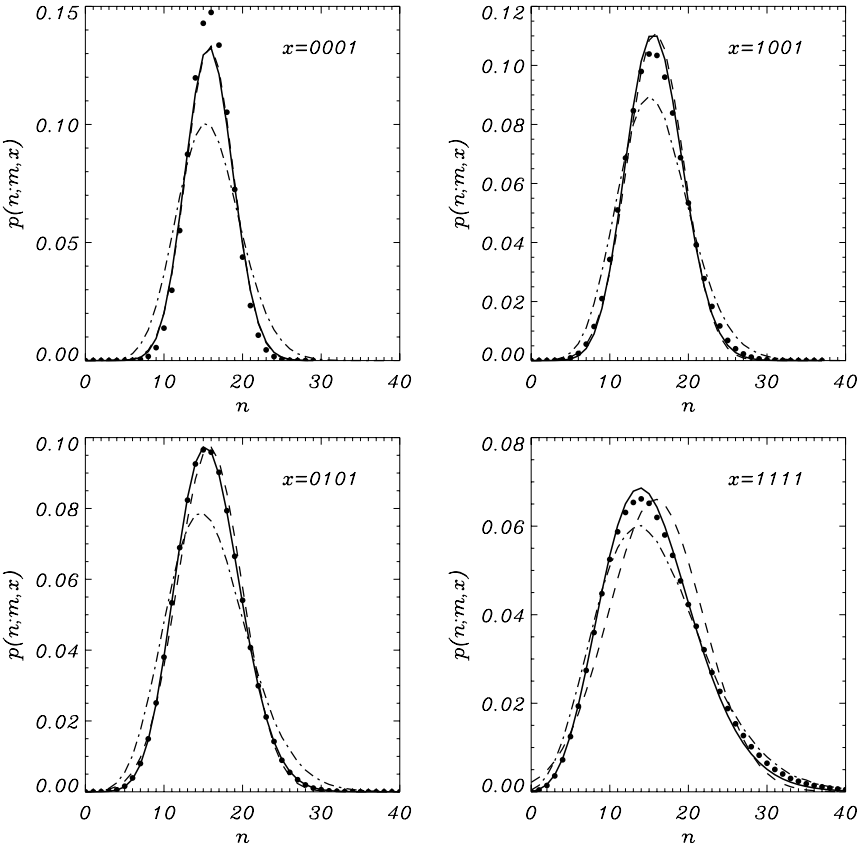

Fig. 1. The distribution of occurrences n of a $l = 4$ letter binary word x inside a random string of length $k = 256$, with uniform i.i.d. distributed letters, $p(\text{“1”}) = p(\text{“0”}) = 1/2$. Shown are the distributions for $x = 0001$ (top left), $x = 1001$ (top right), $x = 1010$ (bottom left) and $x = 1111$ (bottom right). The circles are the exact probabilities, the dashed and dashed-dotted lines correspond to the Gaussian and Compound Poisson approximation (see text for details). The solid line is the analytical result of this paper.

particular how closely the solid line follows the shape of the actual distribution as **x** changes.

We should stress that exact analytical expressions for the probability distribution can be readily written down in the form of either a generating function,[15,19] or a set of recursions.[22] However, since these expressions are implicit, i.e. they do not specify the probability distribution in a closed-form, they are of limited use, particularly if one wishes to understand the interplay between the resulting distribution and its dependence on both the word whose occurrences is sought, as well as the properties of the underlying stochastic process generating the random

letter sequence. The main goal of all of the approaches cited above is to obtain tractable, even if approximate, analytical expression that capture the essentials of this interplay. As remarked above, this has been achieved only for the asymptotic regimes, where the resulting distributions turn out to be Gaussian or Compound Poisson. It is therefore desirable to obtain a single, analytical expression for the probability distribution that captures the behavior of the exact distribution in a broader and not necessarily asymptotic regime, and to recover the asymptotic forms as special limiting cases. Besides the obvious advantage of having a single description, such an approach will naturally identify the regimes of applicability of the various asymptotic approximations, while also pointing out when and how they fail. Our main goal is to show that such an analytical treatment is possible and useful.

The article is organized as follows: After introducing in Sec. 2 the necessary definitions and, rederiving the exact expression for the probability distribution, we show in Sec. 3 that the problem of calculating the probability distribution for the number of occurrences n of a given l letter word in a random string of k letters, is equivalent to the problem of calculating the configurational terms of the grand partition function of a 1d lattice gas. In this mapping the number of particles correspond to the number of occurrences, the "volume" of the gas is the length of the random string, and the correlations between subsequent occurrences turn into pairwise interactions whose nature depends on both the properties of the word to be matched, as well as the underlying stationary stochastic process generating the random string. Thus the dependence of the probability distribution on these parameters is essentially captured in the resulting form of the particle interactions. In Sec. 4 we look more closely at these interactions and show that common to them is a relatively strong and short regime of length l, the core of the interaction, that is followed by a weaker and exponentially decaying tail.

From the equation of state of the lattice gas we obtain in Sec. 5 an analytical expression for the probability distribution that besides reproducing the known asymptotic limits, is also applicable in the intermediate regime, where these asymptotic forms cease to be good approximations, as apparent in Fig. 1. In Sec. 6 we turn to the asymptotic behavior of the distribution and show that the Gaussian asymptotic form can be obtained as the thermodynamical limit of a low density approximation in which the interactions are overall weak, whereas the Compound-Poisson form emerges in a regime where the core of the interaction is prominent and dominates over the weaker tail. We conclude the article with a discussion of our results in Sec. 7. An Appendix contains additional details on some of the results.

We should point out that our method is similar in spirit to recent statistical mechanics based approaches to combinatorial problems such as k-SAT.[23–28]

2. DEFINITIONS

In this section we present the known[15,19] expression for the probability of n occurrences of a given word $\mathbf{x}$ inside a random string. This will form the starting point for the lattice gas approach to be taken up in the following Section.

We denote by $\mathbf{x} = x_1, x_2, x_3, \ldots, x_l$ the word of length l whose occurrence is sought and by $\mathbf{y} = y_1, y_2, y_3, \ldots, y_k$ the random letter string of length k, whose letters have been drawn from an r letter alphabet $\mathcal{A} = \{0, \ldots, r-1\}$. Since $k \geq l$, it is useful to define the excess length $m = k - l$.

The random letters will be assumed to be either independently and identically distributed with probability $p(x)$, such that $\sum_{x\in\mathcal{A}} p(x) = 1$ (M0 model), or according to an s-order Markov chain on $\mathcal{A}$ with transition probabilities

$$\pi(y_1, y_2, \ldots, y_s; y_{s+1}) = \text{Prob}\{Y_i = y_{s+1} | Y_{i-s} = y_1, \ldots Y_{i-1} = y_s\} \tag{2.1}$$

and ergodic stationary distribution $\mu(y_1, y_2, \ldots, y_s)$ (Ms Model). The i.i.d. case can be regarded as a 0-th order Markov Chain, and we will refer to the special case when $p(x) = 1/r$ as the case of uniform i.i.d. letters (M00 Model), in accordance with the terminology of Refs. 19, 29.

Denote by $\mathbf{y}_{a,l} = y_{a+1}, y_{a+2}, \ldots, y_{a+l}$ the substring of length l starting right after position a, $a = 0, 1, \ldots, k-l$ and define the indicator function

$$f_a(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, & \mathbf{x} = \mathbf{y}_{a,l} \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

In other words, $f_a = 1$, if and only if $\mathbf{x}$ matches $\mathbf{y}$ right after position a, and zero otherwise.

Most of the results to follow take on the same form irrespective of the underlying distribution. It is therefore useful to define the following quantities:

Let Prob$\{\mathbf{y}\}$ denote the (stationary) probability Prob$\{Y_{a+1} = y_1, \ldots Y_{a+k} = y_k\}$, then

$$\text{Prob}\{\mathbf{y}\} = \begin{cases} \frac{1}{r^k} & \text{(M00 model)}, \\ \prod_{i=1}^{k} p(y_i) & \text{(M0 model)}, \\ \mu(y_1, y_2, \ldots, y_s) \prod_{i=s+1}^{k} \pi(y_{i-s}, y_{i-s+1}, \ldots, y_{i-1}; y_i) & \text{(M}s\text{ model)}. \end{cases} \tag{2.3}$$

Likewise, denote by $p(\mathbf{x}) = \text{Prob}\{\mathbf{x}\}$ the (unconditional) probability of encountering $\mathbf{x}$,

$$p(\mathbf{x}) = \begin{cases} \frac{1}{r^l} & \text{(M00 model)}, \\ \prod_{i=1}^{l} p(x_i) & \text{(M0 model)}, \\ \mu(x_1, x_2, \ldots, x_s) \prod_{i=s+1}^{l} \pi(x_{i-s}, x_{i-s+1}, \ldots, x_{i-1}; x_i) & \text{(M}s\text{ model)}, \end{cases} \tag{2.4}$$

For the Ms models, we will assume that $s \leq l$.

Let $p(n; m, \mathbf{x})$ be the probability that a randomly drawn k-string $\mathbf{y}$, contains a given l-string $\mathbf{x}$ *precisely* n times, with $m = k - l$. In terms of the indicator functions f_a, $p(n; m, \mathbf{x})$ can be written as

$$p(n; m, \mathbf{x}) = \sum_{\mathbf{y}} \text{Prob}\{\mathbf{y}\} \sum_{a_1 < a_2 < \cdots < a_n} I(a_1, a_2, \ldots a_n; \mathbf{x}, \mathbf{y}), \tag{2.5}$$

where

$$I(a_1, a_2, \ldots a_n; \mathbf{x}, \mathbf{y}) = \left[\prod_{i_1=1}^{a_1-1}(1 - f_{i_1})\right] f_{a_1} \left[\prod_{i_2=a_1+1}^{a_2-1}(1 - f_{i_2})\right] f_{a_2} \cdots \times \left[\prod_{i_n=a_{n-1}+1}^{a_n-1}(1 - f_{i_n})\right] f_{a_n} \left[\prod_{i_{n+1}=a_n+1}^{m}(1 - f_{i_{n+1}})\right]. \tag{2.6}$$

Thus $I(a_1, a_2, \ldots a_n; \mathbf{x}, \mathbf{y})$ is the indicator function for the event that the word $\mathbf{x}$ occurs precisely n times and the occurrences are at positions $a_1 < a_2 < \ldots < a_n$.

Using the stationarity property, the matching probability $p(n; m, \mathbf{x})$ can be shown to factorize as (see Appendix A.1 for details on the following results)

$$p(n; m, \mathbf{x}) = p(\mathbf{x}) \sum_{a_1 < a_2 < \cdots < a_n} d(a_1; \mathbf{x}) \left[\prod_{i=1}^{n-1} h(a_{i+1} - a_i; \mathbf{x})\right] d(m - a_n; \mathbf{x}), \tag{2.7}$$

with $d(b; \mathbf{x})$ and $h(b; \mathbf{x})$ defined as

$$d(b; \mathbf{x}) = \sum_{y_1 \cdots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l} | \mathbf{y}_{0,l} = \mathbf{x}\} f_0 \left[\prod_{a=1}^{b}(1 - f_a)\right] \tag{2.8}$$

$$h(b; \mathbf{x}) = \sum_{y_1 \cdots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l} | \mathbf{y}_{0,l} = \mathbf{x}\} f_0 \left[\prod_{a=1}^{b-1}(1 - f_a)\right] f_b. \tag{2.9}$$

In the above expressions $\text{Prob}\{\mathbf{y}_{0,b+l} | \mathbf{y}_{0,l} = \mathbf{x}\}$ is the conditional probability of generating the string $y_1, \ldots, y_{b+l}$ given that the first l-letters are the word $\mathbf{x}$. Thus $d(b; \mathbf{x})$ is the conditional probability for the event, that given there is an occurrence of $\mathbf{x}$, the next occurrence is more than b positions away down the string, while $h(b; \mathbf{x})$ is the conditional probability for the event that the next occurrence of $\mathbf{x}$ is b positions down the string.

