# Mean-field approximations to the longest common subsequence problem

J. Boutet de Monvel*
*Institute for Hearing and Communication Research, M1:00-ENT, Karolinska Hospital, 17 176 Stockholm, Sweden*
(Received 8 March 2000)

The longest common subsequence (LCS) problem is a fundamental problem of sequence comparison. A natural approximation to this problem is a model in which every pair of letters of two ''sequences'' are matched *independently* of the other pairs with probability $1/S$, $S$ representing the size of the alphabet. This model is analogous to a mean-field version of the LCS problem, which can be solved with a cavity approach [J. Boutet de Monvel, Eur. Phys. J. B **7**, 293 (1999)]. We refine here this approximation by incorporating in a systematic way correlations among the matches in the cavity calculation. We obtain a series of closer and closer approximations to the LCS problem, which we quantify in the large $S$ limit, both with a perturbative approach and by Monte Carlo simulations. We find that, as it happens in the expansion around mean field for other disordered systems, the corrections to our approximations depend upon long-ranged correlation effects that render the large $S$ expansion nonperturbative.

PACS number(s): 02.50.Ga, 64.60.Ak, 64.60.Cn, 75.10.Nr

## I. INTRODUCTION

The longest common subsequence (LCS) problem is a simple and fundamental example of a sequence comparison problem. Such problems arise under various important situations, ranging from biology to combinatorics and computational sciences [1]. A frequent problem of molecular biology is the detection of evolutionary relationships between different molecules [2]: Given two DNA molecules that evolved from a common ancestor through a process of random insertions and deletions, how can one recover the ancestor? A possible approach is to solve a particular instance of the LCS problem, namely, to look for sequences of nucleotides that appear *in the same order* in the two DNA molecules, and to pick such a common subsequence that is as long, i.e., contains as many nucleotides, as possible. Replacing the two DNA molecules by two general sequences $X=(X_1,\ldots,X_N)$ and $Y=(Y_1,\ldots,Y_M)$ (not necessarily of equal lengths) taken from a given alphabet, one obtains a general instance of the LCS problem. As it is natural to expect, when $X$ and $Y$ are very long sequences whose elements are taken at random independently from an alphabet of $S$ letters (with $S\geqslant 2$), there is a definite density of matched points in a LCS of $X$ and $Y$. More precisely, if $L_N$ denotes the length (the number of letters) of a LCS of $(X_1,\ldots,X_N)$ and $(Y_1,\ldots,Y_N)$, one can prove (see, e.g., [3]) that with probability one, $L_N/N$ tend to a nonrandom constant $\gamma_S$ as $N\to\infty$. The determination of $\gamma_S$ and of the rate at which $L_N/N$ approaches this limit are much studied combinatorial problems [4–6]. A connection with statistical physics has been provided by Hwa and Lässig [7] who found that Needleman-Wunsch sequence alignment, a popular comparison scheme for DNA and proteins of which the LCS problem is a special case [8], falls in the universality class of directed polymers in a random medium. This connection is based on a geometric interpretation (explained in the next section) of the LCS problem as a longest path problem [9]. The randomness in the above ''random string'' model can be encoded in variables $\epsilon_{ij}$ defined as occupation numbers for the matches of $X$ and $Y$, namely $\epsilon_{ij}=\delta_{X_i,Y_j}=1$ if $X_i=Y_j$ and 0 otherwise. The presence of long-ranged correlations among the matches (for example, given any indices $i_1,j_1,i_2,j_2$, the variables $\epsilon_{i_1j_1},\epsilon_{i_1j_2},\epsilon_{i_2j_1},\epsilon_{i_2j_2}$ are obviously correlated) complicates the problem very much, and to date the computation of the average length of a LCS has turned out to be intractable. In [10], we studied a related ''Bernoulli matching'' model where the $\epsilon_{ij}$'s are taken to be independent and identically distributed random variables with $P(\epsilon_{ij}=1)=1-P(\epsilon_{ij}=0)=1/S$. It turns out that this model is very analogous to a mean-field version of the LCS problem, which can be solved using a cavity approach. This solution was found to provide a very good approximation (whose precision ameliorates as the size of the alphabet increases) to the average LCS length of two random strings measured from direct Monte Carlo simulations. We pursue here the work of [10] by studying the behavior of the above ''mean-field'' approximation in the limit of large alphabets. We describe a method that allows us to refine the cavity calculation made for the Bernoulli matching model, by taking correlations of the random string model into account in a systematic way. This leads to a series of approximations getting closer and closer to the LCS problem, which we quantify within a perturbative approach valid in the limit $S\to\infty$. We find that, while our perturbative approach provides an excellent approximation to the LCS problem at finite $S$, it leads to a singular expansion (in powers of $1/\sqrt{S}$) around the Bernoulli matching model. In particular, the *leading* corrections to this mean-field approximation depend upon long-ranged correlation effects among the matches and cannot be captured by the method we use.