The probabilities $d(b; \mathbf{x})$ and $h(b; \mathbf{x})$ can be shown to satisfy the following recursions:

$$d(b; \mathbf{x}) = d(b - 1; \mathbf{x}) - h(b; \mathbf{x}), \tag{2.10}$$

$$h(b;\mathbf{x}) = C(b;\mathbf{x}) - \sum_{a=1}^{b-1} h(a;\mathbf{x})C(b-a;\mathbf{x}), \tag{2.11}$$

with $d(0;\mathbf{x}) = 1$, $h(0;\mathbf{x}) \equiv 0$. The function $C(b;\mathbf{x})$ is given by

$$\begin{aligned} C(b;\mathbf{x}) &= \sum_{y_1 \cdots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l} | \mathbf{y}_{0,l} = \mathbf{x}\} f_0(\mathbf{x},\mathbf{y}) f_b(\mathbf{x},\mathbf{y}) \\ &= \begin{cases} c_b \text{ Prob}\{\mathbf{y}_{l,b} = \mathbf{x}_{l-b,b} | \mathbf{y}_{0,l} = \mathbf{x}\}, & 0 < b < l, \\ p(\mathbf{x}) \sum_{y_1 \ldots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l} | \mathbf{y}_{0,l} = \mathbf{y}_{b,l} = \mathbf{x}\} & b \geq l, \end{cases} \end{aligned} \tag{2.12}$$

with $c_b(\mathbf{x})$ defined as

$$c_b(\mathbf{x}) = \sum_{y_1 \cdots y_{b+l}} f_0 f_b = \begin{cases} 1, & \text{if } \mathbf{x}_{0,b} = \mathbf{x}_{l-b,b} \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

We see from Eq. (2.12) that $C(b;\mathbf{x})$ is the conditional probability of the event that there is an occurrence of $\mathbf{x}$ (not necessarily the first one) a distance b down the string from a given occurrence.

From Eqs. (2.10), (2.11), (2.12) and (2.13) it follows that $p(n;m,\mathbf{x})$, Eq. (2.7), is determined by the set of indices $c_b(\mathbf{x})$ together with the known probabilities of generating $\mathbf{x}$, as well as its suffices $\mathbf{x}_{l-b,b} = x_{l-b+1}, x_{l-b+2}, \ldots, x_l$ of length b with $b = 1, 2, \ldots l-1$. The M00 model of uniformly distributed letters forms an exception, since in this case these probabilities depend only on the lengths of the word and its suffixes, but not on the word itself. Thus for the M00 model $p(n;m,\mathbf{x})$ is determined by $\mathbf{c}(\mathbf{x})$ and therefore words with common bit-vector $\mathbf{c}(\mathbf{x})$ have the same probability of occurrences.

As evident from Eq. (2.13), the set of indices $c_b(\mathbf{x}) \in \{0,1\}$, defined for $1 \leq b \leq l-1$, measure the auto-correlations of $\mathbf{x}$. They are referred to as the bit-vector $\mathbf{c} = (c_1, c_2, \ldots, c_{l-1})$ associated with $\mathbf{x}$, and their properties were studied first by Harborth[30] and later in considerable detail by Guibas and Odlyzko.[7–9]

From the definition, Eq. (2.13), it follows that $c_b = 1$ if and only if the string $\mathbf{x}$ shifted by an amount b relative to itself coincides on the overlapping part. Conversely $c_b = 0$, if the overlapping part does not coincide. The set of r^l possible words $\mathbf{x}$ of length l is thus partitioned into sets of words with common bit-vector $\mathbf{c} = (c_1, c_2, \ldots c_{l-1})$. It turns out that the set of possible bit-vectors $\mathbf{c}$ is independent of the number of letters r (excluding the trivial case $r = 1$).[8] Table 1 list the sets of possible bit-vectors upto $l = 8$ along with their number of elements for $r = 2, 3, 4$.

The definition of $c_b(\mathbf{x})$, Eq. (2.13), imposes strong conditions on the possible values of the $l-1$ bits of a bit-vector and the resulting bit-vectors have interesting properties[8,31]: If $c_p = c_q = 1$ with $p < q$ this implies that $c_t = 1$ for all t of the form $t = p + i(q-p)$ with, $i = 0, 1, 2, \ldots$ and $t < l$. This is referred to as

Table I. The possible bit-vectors $\mathbf{c} = (c_1, c_2, \ldots, c_{l-1})$ associated with the words of length l, for $l = 2 - 8$. Note that the set of possible bit-vectors is independent of the size r of the alphabet, ($r \geq 2$, of course), and depends only on the word length l. The number of words having a common bit-vector $\mathbf{c}$ does depend on r, and is given in the adjacent columns for $r = 2, 3$, and 4.

$\mathbf{c}$	$r=2$	$r=3$	$r=4$	$\mathbf{c}$	$r=2$	$r=3$	$r=4$
0	2	6	12	000000	40	1242	11328
1	2	3	4	000001	38	606	3732
00	4	18	48	000010	16	162	768
01	2	6	12	000011	12	72	240
11	2	3	4	000100	8	54	192
000	6	48	180	000101	2	12	36
001	6	24	60	000111	2	6	12
010	2	6	12	001001	6	24	60
111	2	3	4	010101	2	6	12
0000	12	144	720	111111	2	3	4
0001	10	66	228	0000000	74	3678	45132
0010	4	18	48	0000001	82	1866	15108
0011	2	6	12	0000010	26	462	3012
0101	2	6	12	0000011	22	210	948
1111	2	3	4	0000100	16	162	768
00000	20	414	2832	0000101	8	54	192
00001	22	210	948	0000111	6	24	60
00010	6	48	180	0001000	6	48	180
00011	6	24	60	0001001	6	24	60
00100	4	18	48	0010010	4	18	48
00101	2	6	12	0010011	2	6	12
01010	2	6	12	0101010	2	6	12
11111	2	3	4	1111111	2	3	4

the *forward propagation rule*.[8] In particular, $c_p = 1$ implies that $c_{ip} = 1$ for all $i, 1, 2, \ldots$ such that $ip < l$. The latter result shows that p can be considered as a period. We define the *fundamental period* χ of a string $\mathbf{x}$ to be the smallest p, with $0 < p < l$ such that $c_p = 1$. If $\mathbf{x}$ is such that its bit-vector is $000\ldots0$, we define $\chi = l$.

Note that the average number of matches $\langle n \rangle$ as well as its variance σ_n^2 can be readily obtained by writing n as

$$n = \sum_{a=0}^{m} f_a, \tag{2.14}$$

yielding

$$\langle n \rangle = (m+1)p(\mathbf{x}), \tag{2.15}$$

and

$$\sigma_n^2 = (m+1)p(\mathbf{x}) + 2p(\mathbf{x})\sum_{a=1}^{m}(m-a+1)C(a;\mathbf{x}) - (m+1)^2 p^2(\mathbf{x}), \quad (2.16)$$

where $C(a;\mathbf{x})$ is given by Eq. (2.12).[21]

3. THE PROBABILITY OF *n* OCCURRENCES AS THE PARTITION FUNCTION OF A 1D LATTICE GAS

The expression for $p(n;m,\mathbf{x})$ in the form of Eq. (2.7) already resembles the configurational partition function of a gas of n particles with particle boundary interactions proportional to $-\ln d$ and nearest neighbor particle-particle interactions proportional to $-\ln h$. In order to make this analogy work however, we need to consider what we mean by the free-particle, i.e. the no interaction limit.

Recall that $d(b;\mathbf{x})$ and $h(b;\mathbf{x})$ are conditional matching probabilities. Thus $h(b;\mathbf{x})$ is the probability of the event: given an occurrence of $\mathbf{x}$ at position a, the next occurrence is at $a+b$. As we will show below, $h(b;\mathbf{x})$ and $d(b;\mathbf{x})$, Eqs. (3.11) and (3.12), turn out to decay exponentially for large b and this behavior can be interpreted as corresponding to the approximation when correlations inherent in these events are ignored. The ratios $d(b;\mathbf{x})/d_{\text{asy}}(b;\mathbf{x})$ and $h(b;\mathbf{x})/h_{\text{asy}}(b;\mathbf{x})$ thus measure the strength of these correlations and it is natural to define the particle-boundary and particle-particle interactions, $U_b(b)$ and $U(b)$, respectively, as

$$e^{-\beta U_b(b)} = \frac{d(b)}{d_{\text{asy}}(b)} \quad (3.1)$$

$$e^{-\beta U(b)} = \frac{h(b)}{h_{\text{asy}}(b)}, \quad (3.2)$$

obtaining thereby interactions that vanish as $b \to \infty$. Since the interactions do not have an intrinsic scale, a temperature by itself is meaningless and we will write "energies" always with the pre-factor β, i.e. in dimension-less units.

Using the interactions U and U_b as defined above, we will show next that $p(n;m,\mathbf{x})$ can be cast as the (configurational) n particle term in the grand partition function of a $1d$ lattice gas enclosed in a "volume" m. We therefore turn first to the asymptotic behavior of $d(b;\mathbf{x})$ and $h(b;\mathbf{x})$.

Let $f(z) = \sum_{b=0}^{\infty} z^b f(b)$ be the generating function associated with $f(b)$. Then if $f(z)$ is analytic in the region enclosing the origin except for a finite number of poles, $f(b)$ is asymptotically given as[32]

$$f_{\text{asy}}(b) = \sum_{j=1}^{q} \frac{(-1)^j a_{-j}}{z_1^{b+1}} \binom{b+j-1}{j-1}, \quad (3.3)$$

where z_1 is the pole of $f(z)$ of smallest modulus, q is its multiplicity and a_{-j} are the coefficients of the Laurent expansion of $f(z)$ around z_1.

From Eqs. (2.12), (2.10) and (2.11) one finds

$$C(z;\mathbf{x}) = c(z;\mathbf{x}) + p(\mathbf{x})\frac{z^l}{1-z} + p(\mathbf{x})z^l T(z), \tag{3.4}$$

where $c(z;\mathbf{x})$ is a polynomial of degree $l-1$,

$$c(z;\mathbf{x}) = \sum_{b=1}^{l-1} c_b z^b \text{ Prob}\{\mathbf{y}_{l,b} = \mathbf{x}_{l-b,b} | \mathbf{y}_{0,l} = \mathbf{x}\}, \tag{3.5}$$

and $T(z)$ is defined by

$$T(z) = -\sum_{b=l}^{\infty} z^{b-l} \left[1 - \sum_{y_1 \ldots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l} | \mathbf{y}_{0,l} = \mathbf{y}_{b,l} = \mathbf{x}\} \right], \tag{3.6}$$

so that

$$d(z;\mathbf{x}) = \frac{1}{p(\mathbf{x})} \frac{1}{\lambda(z;\mathbf{x})}, \tag{3.7}$$

and

$$h(z;\mathbf{x}) = 1 - \frac{1}{p(\mathbf{x})} \frac{1-z}{\lambda(z;\mathbf{x})}, \tag{3.8}$$

with $\lambda(z;\mathbf{x}) \equiv (1-z)C(z;\mathbf{x})/p(\mathbf{x})$ given by

$$\lambda(z;\mathbf{x}) = z^l + \frac{1}{p(\mathbf{x})}(1-z)\left[1 + c(z;\mathbf{x}) + p(\mathbf{x})T(z)z^l\right]. \tag{3.9}$$

The probability of n-occurrences, Eq. (2.7), is given in terms of its generating function as

$$p(n;z,\mathbf{x}) = p(\mathbf{x})\, d^2(z;\mathbf{x})\, h^{n-1}(z;\mathbf{x}). \tag{3.10}$$

Equations (3.7), (3.8), (3.9), and (3.10), were derived first in the context of $M1$ models by Régnier and Szpankowski.[15] However they are valid for higher order Markov chains as well,[29,19] since apart from requiring that $s \leq l$, no further assumption on s has been made so far.

We see from Eqs. (3.7), (3.8) and (3.10) that the poles of the generating functions of $d(z;\mathbf{x})$, $h(z;\mathbf{x})$ and $p(n;z;\mathbf{x})$ are determined by the zeroes of $\lambda(z;\mathbf{x})$. Therefore, if $\lambda(z;\mathbf{x})$ is a rational function, the asymptotic forms of $h(b;\mathbf{x})$ and $d(b;\mathbf{x})$ can indeed be determined from Eq. (3.3), with z_1 being the zero of $\lambda(z;\mathbf{x})$ of smallest modulus.

Note that for i.i.d. random letters, the M00 and M0 models, $T(z) = 0$ and thus $\lambda(z;\mathbf{x})$ is a polynomial of degree l. Also note that any r-state Markov chain of order s can be converted into an r^s state Markov chain of order 1 by identifying the

new states as the set of all possible words of length s, and using the original state transition probabilities to calculate the new ones. Under this identification the new states $\tilde{y}_a$ correspond to the substrings $y_{a+1}, \ldots, y_{a+s}$, for $a = 0, 1, \ldots, k-s$, with an analogous identification for $\mathbf{x} \to \tilde{x}_0, \ldots \tilde{x}_{l-s}$. Thus the original problem can be reformulated in terms of the new states and a Markov chain of order 1. We will henceforth consider only 1st order Markov chains in which case the state transition probabilities can be represented by a matrix. We will further assume that the state transition matrix is diagonalizable so that $T(z)$ is a rational function (*cf.* Ref. 29, Lemma 1) and thus $\lambda(z; \mathbf{x})$ is a rational function.

Régnier and Szpankowski have proven,[15] that $\lambda(z; \mathbf{x})$ has at least one real zero, and that all zeroes satisfy $|z| > 1$. Nothing can be said in general about the multiplicity of z_1, the zero closest to the circle $|z| = 1$, except in the case of the M00 model, for which Guibas and Odlyzko have proven[7] that z_1 is real and its multiplicity is 1. Fig. 2 shows a plot of the zeroes of $\lambda(z; \mathbf{x})$ for $l = 4, r = 2$ (left) and $l = 8, r = 2$ (right) for the M00 model. In this case $\lambda(z; \mathbf{x})$ is a polynomial of degree l whose coefficients are determined entirely by the components of the bit-vector associated with $\mathbf{x}$. For $l = 4$ (left) the possible bit-vectors are, see Table 1,

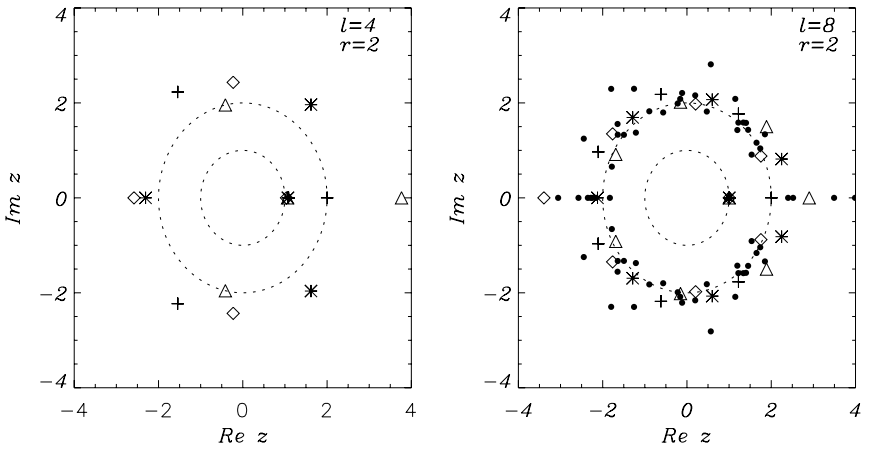

Fig. 2. Plot of the zeroes of $\lambda(z; \mathbf{x})$, Eq. (3.9). The figures are for binary words, $r = 2$, of length $l = 4$ (left) and $l = 8$ (right) with uniform i.i.d. letters (M00 model). Note that for the M00 model $\lambda(z; \mathbf{x})$ is a polynomial of degree l whose coefficients are determined entirely by the components of the bit-vector associated with $\mathbf{x}$. Left: $l = 4$ for which the possible bit-vectors are, see Table 1, $\mathbf{c} = 000$ (+), $\mathbf{c} = 001$ (*), $\mathbf{c} = 010$ (triangles) and $\mathbf{c} = 111$ (diamonds). Right: $l = 8$ (right) with $\mathbf{c} = 0000000$ (+), $\mathbf{c} = 0000001$ (*), $\mathbf{c} = 0000010$ (diamonds), $\mathbf{c} = 0000011$ (triangles). The zeroes associated with the remaining bit-vectors have been shown as small dots. The dashed circles correspond to $|z| = 1$ and $|z| = r = 2$ and have been inserted as a guide to the eye. The polynomial $\lambda(z; \mathbf{x})$ associated with each bit-vector has a single zero near $z = 1$, with the remaining zeroes clustering around and beyond the circle $|z| = r$.

$\mathbf{c} = 000$ (+), $\mathbf{c} = 001$ (*), $\mathbf{c} = 010$ (triangles) and $\mathbf{c} = 111$ (diamonds), while for the $l = 8$ (right) they are $\mathbf{c} = 0000000$ (+), $\mathbf{c} = 0000001$ (*), $\mathbf{c} = 0000010$ (diamonds), $\mathbf{c} = 0000011$ (triangles), and we have shown the zeroes associated with the remaining bit-vectors as small dots. The dashed circles correspond to $|z| = 1$ and $|z| = r = 2$ and have been inserted as a guide to the eye. As remarked above, the polynomial $\lambda(z;\mathbf{x})$ associated with each bit-vector has a single zero near $z = 1$.[7] The remaining zeroes are seen to cluster around the circle $|z| = r$ and beyond.

Returning to the case of general random letter strings, it is readily seen from the general form of $\lambda(z;\mathbf{x})$, Eq. (3.9), that $\lambda(z;\mathbf{x})$ has a real zero z_1 at $z = 1$, with $z_1 - 1$ of order $p(\mathbf{x})$. The case when z_1 has multiplicity greater than one requires more care, but otherwise does not cause additional difficulties. In the following, we will therefore assume that z_1 is real and has multiplicity 1.

From Eq. (3.3) with $q = 1$, we see that the asymptotic behavior of $h(b;\mathbf{x})$ and $d(b;\mathbf{x})$ is given by

$$h_{\text{asy}}(b) = p(\mathbf{x})\frac{A_1}{z_1}\left[\frac{z_1 - 1}{p(\mathbf{x})}\right]^2\left(\frac{1}{z_1}\right)^b \equiv e^{\beta\mu}\left(\frac{1}{z_1}\right)^b, \tag{3.11}$$

and

$$d_{\text{asy}}(b) = \left[\frac{A_1}{z_1 p(\mathbf{x})}e^{\beta\mu}\right]^{\frac{1}{2}}\left(\frac{1}{z_1}\right)^b, \tag{3.12}$$

where z_1 is the root of smallest magnitude of the polynomial $\lambda(z;\mathbf{x})$, which to (leading) order $p(\mathbf{x})$ is given by[3]

$$z_1 = 1 + \frac{p(\mathbf{x})}{1 + c(1;\mathbf{x}) + p(\mathbf{x})T(1)} + \mathcal{O}\left(p(\mathbf{x})^2\right), \tag{3.13}$$

[3] Infact, using the Lagrange Inversion Formula,[32] $z_1 - 1$ can be expanded in a power series in $p(\mathbf{x})$: Letting $u = z - 1$ and $t = p(\mathbf{x})[1 + c(1;\mathbf{x}) + p(\mathbf{x})T(1)]^{-1}$, the equation $\lambda(z;\mathbf{x}) = 0$, Eq. (3.9) can be written in the form

$$u = t\,\phi(u),$$

where

$$\phi(u) = (1+u)^l\frac{1 + c(1)}{1 + c(1+u;\mathbf{x}) + p(\mathbf{x})T(1+u)(1+u)^l}.$$

is a formal power series in u. Thus

$$z_1 = 1 + u(t) = 1 + \sum_{i=1}^{\infty} u_i t^i,$$

with

$$u_i = \frac{1}{i!}\left.\frac{\mathrm{d}^{i-1}\phi^i}{\mathrm{d}u^{i-1}}\right|_{u=0}.$$

Since $u_1 = 1$, the leading order result for z_1, Eq. (3.13), follows.

and

$$A_1 = -\frac{1}{(z_1 - 1)\lambda'(z_1; \mathbf{x})}. \tag{3.14}$$

It can be shown that $A_1 = 1 + \mathcal{O}(p(\mathbf{x}))$.

Note that Eqs. (3.11) and (3.12) imply that the events "$\mathbf{x}$ does not occur at position b" are asymptotically independent and have probability $1/z_1$, as mentioned at the beginning of this section.

The probability of n occurrences, Eq. (2.7), can finally be cast in the form of a partition function,

$$p(n; m, \mathbf{x}) = \frac{A_1}{z_1^{m+1}} e^{\beta\mu n} \sum_{a_1 < a_2 < \cdots < a_n} e^{-\beta \mathcal{H}_n(a_1, \ldots, a_n)}, \tag{3.15}$$

with the fugacity $e^{\beta\mu}$ as defined in Eq. (3.11), and the n-particle Hamiltonian given by

$$\mathcal{H}_n(a_1, \ldots, a_n) = U_b(a_1) + U_b(m - a_n) + \sum_{i=1}^{n-1} U(a_{i+1} - a_i) \tag{3.16}$$

Thus $p(n; m, \mathbf{x})$ can be regarded as the configurational n particle term in the grand partition function of a 1 dimensional lattice gas with chemical potential μ, enclosed in a "volume" m, and whose particles interact with each other and the boundaries via pairwise nearest-neighbour interactions, U and U_b, respectively.

In the probability description, the no-interaction limit, $U = U_b = 0$, corresponds to the case where all correlations are ignored and can be readily worked out. As one might expect, the approximation turns out to be very poor, indicating that the interactions cannot be ignored. We therefore turn next to the properties of the interactions.

4. INTERACTIONS

Consider the particle-particle interaction first. From Eqs. (3.2) and (3.11) we find that

$$\beta U(b) = -\ln h(b) - b \ln z_1 + \ln p(\mathbf{x}) + \beta U_0, \tag{4.1}$$

where

$$\beta U_0 = \ln \left[\frac{A_1}{z_1} \left(\frac{z_1 - 1}{p(\mathbf{x})} \right)^2 \right]. \tag{4.2}$$

Treating $p(\mathbf{x})$ as a small parameter, it can be shown that the term in square brackets is one to this order, *cf.* Eqs. (3.13) and (3.14), and hence βU_0 is of order $p(\mathbf{x})$.

Two regimes of the interaction $\beta U(b)$ emerge: the asymptotically decaying tail $b \gg l$ and the core region $b < l$. The asymptotic behavior of $\beta U(b)$ is due to the second dominant pole $|z_2|$ of $h(z;\mathbf{x})$ and thus gives rise to asymptotic exponential decay with characteristic length scale equal to $[\ln(|z_2|/z_1)]^{-1}$.

In the core-region $b < l$, the values of $h(b;\mathbf{x})$ can be obtained from the recursion relation, Eqs. (2.11) and (2.12). One finds that for $b < l$ the non-zero values of h are determined by the bit-vector $\mathbf{c}$ associated with $\mathbf{x}$ as

$$h(b;\mathbf{x}) = \begin{cases} c_b \text{Prob}\{\mathbf{y}_{l,b} = \mathbf{x}_{l-b,b} | \mathbf{y}_{0,l} = \mathbf{x}_{0,l}\}, & \text{if } \chi \text{ does not divide } b, \\ \text{Prob}\{\mathbf{y}_{l,\chi} = \mathbf{x}_{l-\chi,\chi} | \mathbf{y}_{0,l} = \mathbf{x}_{0,l}\}, & \text{if } b = \chi, \\ 0, & \text{otherwise,} \end{cases} \tag{4.3}$$

where χ is the fundamental period associated with $\mathbf{c}$ that was defined at the end of Sec. 2, and by definition $h(0) = 0$. The set of b, with $b < l$, for which $h(b;\mathbf{x})$ is non-zero is referred to as the set of *principal periods* of $\mathbf{x}$, $\mathcal{P}'(\mathbf{x})$.[18,19] We therefore see that for $b < l$, the interaction becomes $+\infty$, whenever b does not belong to the principal period set, and hence $h(b) = 0$. In particular, when $b < \chi$ we have $h = 0$. Thus the interaction has a repulsive hard-core for $b < \chi$, while finite values of $U(b)$ in the core-region occur only at points b belonging to the principal period set,

$$\begin{aligned} \beta U(b) = &-b \ln \left[z_1 \text{Prob}^{1/b}\{\mathbf{y}_{l,b} = \mathbf{x}_{l-b,b} | \mathbf{y}_{0,l} = \mathbf{x}_{0,l}\} \right] \\ &+ \ln p(\mathbf{x}) + \beta U_0, \quad b \in \mathcal{P}'(\mathbf{x}). \end{aligned} \tag{4.4}$$

It is instructive to consider first the M00 model, for which one finds

$$\beta U(b) = b \ln \left[\frac{r}{z_1} \right] - l \ln r + \beta U_0, \quad b \in \mathcal{P}'(\mathbf{x}). \tag{4.5}$$

Since $r/z_1 > 1$ (except when $r = l = 2$ which can be solved exactly), finite values of $U(b)$ increase with increasing b. The first finite value of $U(b)$ occurs at $b = \chi$ and from Eq. (4.5) we find

$$\beta U(\chi) = \chi \ln \left(\frac{r}{z_1} \right) - l \ln r + \mathcal{O}\left(\frac{1}{r^l} \right). \tag{4.6}$$

Thus it is apparent that for fixed χ, $\beta U(\chi)$ becomes more negative as either l or r increase. In fact we see that to leading order, the dependence of $\beta U(\chi)$ on l is linear, while its dependence on r is logarithmic, with the overall energy scale given by $-l \ln r = \ln p(\mathbf{x})$. When $\chi = l$, corresponding to $\mathbf{c} = 00\ldots0$, there is a genuine hard-core for $b < l$, since the set of principal periods is empty. The first finite value occurs at $b = \chi = l$, which is outside of the core-region and will be discussed below.

Returning now to the case of general random letter processes, the argument in square brackets of Eq. (4.4) is not necessarily smaller (or larger) than one and

thus it is not necessarily true that finite values of $U(b)$ in the core region are increasing (decreasing) with b. Instead, the behavior of these values of $U(b)$ can depend on the subtle interplay of the overall word matching probability $p(\mathbf{x})$ (which determines z_1) with the generally larger probabilities of generating its suffixes, $\text{Prob}\{\mathbf{y}_{l,b} = \mathbf{x}_{l-b,b} | \mathbf{y}_{0,l} = \mathbf{x}_{0,l}\}$. Thus such cases will depend in general on the choice of $\mathbf{x}$ as well as the stochastic model for the letter generation. Nevertheless, from Eq. (4.4) it is readily seen that

$$\beta U(\chi) = -\chi \ln z_1 + \ln p(\mathbf{x}_{0,l-\chi}) + \mathcal{O}(p(\mathbf{x})). \tag{4.7}$$

Thus $\beta U(\chi) < 0$, while the overall scale of finite energies in the core region goes as $\ln p(\mathbf{x})$, for $\chi < l$. When $\chi = l$, we again have a genuine hard core for $b < l$ and the first finite value occurs at $b = \chi = l$, outside of the core-region, to which we turn next.

Note that for $b > l$, $h(b; \mathbf{x})$ is always non-zero, and thus $U(b)$ is finite, as can be shown using Eqs. (2.11), (2.12) and (3.2). The case $b = l$ is a little bit more subtle, and $h(l; \mathbf{x})$ can be zero if $\chi < l/2$. We will show that for $b > l$, $h(b; \mathbf{x})$ is of the form $h(b; \mathbf{x}) = \alpha(b)p(\mathbf{x})$, where $\alpha(b)$ is to leading order of order one with respect to $p(\mathbf{x})$. Substituting this expression for $h(b; \mathbf{x})$ into Eq. (4.1), one finds that the $\ln p(\mathbf{x})$ terms cancel and we are left with

$$\beta U(b) = -\ln \alpha(b) - b \ln z_1 + \beta U_0, \tag{4.8}$$

which upon expanding the arguments of the logarithms, which are all of order one now, is seen to be of order $p(\mathbf{x})$. To prove that $h(b; \mathbf{x}) \sim p(\mathbf{x})$ for $b > l$, note that from Eq. (2.12) and with $b > l$, $C(b; \mathbf{x})$ is already of order $p(\mathbf{x})$ and determines $h(b; \mathbf{x})$ via the recursion relation, Eq. (2.11). It can be shown that the convolution term is always of the from $\gamma(b)p(\mathbf{x})$, where $\gamma(b)$ is again of order one to leading order. This can be proven by induction with the only case requiring care being $l < b < 2l$, in which case the convolution sum contains terms $h(a; \mathbf{x})C(b - a; \mathbf{x})$ where both $a < l$ and $b - a < l$. Taking into account the periodicity a of $\mathbf{x}$, which is necessary for a non-zero term, the product of the probabilities of the corresponding partial words, Eqs. (2.13) and (4.3), combine to yield a factor $p(\mathbf{x})$. Thus an overall term $p(\mathbf{x})$ can be factored out from the RHS of the recursion, Eq. (2.11), and $h(b; \mathbf{x}) \sim p(\mathbf{x})$ for $b > l$.

In summary, we find that the core-region of the inter-particle interaction contains infinite repulsive segments as well as finite values with $U < 0$, and that the characteristic energy scales of these is of order $\ln p(\mathbf{x})$. The tail region on the other hand is free of terms $\ln p(\mathbf{x})$, and behaves to leading order as $p(\mathbf{x})$, with an exponential asymptotic decay. This means that as $p(\mathbf{x}) \to 0$, the core due to its logarithmic dependence on $\ln p(\mathbf{x})$, becomes stronger, with $U(\chi)$ becoming more strongly attractive, while the overall strength of the tail, which goes as $p(\mathbf{x})$, weakens. We have plotted in Fig. 3 the behavior of the inter-particle interaction associated with words belonging to the bit-vectors $\mathbf{c} = 0 \ldots 0$ (left) and $\mathbf{c} = 1 \ldots 1$

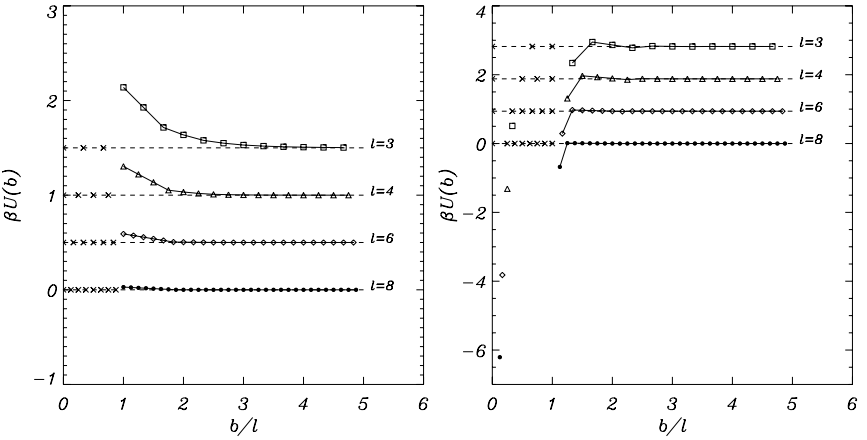

Fig. 3. Plot of the inter-particle interactions $\beta U(b)$, Eq. (4.1), associated with words belonging to the bit-vector $\mathbf{c} = 0 \ldots 0$ and $1 \ldots 1$ in their dependence on the word lengths l for the M00 model with $r = 2$. The potentials are plotted against distance measured in units of the word length l and have been vertically offset for clarity. The dashed lines represent the $U = 0$ line for each potential with crosses on a line indicating that the corresponding potential at that point is $+\infty$. Left: Interparticle potentials associated with the bit-vector $\mathbf{c} = 0 \ldots 0$ for words of length $l = 3, 4, 6$ and 8. Note that the interactions have a hard-core of size $b/l = 1$ followed by a repulsive tail. The strength of the tail weakens with increasing l. Right: Inter-particle potentials associated with the bit-vector $\mathbf{c} = 1 \ldots 1$ for words of length $l = 3, 4, 6$ and 8. Note that these interactions have an attractive part at $b = 1$, followed by a hard-core for $b/l < 1$, and a weak, oscillatory decaying tail. Also note the opposite behavior of the strength of the core and the tail: With increasing l, the strength of the attractive part of the core increases, while the strength of the tail decreases.

(right) in their dependence on the word length l for the uniform i.i.d. letter model M00 with $r = 2$. The potentials are plotted against distance measured in units of the word length l and have been vertically offset for clarity. The dashed lines represent the $U = 0$ line for each potential and crosses on a line indicate that the corresponding potential is $+\infty$ at that point. Since $p(\mathbf{x}) = 1/2^l$, the tails of the interaction become weaker as l increases. The core of the $\mathbf{c} = 1 \ldots 1$ interaction has a single finite value at $b = 1$, since its principal period set contains only $b = 1$. We see that the cores of the $\mathbf{c} = 1 \ldots 1$ interactions are attractive and become stronger with increasing l. The core of the $\mathbf{c} = 0 \ldots 0$ interaction on the other hand is genuine hard-core, since the corresponding principal period set is empty.

Turning to the particle-boundary interactions, notice that Eq. (2.10), written as

$$d(b; \mathbf{x}) = 1 - \sum_{a=1}^{b} h(a; \mathbf{x}), \tag{4.9}$$

relates the properties of h to those of d. It turns out that $U_b(b)$ is singularity free and overall behaves like the tail of the particle-particle interactions.

Lastly, let us remark that the case of uniform i.i.d.letters is very special, since $p(\mathbf{x}) = 1/r^l \leq 1/2$. For the general letter distributions, both the distribution as well as the word $\mathbf{x}$ can be chosen arbitrarily and thus there is no constraint on the values that $0 < p(\mathbf{x}) < 1$ can take. This means in particular that the relative strength of the core and tail regions of the inter-particle interactions can vary within a wider range.

5. THE GRAND PARTITION FUNCTION

Having discussed the features of the interactions, the next step is to evaluate the n-particle terms of the grand partition sum, Eq. (3.15) and (3.16). Our approach is a lattice version of a method due to Gürsey[35] for treating the $1d$ continuum gas of n particles that only interact with their neighbours.

This section is organized as follows: In 5.1 we will rewrite the n-particle configurational sum in terms of a contour-integral over a product of generating functions. Since the interactions decay exponentially at large distances, and we have argued in Sec. 4 that in general the core region $b < l$ of the interaction is stronger than its tail, we will introduce in 5.2 a cut-off distance $\Lambda > l$, beyond which all interactions will be eventually set to zero. The larger Λ, the more of the tail of the interactions is kept and thus Λ can be used as a parameter to control the quality of the approximation. We next evaluate the contour-integral with the interactions cut-off at Λ by using a stationary phase approximation to obtain an approximate analytical expression for $p(n; m, \mathbf{x})$. Comparing with the exact distributions, we show in 5.3 that our approximation performs rather well. We also find that our approximation performs generally better than the Gaussian and Compound Poisson asymptotic forms especially for words $\mathbf{x}$, for which the tails of the corresponding interactions are relatively strong. This implies that the asymptotic forms must have been obtained by suppressing the tails of the interactions and we will return to this in Sec. 6, where we will discuss asymptotic behavior. As in the continuum gas, the dependence of the point of stationary phase on the "volume" m and number of particles n, turns out to furnish the equation of state. We will discuss this in 5.4 and show that in the large m limit the equation of state has a virial expansion. We will determine the first two virial coefficients, which in complete analogy to the continuum model, are given as certain sums over the Boltzmann factors of the interactions.

Our method of calculating the probability of occurrences by making first an analogy with the grand partition function of a lattice gas and then evaluating the latter approximately, is an approach based on liquid-theory. We will briefly remark in 5.5 on its advantages.

5.1. The Probability of Occurrence as a Stationary Phase Integral

Define the generating functions associated with the Boltzmann factors, Eqs. (3.1) and (3.2), as

$$D(z) = \sum_{b=0}^{\infty} z^b e^{-\beta U_b(b)}, \tag{5.1}$$

$$H(z) = \sum_{b=0}^{\infty} z^b e^{-\beta U(b)}, \tag{5.2}$$

which in terms of the generating functions of $d(b;\mathbf{x})$ and $h(b;\mathbf{x})$, are given by

$$D(z) = \left[p(\mathbf{x})e^{-\beta\mu}\frac{z_1}{A_1}\right]^{1/2} d\,(zz_1;\mathbf{x}) \tag{5.3}$$

$$H(z) = e^{-\beta\mu}h\,(zz_1;\mathbf{x})\,. \tag{5.4}$$

Using the convolution property, Eq. (3.15) can be written in terms of the generating functions $D(z)$ and $H(z)$ as

$$p(n;m,\mathbf{x}) = \frac{A_1}{z_1^{m+1}}e^{\beta\mu n}\frac{1}{2\pi i}\oint_{\partial D} dz\frac{1}{z^{m+1}}\;\; D^2(z)H^{n-1}(z), \tag{5.5}$$

where the contour is the boundary of a domain enclosing the origin inside of which $D^2(z)H^{n-1}(z)$ is analytic.[32] Equation (5.5) is the lattice version of the partition function of a gas in a 1d continuum with pairwise nearest neighbor interactions which has been treated in detail by Gürsey[35] and Fisher[36] by evaluating the contour integral, Eq. (5.15), by the method of stationary phase, which in the context of generating functions is also known as Hayman's method[32]: Writing the integral in Eq. (5.15) as

$$I = \frac{1}{2\pi i}\oint_{\partial D} dz\frac{1}{z^{m+1}}\;\; f(z), \tag{5.6}$$

the value of the integral for large m is given approximately by Ref. 32

$$I \approx \left(\frac{1}{u_m}\right)^m \frac{f(u_m)}{\sqrt{2\pi\, b_m}}, \tag{5.7}$$

where u_m is the positive real root of the equation

$$m = u\frac{\mathrm{d}}{\mathrm{du}}\ln f(u) \tag{5.8}$$

and b_m is given by

$$b_m = \left[u\frac{\mathrm{d}}{\mathrm{du}}\ln f(u) + u^2\frac{\mathrm{d}^2}{\mathrm{du}^2}\ln f(u)\right]_{u=u_m}. \tag{5.9}$$

5.2. The Λ Cut-off Approximation

As discussed in the beginning of this section, it is useful to introduce a cut-off distance Λ so that

$$D_\Lambda(z) = \sum_{b=0}^{\Lambda-1} z^b e^{-\beta U_b(b)}, \tag{5.10}$$

$$H_\Lambda(z) = \sum_{b=0}^{\Lambda-1} z^b e^{-\beta U(b)}. \tag{5.11}$$

We define next a new interaction $U_\Lambda(b)$ as

$$e^{-\beta U_\Lambda(b)} = \begin{cases} e^{-\beta U(b)}, & b < \Lambda \\ 1 + \Gamma_\Lambda(b), & b \geq \Lambda, \end{cases} \tag{5.12}$$

where $\Gamma_\Lambda(b)$ is a cut-off function regulating the behavior of $e^{-\beta U_\Lambda(b)}$ beyond the cut-off distance. Eventually we will set $\Gamma_\Lambda = 0$, which means that the interaction have finite range Λ. Since, by construction, the interactions decay to zero at large distances, introducing a finite cut-off Λ will introduce only a small and controllable error in the overall calculation. Moreover, this error vanishes, when we let $\Lambda \to \infty$. Thus Λ can be used to both set up a perturbation expansion for the probability distribution as well as to control the resulting error.[4]

The generating function for the approximate inter-particle interaction becomes

$$\hat{H}_\Lambda(z) = \sum_{b=0}^{\infty} z^b e^{-\beta \hat{U}_\Lambda(b)} = H_\Lambda(z) + \frac{z^\Lambda}{1-z} + \Gamma_\Lambda(z). \tag{5.13}$$

Likewise, one finds that

$$\hat{D}_\Lambda(z) = \sum_{b=0}^{\infty} z^b e^{-\beta \hat{U}_\Lambda(b)} = D_\Lambda(z) + \frac{z^\Lambda}{1-z} + \Gamma_\Lambda(z). \tag{5.14}$$

---

[4] Likewise, by setting the potential beyond the cut-off to constant values U_+ and U_- choosen such that

$$U_+ = \max_{b \geq \Lambda} \{U(b), U_b(b)\},$$

$$U_- = \min_{b \geq \Lambda} \{U(b), U_b(b)\}$$

the variation in the probability distribution can be controlled, since it follows that

$$p(n; U_+, m, \mathbf{x}) \leq p(n; m, \mathbf{x}) \leq p(n; U_-, m, \mathbf{x}),$$

where $p(n; U_\Lambda, m, \mathbf{x}$, is the distribution, Eq. (5.5), with $U(b) = U_b(b) = U_\Lambda$ for $b \geq \Lambda$.

and the approximate distribution function is given by

$$\hat{p}(n;\Gamma_\Lambda,m,\mathbf{x})=\frac{A_1e^{\beta\mu n}}{z_1^{m+1}}\frac{1}{2\pi i}\oint_{\partial D}dz\frac{1}{z^{m+1}}\ \hat{D}_\Lambda^2(z)\hat{H}_\Lambda^{n-1}(z). \tag{5.15}$$

Applying now Hayman's method to the integral, Eq. (5.15), we let $f(u)=\hat{D}_\Lambda^2(u)\hat{H}_\Lambda^{n-1}(u)$ and find after a little bit of algebra

$$\begin{aligned}m=&\frac{2}{x}\frac{1+\Lambda x+x^2(1+x)^{\Lambda-2}\left[D'_\Lambda\left(\frac{1}{1+x}\right)+\Gamma'\left(\frac{1}{1+x}\right)\right]}{1+x(1+x)^{\Lambda-1}\left[D_\Lambda\left(\frac{1}{1+x}\right)+\Gamma\left(\frac{1}{1+x}\right)\right]}\\&+\frac{n-1}{x}\frac{1+\Lambda x+x^2(1+x)^{\Lambda-2}\left[H'_\Lambda\left(\frac{1}{1+x}\right)+\Gamma'\left(\frac{1}{1+x}\right)\right]}{1+x(1+x)^{\Lambda-1}\left[H_\Lambda\left(\frac{1}{1+x}\right)+\Gamma\left(\frac{1}{1+x}\right)\right]},\end{aligned} \tag{5.16}$$

where we have written u as

$$u=\frac{1}{1+x}. \tag{5.17}$$

and it is assumed that $\Gamma_\Lambda(z)$ has been chosen such that $(z-1)\Gamma_\Lambda(z)$ has no pole for $|z|\leq 1$.

For large m, it is seen from Eq. (5.16), that to leading order $x\sim 1/m$ and a power series expansion of x can be obtained by multiplying both sides of the above equation by x and expanding the fractions in a Taylor series around $x=0$,

$$mx=(n+1)\left\{1+\epsilon_1x+\epsilon_2x^2+\ldots\right\}. \tag{5.18}$$

The first two orders can be worked out, yielding

$$\epsilon_1=\Lambda-\Gamma_\Lambda(1)-\frac{2D_\Lambda(1)+(n-1)H_\Lambda(1)}{n+1} \tag{5.19}$$

and

$$\begin{aligned}\epsilon_2=&2\Gamma'_\Lambda(1)-(2\Lambda-1)\Gamma_\Lambda(1)\\&+\frac{1}{n+1}\{2[D_\Lambda(1)+\Gamma_\Lambda(1)]^2+(n-1)[H_\Lambda(1)+\Gamma_\Lambda(1)]^2\}\\&+\frac{2}{n+1}\{2D'_\Lambda(1)+(n-1)H'_\Lambda(1)\}\\&-\frac{2\Lambda-1}{n+1}\{[2D_\Lambda(1)+(n-1)H_\Lambda(1)]\}.\end{aligned} \tag{5.20}$$

Rewriting Eq. (5.18) in a form suitable for Lagrange's Inversion Formula,[32]

$$x=\frac{n+1}{m-(n+1)\epsilon_1}\left\{1+\epsilon_2x^2+\cdots\right\}, \tag{5.21}$$

an expansion of x in powers of $(n+1)/[m-(n+1)\epsilon_1]$ is obtained as

$$x = \frac{n+1}{m-(n+1)\epsilon_1} + \left[\frac{n+1}{m-(n+1)\epsilon_1}\right]^3 \epsilon_2 \ldots \ . \tag{5.22}$$

The expansion of b_m can be worked out in a similar manner and one finds

$$b_m = m + \frac{n+1}{x^2} - (n+1)(\epsilon_1+\epsilon_2) + \cdots \ , \tag{5.23}$$

where the omitted terms are of order x and higher.

5.3. Distributions

Setting $\Gamma_\Lambda = 0$, and evaluating the integral by the stationary phase approximation we obtain, *cf.* Eq.(5.15),

$$\hat{p}(n;0,m,\mathbf{x}) \approx \frac{A_1 e^{\beta\mu n}}{z_1^{m+1}} (1+x)^m \hat{D}_\Lambda^2 \left(\frac{1}{1+x}\right) \hat{H}_\Lambda^{n-1} \left(\frac{1}{1+x}\right) \frac{1}{\sqrt{2\pi b_m}}. \tag{5.24}$$

Because of its overall similarity, this result will be referred to as the liquid theory approximation.

Unless other precautions are taken (see Sec. 6 below), the finite-cut off along with the stationary phase approximation will generally destroy the normalization of the distribution. We will compensate for this by normalizing by an overall constant, which will be the closer to unity the better the approximation is.

The solid lines in Fig. 1 show the approximate distribution, Eq. (5.24), for the four equivalence classes associated with words of length $l = 4$, $r = 2$, and $m = 252$ for the M00 model. The cut-off was chosen as $\Lambda = 3l = 12$ and rather then solving numerically for x from Eq. (5.16), we used the expansion of x to second order, which turns out to be a good approximation in this case. The normalization is not perfect and is found to vary by a few percent from unity. The dashed lines in Fig. 1 are the Gaussian approximation of Kleffe and Borodovsky (KB)[21] with the distribution mean and variance given by Eqs. (2.15) and (2.16). The dot-dashed lines are the compound Poisson (CP) approximation.[13,14,18] For $\mathbf{x} = 0001$, which corresponds to $\mathbf{c} = 000$, both the Gaussian as well as the liquid theory approximation perform comparably, while the Compound-Poisson approximation performs poorly. For $\mathbf{x} = 1111$, corresponding to $\mathbf{c} = 111$, neither the Gaussian nor the Compound-Poisson approximation perform well. The liquid theory approximation, on the other hand, tracks rather well the actual distribution for all four possible bit-vectors $\mathbf{c}$.

The variation between actual and approximate distributions, $p(n)$ and $\hat{p}(n)$, can be quantified by the *total variational distance*[33] between the two distributions

Table II. Total variational distance between the actual distribution and the approximate distributions for the case $l = 4$, $r = 2$, $k = 256$ and the M00 Model: liquid theory approximation (L), Eq. (5.24), the liquid theory approximation normalized by an overall constant (NL), the compound poisson approximation (CP) and the gaussian approximation (KB).

c	d_{TV}^{L}	d_{TV}^{NL}	d_{TV}^{CP}	d_{TV}^{KB}
000	0.052	0.053	0.189	0.052
001	0.035	0.031	0.079	0.040
010	0.009	0.004	0.108	0.031
111	0.032	0.021	0.047	0.083

and is defined as

$$d_{TV}(p, \hat{p}) = \frac{1}{2} \sum_{n} |\hat{p}(n) - p(n)|. \tag{5.25}$$

Table 2 shows the variational distances between the actual and approximate distributions of Fig. 1 for the bit-vectors $\mathbf{c} = 000, 001, 010, 111$ associated with $\mathbf{x} = 0001, 1001, 0101, 1111$, respectively.

We see that the (un-normalized) liquid theory approximation, Eq. (5.24) (L), as well as the liquid theory approximation normalized by an overall constant (NL) perform better than the compound poisson (CP) and gaussian approximation (KB). Appendix A.2 contains the total variational distances for $l = 3, 4, 5, 6, 7$ and 8 for $r = 2$ and the M00 model along with some further remarks on the quality of the liquid theory approximation.

5.4. The Equation of State and its Virial Expansion

The expansion of x, Eq. (5.22), is in fact the virial expansion of the equation of state, Eq. (5.16), for the (discrete) lattice gas. To see this, note that the parameter x is related to u as $x = 1/u - 1$, Eq. (5.17). In the continuous 1d gas of n particles in a "volume" L and nearest-neighbor interactions, the partition function can be written as[35,36]

$$Q(n, L) = \frac{1}{2\pi i} \oint ds e^{sL} D^2(s) H^{n-1}(s) \tag{5.26}$$

where $D(s)$ and $H(s)$ are the Laplace transforms of the Boltzmann factor for the particle-boundary and particle-particle interactions, and Eq. (5.26) is the inverse Laplace transform with appropriately chosen contour. For physical interactions and in the thermodynamical limit, the integral in the above equation can be evaluated by a saddle point expansion around the point s_0[36] and, it turns out that $s_0 = \beta P$, where β is the Boltzmann factor and P is the pressure.[35,36] Comparing with Eq. (5.15) we see that upon discretizing the length of the container by letting

$L = (m + 1)\Delta$, and assuming that the interactions vary slowly with respect to Δ, Eq. (5.15) can be recovered under the identification

$$e^{s_0\Delta} = \frac{1}{u} = 1 + x, \tag{5.27}$$

which for small Δ implies that x = $s_0\Delta = \beta P\Delta$. We thus see that the virial expansion Eq. (5.22) leads to a van der Waals type equation of state.[37]

Figure 4 shows the "$P - V$ isotherms" for the four equivalence classes $\mathbf{c} = 000, 001, 010$ and 111 (from top to bottom) of the lattice gas corresponding to the M00 model with $l = 4$, $r = 2$ and fixed particle number $n = 15$. The thick solid line is the "ideal gas" law $xm = n$. The data points have been obtained from numerically solving Eq. (5.16). Note that the "compressibility," $-V^{-1}\partial V/\partial P$[37] is positive throughout and that for small m, the compressibility increases with decreasing χ. This is due to the possibility of overlapping occurrences of the words: at small m, a gas corresponding to words of the type $\mathbf{c} = 111$ is more compressible then one with $\mathbf{c} = 000$. In the limit $m \to \infty$, all isotherms approach the ideal gas law, since the average separation between particles is large and thus the interactions, which depend on $\mathbf{c}$, become increasingly negligible.

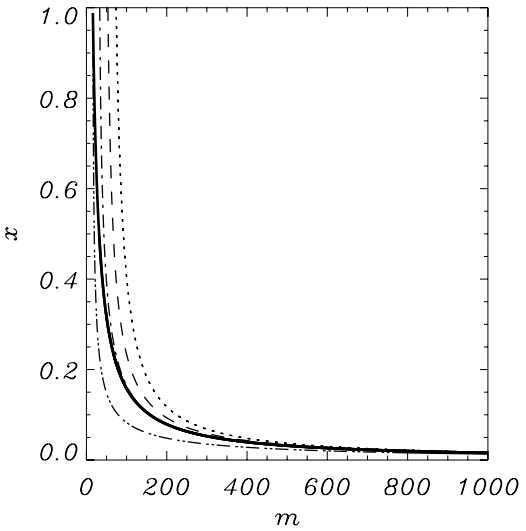

Fig. 4. The "P-V diagram" of the lattice gas with $l = 4$, $r = 2$ and fixed particle number $n = 15$ for the four possible interactions: $\mathbf{c} = 000, 001, 010$ and 111 (from top to bottom). The thick solid line corresponds to the "ideal gas" law *x=n/m* (refer to text for details).

Also note that the virial coefficients ϵ_1 and ϵ_2 remain finite as $\Lambda \to \infty$. Rewriting Eq. (5.19) as, *cf.* Eqs. (5.10) and (5.11),

$$\epsilon_1 = \Gamma_\Lambda(1) + \sum_{b=0}^{\Lambda-1} \left[1 - \frac{2}{n+1} e^{-\beta U_b(b)} - \frac{n-1}{n+1} e^{-\beta U(b)} \right], \tag{5.28}$$

it is seen that the quantity in rectangular brackets approaches zero as b increases. Since the interactions decay exponentially, the overall sum remains finite for all values of Λ, while Γ_Λ goes to zero in the same limit. In fact, ϵ_1 establishes an effective particle size and the term $(n+1)\epsilon_1$ in the expansion of x is simply the excluded volume, similar to a van der Waals type of equation of state.[37] By noting that both $D'_\Lambda(1)$ and $H'_\Lambda(1)$ grow as $\Lambda(\Lambda-1)/2$, one can similarly shown that ϵ_2 remains finite as $\Lambda \to \infty$.

5.5. Remarks

In concluding this section we should note that the approximation method presented here is very different from an approach of Régnier and Szpankowski [15] to calculate the probability of n occurrences in the limit when $n \sim \mathcal{O}(1)$. Upon deriving the generating function of the distribution, Eq. (3.10), the authors perform a Laurent expansion around its dominant $n+1$ order pole at z_1. In terms of the lattice gas description, such an expansion is asymptotic in the range of the interactions and thus might capture more accurately the tail of the interaction rather than its core. Or, in order to circumvent this, many terms must be kept in the analytic part of the expansion so that the core part is captured to a sufficient degree of accuracy.[32] However, as we have shown the characteristic core energies are of order $\ln p(\mathbf{x})$, while those of the tail go as $p(\mathbf{x})$. Therefore the core of the interaction should be retained as much as possible and only the tail should be treated perturbatively. This is precisely what our approximation achieves by introducing a cut-off distance Λ and thus removing the restriction on n.

6. ASYMPTOTICS

We now consider the asymptotic form of the n-match distributions in the limit that the length $k = m + l$ of the random string is large while $p(\mathbf{x})$, the probability of encountering $\mathbf{x}$ is small and will be used as the expansion parameter. Define the generating function $p(\zeta, m; \mathbf{x})$ of $p(n, m; \mathbf{x})$ as

$$p(\zeta; m, \mathbf{x}) = \sum_{n=0}^{\infty} p(n; m, \mathbf{x}) \zeta^n. \tag{6.1}$$

One finds, using Eq. (5.5),

$$p(\zeta; m, \mathbf{x}) = \frac{A_1}{z_1^{m+1}} + \frac{A_1}{z_1^{m+1}} \sum_{n=1}^{\infty} \left(\zeta e^{\beta\mu}\right)^n \frac{1}{2\pi i} \oint_{\partial D} dz \frac{1}{z^{m+1}} \quad D^2(z) H^{n-1}(z). \tag{6.2}$$

Recall that z_1 is the zero of smallest modulus of $\lambda(z; \mathbf{x})$, *cf.* Eqs. (3.9) and (3.13), $e^{\beta\mu}$ and A_1 are as defined in Eqs. (3.11) and (3.14), while $D(z)$ and $H(z)$ are the generating functions of the Boltzmann factors of the interactions which are given by Eqs. (5.1) and (5.2), respectively. In the above expression we have used the asymptotic form $p(0; m, \mathbf{c}) = A_1/z_1^{m+1}$ for the $n = 0$ term, since m is assumed to be large. The order of summation and integration can be exchanged if the integrand is uniformly converging in the region of integration. It is not hard to show that this is the case by considering a circular path $|z| = R$, with a suitably chosen R. Thus carrying out the sum first, we obtain

$$p(\zeta; m, \mathbf{x}) = \frac{A_1}{z_1^{m+1}} + \frac{A_1}{z_1^{m+1}} \zeta e^{\beta\mu} \frac{1}{2\pi i} \oint_{\partial D} dz \frac{1}{z^{m+1}} \quad \frac{D^2(z)}{1 - \zeta e^{\beta\mu} H(z)}. \tag{6.3}$$

Substituting the approximate forms for $H(z)$ and $D(z)$, Eqs. (5.13) and (5.14), we find

$$\begin{aligned} \hat{p}\,(\zeta; \Gamma_\Lambda, m, \mathbf{x}) = {} & \frac{A_1}{z_1^{m+1}} \\ & + \frac{A_1}{z_1^{m+1}} \frac{\zeta e^{\beta\mu}}{2\pi i} \oint_{\partial D} \frac{dz}{z^{m+1}} \frac{1}{1 - z} \\ & \times \frac{[z^\Lambda + (1 - z)(D_\Lambda(z) + \Gamma_\Lambda(z))]^2}{(1 - z)[1 - \zeta e^{\beta\mu}(H_\Lambda(z) + \Gamma_\Lambda(z))] - \zeta e^{\beta\mu} z^\Lambda}. \end{aligned} \tag{6.4}$$

Define $\bar{\lambda}(z; \zeta, \mathbf{x})$ as

$$\bar{\lambda}(z; \zeta, \mathbf{x}) = (1 - z)\left[1 - \zeta e^{\beta\mu}(H_\Lambda(z) + \Gamma_\Lambda(z))\right] - \zeta e^{\beta\mu} z^\Lambda. \tag{6.5}$$

Since $\exp(\beta\mu)$, is of order $p(\mathbf{x})$, *cf.* Eq. (3.11), $\bar{\lambda}(z; \zeta, \mathbf{x})$ has a root near $z = 1$. It turns out again that this is the root closest to the origin and that all other roots are of order $|z|^\Lambda \zeta \exp(\beta\mu) \sim 1$. Denoting the root of smallest magnitude by $\bar{z}_1$, an expansion of $\bar{z}_1$ in powers of $p(\mathbf{x})$ can be performed and one finds to lowest order

$$\bar{z}_1 = 1 - \frac{\zeta e^{\beta\mu}}{1 - \zeta e^{\beta\mu} H_\Lambda(1) - \zeta e^{\beta\mu} \Gamma_\Lambda(1)}. \tag{6.6}$$

For large m, the contour integral, Eq. (6.4), can be evaluated approximately by pushing the countour out to infinity and keeping only the residues from the

dominant poles at $z = 1$ and $z = \bar{z}_1$ so that

$$\hat{p}(\zeta;\Gamma_\Lambda,m,\mathbf{x}) = \frac{A_1}{(z_1\bar{z}_1)^{m+1}}\frac{\zeta e^{\beta\mu}}{1-\bar{z}_1}\left(-\frac{1}{\bar{\lambda}'(\bar{z}_1;\zeta,\mathbf{x})}\right) \times\left[(1-\bar{z}_1)\left(D_\Lambda(\bar{z}_1)+\Gamma_\Lambda(\bar{z}_1)\right)+\bar{z}_1^\Lambda\right]^2. \tag{6.7}$$

Notice that the m dependence is entirely confined to the term $1/(z_1\bar{z}_1)^{m+1}$. Thus this term alone is responsible for the large m behavior. The term in square brackets is the effect due to the boundaries of the string. When m is large, boundary effects should not matter and we will set this term to 1. Alternatively, we can assume that the random string is circular in which case the boundary term will not arise.

Apart from the cut-off assumption on the behavior of the tails, and the assumption of large m leading to the m-asymptotic expression, Eq. (6.7), we have not made any assumptions on $p(\mathbf{x})$ so far. To proceed further, we will assume that $p(\mathbf{x}) \ll 1$ so that the lowest order expressions for z_1 and $\bar{z}_1$, Eqs. (3.13) and (6.6), will provide the leading order approximation to Eq. (6.7).

Substituting the lowest order expression for $\bar{z}_1$, Eq. (6.6), and noting that to this order $-\bar{\lambda}'(\bar{z}_1;\zeta,\mathbf{x}) = 1-\zeta\exp(\beta\mu)H_\Lambda(1)-\zeta\exp(\beta\mu)\Gamma_\Lambda(1)$, the result simplifies to

$$\hat{p}(\zeta;\Gamma_\Lambda,m,\mathbf{x}) = \frac{A_1}{(z_1\bar{z}_1)^{m+1}}. \tag{6.8}$$

6.1. The Compound Poisson Approximation

The compound poisson distribution arises in the limit when $m\to\infty$ and $p(\mathbf{x})\to 0$ such that $\langle n\rangle = (m+1)p(\mathbf{x})$, Eq. (2.15), is finite. Obviously, a zero matching probability renders the problem meaningless, instead we are interested in small but non-zero $p(\mathbf{x})$ as can be obtained varying the letter distribution or increasing the length l of the word $\mathbf{x}$. This implies that $p(\mathbf{x})\sim 1/(m+1)$, or if one lets the length l of $\mathbf{x}$ increase, $l\sim\ln m$. From the properties of the inter-particle interactions that were derived in Sec. 4, we see that in this limit the strength of the tail is of order $1/m$ and hence very weak, while the strength of the core is of order $\log m$ and therefore relatively strong. Thus it is permissible to set $\Lambda = l$ and ignore the tails, $\Gamma_\Lambda = 0$. Since $p(\mathbf{x})\sim 1/m$, to lowest order $A_1 = 1$, and to the same order we find from Eqs. (3.11) and (3.13) that

$$e^{\beta\mu} = \frac{p(\mathbf{x})}{[1+c(1;\mathbf{x})]^2}, \tag{6.9}$$

where $c(z;\mathbf{x})$ is given by Eq. (3.4). We obtain

$$\hat{p}(\zeta;0,m,\mathbf{x}) = \left[\left(1+\frac{p(\mathbf{x})}{1+c(1;\mathbf{x})}\right) \times\left(1-\zeta\frac{p(\mathbf{x})}{[1+c(1;\mathbf{x})]^2}\frac{1}{1-\zeta e^{\beta\mu}H_l(1)}\right)\right]^{-(m+1)}, \quad (6.10)$$

which can be further simplified by noting that from Eq. (3.8) to order lowest order in $p(\mathbf{x})$

$$\frac{1}{1+c(1;\mathbf{x})} = 1-\sum_{b=1}^{l-1} h(b;\mathbf{x}) \equiv 1-\alpha_l(\mathbf{x}). \quad (6.11)$$

Likewise, using Eqs. (5.2), (5.4) and (5.11), one finds to lowest order ($z_1 \approx 1$) that $e^{\beta\mu}H_l(1) = \alpha_l(\mathbf{x})$.

Multiplying out the product in Eq. (6.10) and keeping only terms to order $p(\mathbf{x}) \sim 1/m$, we obtain

$$\hat{p}(\zeta;0,m,\mathbf{x}) = \left[1+p(\mathbf{x})\,(1-\alpha_l(\mathbf{x}))^2\left(\frac{1}{1-\alpha_l(\mathbf{x})}-\frac{\zeta}{1-\zeta\alpha_l(\mathbf{x})}\right)\right]^{-(m+1)}, \quad (6.12)$$

which upon taking the limit $m\to\infty$ such that $(m+1)p(\mathbf{x}) = \langle n\rangle$ is finite, is readily brought to the form

$$\hat{p}(\zeta;0,m,\mathbf{x}) = \exp\left(-\sum_{j=1}^{\infty}\left(1-\zeta^j\right)\bar{\lambda}_j\right) \quad (6.13)$$

with

$$\bar{\lambda}_j = \langle n\rangle\,[1-\alpha_l(\mathbf{x})]^2\,\alpha_l(\mathbf{x})^{j-1}. \quad (6.14)$$

Note that from Eqs. (4.3) and (6.11) we have

$$\alpha_l(\mathbf{x}) = \sum_{b\in\mathcal{P}'}\mathrm{Prob}\{\mathbf{y}_{l,b} = \mathbf{x}_{l-b,b}\,|\,\mathbf{y}_{0,l} = \mathbf{x}_{0,l}\}, \quad (6.15)$$

where $\mathcal{P}'$ is the set of principal periods of $\mathbf{x}$. Thus $\alpha_l(\mathbf{x})$ is the probability that the next occurrence of $\mathbf{x}$ is less than a distance l apart, i.e. overlapping.

Eq. (6.13) is the generating function of a compound poisson distribution[34] and has been derived by various other methods, by Chrysaphinou and Papastavridis,[13] Geske *et al.*[14] and Schbath[18] for the different models of letter distributions.

Note that setting the tails ($b \geq l$) of the interactions to zero implies that given the next occurrence of $\mathbf{x}$ is a distance at least l away, it can occur with equal probability at any $b \geq l$. Since nearest neighbour match separations b with $b < l$

define an overlapping cluster, this means that the starting points of the clusters, are distributed like the arrivals of a poisson process.[13,17,19] We therefore see that the liquid theory description in terms of interactions along with the separation of cores and tails provides an alternative and very simple explanation of this property. Conversely, strong tails imply significant deviations from the poissonian occurrences of cluster initiation sites, meaning that the locations of the clusters themselves are correlated.

6.2. The Gaussian Approximation

We now consider the limit $m, n \to \infty$, such that the number density $n/(m+1) = p(\mathbf{x})$ remains constant and is small. The tails of the interaction are therefore weak, and to this order $A_1 = 1$, so that Eq. (6.8) becomes $\hat{p}(\zeta; \Gamma_\Lambda, m, \mathbf{x}) = 1/(z_1 \bar{z}_1)^{m+1}$, where $\bar{z}_1$ is given by Eq. (6.6). For the distribution to be normalized, we must have $\hat{p}(\zeta; \Gamma_\Lambda, m, \mathbf{x}) = 1$, when we set $\zeta = 1$. Noting that the ζ dependence is contained entirely in $\bar{z}_1$, this implies that to lowest non-trivial order in $p(\mathbf{x})$

$$[z_1 \bar{z}_1]_{\zeta=1} = 1 + \mathcal{O}(p(\mathbf{x})). \tag{6.16}$$

Choosing therefore the cut-off function $\Gamma_\Lambda(1)$ contained in $\bar{z}_1$ so that the above condition is satisfied, we obtain

$$\hat{p}(\zeta; m, \mathbf{x}) = \left[(1 + \eta p(\mathbf{x})) \left(1 - \zeta p(\mathbf{x}) \frac{\eta^2}{z_1} \frac{1}{1 - \zeta \left(1 - \frac{\eta}{z_1} \right)} \right) \right]^{-(m+1)}, \tag{6.17}$$

where we have defined

$$\eta = \frac{z_1 - 1}{p(\mathbf{x})}, \tag{6.18}$$

which from Eq. (3.13) is given to leading order in $p(\mathbf{x})$ as $1/\eta = 1 + c(1; \mathbf{x}) + p(\mathbf{x}) T(1)$.

The large n limit can again be obtained using Hayman's method[32] introduced in Sec. 5.1. Choosing ζ_0 such that

$$n = \left(\zeta \frac{\mathrm{d}}{\mathrm{d}\zeta} \ln \hat{p}(\zeta; m, \mathbf{x}) \right) \Bigg|_{\zeta=\zeta_0} \tag{6.19}$$

we find to lowest order in $\zeta - 1$

$$\zeta_0 - 1 = \frac{n - \langle n \rangle}{\langle n \rangle \left(\frac{2}{\eta} - 1 \right)}, \tag{6.20}$$

where $\langle n \rangle = (m+1) p(\mathbf{x})$, Eq. (2.15). Using this approximation for ζ_0, and expanding the resulting approximation $\ln \hat{p}(n; m, \mathbf{x})$ as a Taylor series in $\zeta_0 - 1$

around $\zeta = 1$, which turns out to be a cumulant expansion, one finds that the third and higher order terms vanish in the limit $m, n \to \infty$, if the fluctuations of n around its average $\langle n \rangle$ are of order $\sqrt{n}$. Thus requiring $n - \langle n \rangle$ to be of order $\sqrt{n}$ or less, the distribution in the limit $m, n \to \infty$ is Gaussian. Taking care to collect all relevant contributions, one finds after some tedious but otherwise straight-forward algebra that

$$\hat{p}(n; m, \mathbf{x}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}} \exp\left(-\frac{(n - \langle n \rangle)^2}{2\hat{\sigma}_n^2}\right), \tag{6.21}$$

with its variance given by

$$\hat{\sigma}_n^2 = \langle n \rangle \left(\frac{2}{\eta} - 1\right). \tag{6.22}$$

Substituting for η, Eq. (6.18), and using the leading order expression for z_1, Eq. (3.13), this reduces to

$$\hat{\sigma}_n^2 = \langle n \rangle \ [1 + 2c(1; \mathbf{x}) + 2p(\mathbf{x})T(1)], \tag{6.23}$$

which, as can be readily verified, is the leading order term of the exact variance, Eq. (2.16). The Gaussian form of the distribution, Eq. (6.21), with variance given by Eq. (6.23) is the first order perturbation result of the low density expansion of the lattice gas in the thermodynamical limit, $m, n \to \infty$. As evident from the result above, the correct mean of the distribution is already established at this order, while higher order perturbations apparently only add corrections to the variance, so that it eventually converges to the exact expression, Eq. (2.16).

Note that the derivations of the Compound Poisson and Gaussian asymptotic forms as presented in this section are based on determing the dominant root of $\bar{\lambda}(z; \zeta, \mathbf{x})$, Eq. (6.5), which in turn emerges as a result of introducing a cut-off Λ and approximating the interactions beyond Λ. Hence, the introduction of a cut-off turns out to be a convenient way to handle the tail of the interactions.

7. DISCUSSION

We have presented a new approach to calculating the probability distribution for the number of occurrences of a given word inside a random string of letters. Our approach rests on the observation that the probability distribution can be interpreted as the n-particle term of the grand partition function for a gas of particles on a linear lattice, with pairwise nearest neighbor interactions. By exploiting this analogy and focusing on the generic properties of the interactions, we have derived an equation of state for the lattice gas and thereby obtained an analytical

expression for the probability distribution of n occurrences, which besides interpolating between the known asymptotic forms of the distribution, also provides a good approximation in the intermediate regime.

The identification and subsequent analysis of the effective inter-particle interactions of the lattice gas description turns out to be key in our approach to this problem. The interactions are characterized by a strong core-region of the size of the word length, whose energy scale is logarithmic in the probability of word occurrence, followed by a relatively weak and exponentially decaying tail whose characteristic energy is of the order of the probability itself. Our results are valid for a broad class of random letter sequences, including those generated from non-uniform i.i.d. letter distributions as well as those generated by Markov chains of order s with $s \leq l$. We have shown that the details of these underlying stochastic processes only affect the form of the interactions of the lattice gas, which remain pairwise and of nearest neighbor type. We also have shown that the generic features of these interactions, namely a relatively strong core and an exponentially decaying weak tail, are robust. Furthermore, the core of the interaction is found to depend only on the probability of occurrences of the word $\mathbf{x}$ and its suffixes along with the overlap properties of $\mathbf{x}$ as given by the bit-vector of Guibas and Odlyzko.[8]

The lattice gas description can be extended to calculate the probability of the number of occurrences of a given set of words.[38,39] In terms of the lattice gas analogy, this corresponds to a mixture, where each word is a different kind of particle and one has different type of interactions between the kinds of particles.

Lastly, in the theory of liquids, which underlies the lattice gas description of our approach, the spatial correlation functions and their behavior are a result of the particle interactions and can be used to recover the latter from the former. In terms of the lattice gas description of the string matching problem this means that one can similarly determine the effective particle interactions from spatial correlations of word occurrences, as provided for example by the pair correlation function. General assumptions such as stationarity and the Markov property, result in interactions that are pair-wise and of nearest-neighbour type, with the properties of the underlying stochastic process determining only the functional form of the interactions. Thus determining interactions from such correlation constitutes a way of estimating or determining the underlying stochastic model. In many applications of sequence matching, such as the analysis of DNA sequences, it is essential to have an accurate stochastic model in order to obtain good estimates for the statistical significance of certain events.[19,29] Given a randomly generated sequence of letters, the lattice gas approach to string matching should therefore also be applicable to the determination of the underlying stochastic model or the estimation of its parameters.

APPENDIX A

A.1. The Factorized Form of the Probability of Occurrences

In this appendix we outline the derivation of Eqs. (2.7), (2.10), (2.11), (2.12), and (2.13).

The probability of n occurrences can be written as, *cf.* Eqs. (2.5) and (2.6),

$$p(n;m,\mathbf{x}) = \sum_{\mathbf{y}} \text{Prob}\{\mathbf{y}\} \sum_{a_1<a_2<\cdots<a_n} I(a_1,a_2,\ldots a_n;\mathbf{x},\mathbf{y}), \tag{A.1}$$

where

$$I(a_1,a_2,\ldots a_n;\mathbf{x},\mathbf{y}) = \left[\prod_{i_1=1}^{a_1-1}(1-f_{i_1})\right] f_{a_1} \left[\prod_{i_2=a_1+1}^{a_2-1}(1-f_{i_2})\right] f_{a_2}\cdots \times \left[\prod_{i_n=a_{n-1}+1}^{a_n-1}(1-f_{i_n})\right] f_{a_n} \left[\prod_{i_{n+1}=a_n+1}^{m}(1-f_{i_{n+1}})\right]. \tag{A.2}$$

with $I(a_1,a_2,\ldots a_n;\mathbf{x},\mathbf{y})$ being the indicator function for the event that the word $\mathbf{x}$ occurs precisely n times and the occurrences are at positions $a_1 < a_2 < \ldots < a_n$.

Since $f_a(\mathbf{x},\mathbf{y}) \in \{0,1\}$ we can write,

$$f_a(\mathbf{x},\mathbf{y}) = f_a(\mathbf{x},y_1,\ldots,y_k) f_a(\mathbf{x},y_1\ldots,y_a,\tilde{y}_{a+1},\ldots,\tilde{y}_{a+l},y_{a+l+1},\ldots,y_k), \tag{A.3}$$

where we have introduced another set of random variables $\tilde{\mathbf{y}}_{a,l} = \tilde{y}_{a+1},\ldots,\tilde{y}_{a+l}$ for the segment at a. Note that for a non-zero value of f_a we must have $\mathbf{x} = \mathbf{y}_{a,l}$ as well as $\mathbf{x} = \tilde{\mathbf{y}}_{a,l}$, and therefore by implication $\mathbf{y}_{a,l} = \tilde{\mathbf{y}}_{a,l}$.

Assuming that $s \le l$ and using the stationarity property, the matching probability $p(n;m,\mathbf{x})$ thus factorizes as

$$p(n;m,\mathbf{x}) = p(\mathbf{x}) \sum_{a_1<a_2<\cdots<a_n} d_L(a_1;\mathbf{x}) \left[\prod_{i=1}^{n-1} h(a_{i+1}-a_i;\mathbf{x})\right] d_R(m-a_n;\mathbf{x}), \tag{A.4}$$

with

$$d_L(b;\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_{y_1\ldots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l}\} f_0 \left[\prod_{a=1}^{b}(1-f_a)\right], \tag{A.5}$$

$$h(b;\mathbf{x}) = \sum_{y_1\ldots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l}|\mathbf{y}_{0,l}=\mathbf{x}\} f_0 \left[\prod_{a=1}^{b-1}(1-f_a)\right] f_b, \tag{A.6}$$

and

$$d_R(b;\mathbf{x}) = \sum_{y_1\cdots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l}|\mathbf{y}_{0,l}=\mathbf{x}\} f_0 \left[\prod_{a=1}^{b}(1-f_a)\right]. \tag{A.7}$$

With $d_L(0;\mathbf{x}) = d_R(0;\mathbf{x}) = 1$, it is readily shown that $d_L(b;\mathbf{x}) = d_R(b;\mathbf{x}) \equiv d(b;\mathbf{x})$ and thus Eq. (A.4) becomes

$$p(n;m,\mathbf{x}) = p(\mathbf{x}) \sum_{a_1<a_2<\cdots<a_n} \frac{1}{r^{m+l}} d(a_1;\mathbf{x}) \left[\prod_{i=1}^{n-1} h(a_{i+1}-a_i;\mathbf{x})\right] d(m-a_n;\mathbf{x}), \tag{A.8}$$

which is Eq. (2.7).

A recursion relation for $d(b;\mathbf{x})$ can be obtained by factoring out the $a = b$ term in Eq. (A.7),

$$\begin{aligned} d(b;\mathbf{x}) = &\sum_{y_1\cdots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l}|\mathbf{y}_{0,l}=\mathbf{x}\} f_0 \left[\prod_{a=1}^{b-1}(1-f_a)\right] \\ &- \sum_{y_1\cdots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l}|\mathbf{y}_{0,l}=\mathbf{x}\} f_0 \prod_{a=1}^{b-1}[(1-f_a)]\, f_b. \end{aligned} \tag{A.9}$$

The argument of the first sum does not contain the variable y_{b+l} and the sum over the remaining variables yields $d(b-1;\mathbf{x})$, while the second sum is $h(b;\mathbf{x})$, Eq. (A.6). Thus,

$$d(b;\mathbf{x}) = d(b-1;\mathbf{x}) - h(b;\mathbf{x}). \tag{A.10}$$

We next seek a recursion relation for $h(b;\mathbf{x})$. Using the algebraic identity,

$$\prod_{a=0}^{m}(1-f_a) = 1 - \sum_{b=0}^{m} f_b \prod_{a=0}^{b-1}(1-f_a), \tag{A.11}$$

which is readily proven by induction, we find from Eq. (A.6) that

$$h(b;\mathbf{x}) = \sum_{y_1\cdots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l}|\mathbf{y}_{0,l}=\mathbf{x}\} \left\{ f_0 f_b - \sum_{c=1}^{b-1} f_0 \left[\prod_{a=1}^{c-1}(1-f_a)\right] f_c f_b \right\}. \tag{A.12}$$

Using the factorization property, Eq. (A.3) on the f_c term this becomes

$$h(b;\mathbf{x}) = C(b;\mathbf{x}) - \sum_{a=1}^{b-1} h(a;\mathbf{x}) C(b-a;\mathbf{x}), \tag{A.13}$$

with the function $C(b;\mathbf{x})$ defined as, *cf.* Eq. (2.12),

$$C(b;\mathbf{x}) = \sum_{y_1\cdots y_{b+l}} \text{Prob}\{\mathbf{y}_{0,b+l}|\mathbf{y}_{0,l} = \mathbf{x}\} f_0(\mathbf{x},\mathbf{y}) f_b(\mathbf{x},\mathbf{y}). \tag{A.14}$$

It can be easily shown that for $b \geq l$

$$\begin{aligned} C(b;\mathbf{x}) &= p(\mathbf{x}) \sum_{\mathbf{y}_{l,b-l}} \text{Prob}\{\mathbf{y}_{0,b+l}|\mathbf{y}_{0,l} = \mathbf{y}_{b,l} = \mathbf{x}\}, \\ &= \frac{p(\mathbf{x})}{\mu(\mathbf{x}_{0,s})} \Pi^{(b-l)}(\mathbf{x}_{l-s,s} \to \mathbf{x}_{0,s}), \end{aligned} \tag{A.15}$$

where $\Pi^{(a)}(\mathbf{x}_{l-s,s} \to \mathbf{x}_{0,s})$ is the a step transition probability from $\mathbf{x}_{l-s,s}$ to $\mathbf{x}_{0,s}$.

A.2. Total Variational Distances

Table 3 below shows the variational distances between the actual and approximate distributions for words of length $l = 3, 4, 5, 6, 7, 8$ and their associated sets of bit-vectors $\mathbf{c}$. The letters of the random string were generated according to the M00 model with $r = 2$, for lengths $k = 128, 256, 512, 1024, 2048$ and 4098, chosen so that the average number of occurrences is about the same. The liquid theory distributions were calculated with $\Gamma_\Lambda = 0$ and $\Lambda = 3l$. The values associated with the distributions of Fig. 1, correspond to $\mathbf{c} = 000, 001, 010, 111$ for $\mathbf{x} = 0001, 1001, 0101, 1111$, respectively.

The (un-normalized) liquid theory approximation, Eq. (5.24) (L), as well as the liquid theory approximation normalized by an overall constant (NL) perform better than or at worst comparably with the compound poisson (CP) and gaussian approximation (KB). Also, the liquid theory approximations (XL) with x determined numerically from Eq. (5.18), does not overall perform distinctly better than the distributions obtained from the expansion of x to second order. The only exception is the case $l = 3$ with $\mathbf{c} = 00$, where due to the short length of the string, the tail of the inter-particle interaction turns out to be rather strong, as evident from Fig. 3, and one therefore has to expand x to higher order.

We have checked that using a larger cut-off does not improve the distributions very much, thereby justifying setting the interactions to zero beyond the cut-off. It turns out that for large χ and l, the first order expression for x is often sufficient. It is however almost always insufficient for small χ and in particular when $\chi = 1$, i.e. x belongs to the equivalence class $\mathbf{c} = 11\ldots1$.

Note from Fig. 1 that for $\mathbf{x} = 0001$ ($\mathbf{c} = 000$) none of the approximations captures the height of the peak of the distribution accurately. This discrepancy is persistent: it does not improve with increasing Λ, solving numerically for x from Eq. (5.16) rather than approximating it through an expansion, or by taking the stationary phase approximation to higher order, which turns out to be a $1/n$

Table III. Total variational distance between the actual distribution and the various approximate distribution for the case $r = 2$, and the M00 Model: Shown are the cases for words of length $l = 3,4,5,6,7,8$ inside random strings of length $k = 128,256,512,1024,2048\ 4098$. The liquid theory distributions were calculated with $\Gamma_\Lambda = 0$ and $\Lambda = 3l$. The values correspond to (L): liquid theory approximation with x determined by the 2nd order expansion, Eq. (5.22), (NL): the distribution (L) normalized by an overall constant, (XL): liquid theory approximation with x determined from solving numerically Eq. (5.18), (CP) the compound poisson approximation and (KB): the gaussian approximation.

c	d_{TV}^{L}	$\bar{d}_{TV}^{NL}$	$\bar{d}_{TV}^{XL}$	d_{TV}^{CP}	d_{TV}^{KB}
00	******	0.961	0.006	0.227	0.006
01	0.030	0.008	0.010	0.156	0.013
11	0.019	0.016	0.022	0.121	0.073
000	0.052	0.053	0.053	0.189	0.052
001	0.035	0.031	0.031	0.079	0.040
010	0.009	0.004	0.004	0.108	0.031
111	0.032	0.021	0.022	0.047	0.083
0000	0.009	0.010	0.010	0.090	0.018
0001	0.018	0.016	0.016	0.056	0.031
0010	0.010	0.008	0.008	0.061	0.031
0011	0.041	0.036	0.037	0.034	0.054
0101	0.021	0.024	0.022	0.075	0.043
1111	0.044	0.026	0.030	0.012	0.089
00000	0.013	0.011	0.011	0.034	0.028
00001	0.006	0.004	0.004	0.040	0.028
00010	0.009	0.011	0.011	0.053	0.028
00011	0.018	0.019	0.019	0.061	0.032
00100	0.013	0.011	0.011	0.032	0.037
00101	0.010	0.006	0.007	0.037	0.039
01010	0.019	0.011	0.008	0.042	0.047
11111	0.049	0.027	0.031	0.011	0.090
000000	0.025	0.024	0.024	0.004	0.037
000001	0.003	0.003	0.003	0.028	0.027
000010	0.004	0.002	0.002	0.023	0.030
000011	0.005	0.006	0.006	0.031	0.031
000100	0.004	0.002	0.002	0.023	0.034
000101	0.004	0.003	0.003	0.029	0.034
000111	0.004	0.003	0.003	0.029	0.036
001001	0.011	0.012	0.012	0.033	0.039
010101	0.023	0.013	0.009	0.026	0.049
111111	0.052	0.022	0.025	0.015	0.088
0000000	0.023	0.022	0.022	0.009	0.040
0000001	0.022	0.021	0.021	0.006	0.039
0000010	0.004	0.002	0.002	0.013	0.031
0000011	0.018	0.017	0.017	0.004	0.038
0000100	0.003	0.002	0.002	0.014	0.032

Table III. Continued.

c	d_{TV}^{L}	$\bar{d}_{TV}^{NL}$	$\bar{d}_{TV}^{XL}$	d_{TV}^{CP}	d_{TV}^{KB}
0000101	0.010	0.008	0.008	0.007	0.035
0000111	0.003	0.002	0.002	0.014	0.035
0001000	0.003	0.003	0.003	0.017	0.034
0001001	0.005	0.003	0.003	0.012	0.036
0010010	0.005	0.005	0.005	0.017	0.040
0010011	0.010	0.007	0.007	0.007	0.043
0101010	0.025	0.009	0.003	0.012	0.052
1111111	0.054	0.020	0.020	0.015	0.086

expansion. The discrepancy for $\mathbf{c} = 000$ does not seem to be a finite-size effect either. We have checked that increasing the string length to $m = 4092$ does not remove this discrepancy. It thus appears that for the case $\mathbf{c} = 000$ and $r = 2$, the stationary phase approximation around the single point $u \approx 1$ is not fully capturing the probability distribution.

ACKNOWLEDGMENTS

I would like to thank Ayşe Erzan for initially bringing to my attention the string matching problem as well as for her critical reading of the manuscript. This work was supported in part by the Nahide and Mustafa Saydan Foundation and Tübitak, the Turkish Science and Technology Research Council.

REFERENCES

1. S. B. Boyer and J. S. Moore, *Comm. ACM* **20**:762 (1977).
2. D. E. Knuth, J. H. Morris and V. R. Pratt, *SIAM J. Comput.* **6**:323 (1977).
3. P. Pevzner, M. Bordovsky and A. Mironov, *J. Biomol. Struct. Dyn.* **6**:1013 (1991).
4. B. Prum, F. Rodolphe and E. Turckheim, *J. Roy. Stat. Soc. Ser.* **B57**:205 (1995).
5. S. Karlin and S. F. Altshul, *Proc. Natl. Acad. Sci. USA* **90**:5873 (1993).
6. A. Dembo and S. Karlin, *Ann. Prob.* **19**:1773 (1991).
7. L. J. Guibas and A. M. Odlyzko, *SIAM J. Appl. Math.* **35**:401 (1978).
8. L. J. Guibas and A. M. Odlyzko, *J. Comb. Theory* **30A**:19 (1981).
9. L. J. Guibas and A. M. Odlyzko, *J. Comb. Theory* **30A**:183 (1981).
10. O. Chrysaphinou and S. Papastavridis, *J. Appl. Prob.* **25**:428 (1988).
11. O. Chrysaphinou and S. Papastavridis, *Probab. Th. Rel. Fields* **79**:129 (1988).
12. I. Fudos, E. Pitoura and W. Szpankowski, *Inform. Process. Lett.* **57**:307 (1996).
13. O. Chrysaphinou and S. Papastavridis, *Theory Probab. Appl.* **35**:145 (1990).
14. M. X. Geske, A. P. Godbole, A. A. Schaffner, A. M. Skolnick and G. L. Wallstrom, *J. Appl. Prob.* **32**:877 (1995).
15. M. Régnier and W. Szpankowski, *Algorithmica* **22**:631 (1998).
16. L. Goldstein and M. S. Waterman, *Bull. Math. Biol.* **54**:785 (1992).
17. M. S. Waterman, *Introduction to Computational Biology* (Chapman & Hall, Boca Raton, 1995).

18. S. Schbath, *ESAIM Prob. Stat.* **1**:1 (1995).
19. G. Reinert, S. Schbath and M. S. Waterman, *J. Comp. Biol.* **7**:1 (2000).
20. S. Robin and S. Schbath, *J. Comp. Biol.* **8**:349 (2001).
21. J. Kleffe and M. Borodovsky, *Comp. Appl. Biosci.* **8** (1992) 433.
22. Jane F. Gentleman and R. C. Mullins, *Biometrics* **45**:32 (1999).
23. Y. Fu and P. W. Anderson, *J. Phys.* **19A**:1605 (1986).
24. R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman and L. Troyansky, *Nature* **400**:137 (1999).
25. S. Kirkpatrick and B. Selman, *Science* **264**:1297 (1994).
26. M. Mézard, G. Parisi and R. Zecchina, *Science* **297**:812 (2002).
27. S. Mertens, M. Mézard and R. Zecchina, cs.CC/0309020.
28. D. Achlioptas, A. Naor and Y. Peres, *Nature* **435**:759 (2005).
29. S. Robin and J. J. Daudin, *J. Appl. Prob.* **36**:179 (1999).
30. H. Harborth, *Zeits. f. Reine Angew. Math.* **271**:139 (1974).
31. E. Rivals and S. Rahmann, *J. Comb. Theory* **104**:95 (2003).
32. H. S. Wilf, *generatingfunctionology* (Academic Press, Boston, 1994).
33. A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation* (Oxford University Press, Oxord, 1992).
34. W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1971).
35. F. Gürsey, *Proc. Cambr. Phil. Soc.* **46**:182 (1950).
36. I. Z. Fisher, *Statistical Theory of Liquids* (University of Chicago Press, Chicago, 1964).
37. G. E. Uhlenbeck and G. W. Ford, *Lectures in Statistical Mechanics* (American Mathematical Society, Providence, 1963).
38. M. Régnier, *Discr. Appl. Math.* **104**:259 (2000).
39. S. Robin and J. J. Daudin, *Ann. Inst. Stat. Math.* **36**:895 (2001).