## II. THE CAVITY SOLUTION TO THE BERNOULLI MATCHING MODEL

Consider the lattice $\mathcal{C}_{NM}$ formed by the integer points $(ij)$, $0\leqslant i\leqslant N, 0\leqslant j\leqslant M$ together with nearest-neighbor bonds, and add a diagonal bond $\{(i-1,j-1),(ij)\}$ for each

*Electronic address: j.boutet.de.monvel@ihk.ki.se

 

point $(ij)$ such that $\epsilon_{ij}=1$ (we call such a point a *match*). Define the weight of any path on $\mathcal{C}_{NM}$ to be the number of diagonal bonds that it contains, and let $L_{ij}$ be the maximum possible weight of a directed path joining the point (0,0) to $(ij)$. In the random string model, $L_{ij}$ is just the length of a LCS of the substrings $(X_1,\ldots,X_i)$ and $(Y_1,\ldots,Y_j)$. Setting $L_{i,0}=L_{0,j}=0$, the $L_{ij}$'s satisfy the following recursion relation:

$$L_{ij}=\max(L_{i-1,j},L_{i,j-1},L_{i-1,j-1}+\epsilon_{ij}), \tag{1}$$

which follows from the fact that any directed path ending at $(ij)$ must visit one of the points $(i-1,j),(i,j-1)$, or $(i-1,j-1)$. It turns out to be more convenient to work with the *local gradient* variables $\nu_{ij}=L_{ij}-L_{i-1,j}$ and $\mu_{ij}=L_{ij}-L_{i,j-1}$, rather than with $L_{ij}$ itself. It is obvious from Eq. (1) that $\nu_{ij}$ and $\mu_{ij}$ can take only the values 0 or 1. Writing $\overline{x}=1-x$ if $x\in\{0,1\}$, the recursion relations for $\nu_{ij}$ and $\mu_{ij}$ can be written in algebraic form:

$$\nu_{ij}=(1-\overline{\epsilon}_{ij}\overline{\nu}_{i,j-1})\overline{\mu}_{i-1,j},$$

$$\mu_{ij}=(1-\overline{\epsilon}_{ij}\overline{\mu}_{i-1,j})\overline{\nu}_{i,j-1}, \tag{2}$$

with $\nu_{i,0}=\nu_{0,i}=\mu_{i,0}=\mu_{0,i}=0$. The key property that was used (but left unjustified) in [10] is that in the Bernoulli matching model the variables $\nu_{ij}$ and $\mu_{ij}$ along $i+j=t$ become *independent* in the limit $t\to\infty$. This can be viewed as a consequence of the directed polymer picture of [7], if we interpret $L_{ij}$ as the height profile $L(x,t)$ (as a function of $x=i-j$ and $t=i+j$) of a growing one-dimensional interface, described in a continuum limit by the Kardar-Parisi-Zhang equation (KPZ) [11]. In this limit, it is known [12] that the gradient of $L(x,t)$ become decorrelated along $x$ as $t\to\infty$. The $\nu_{ij}$'s and $\mu_{ij}$'s could still have finite-ranged correlations along the $x$ direction at the *discrete* level of the model. However this does not happen here. This can be seen from a Markov chain approach that we present in the Appendix. The consequence of this decorrelation property is that we can use Eqs. (2) in a self-consistent way in order to compute the probabilities $p_{ij}=P(\nu_{ij}=1)$ and $p'_{ij}=P(\mu_{ij}=1)$ for $i,j$ large. In this sense we may view the Bernoulli matching model as a mean-field model in which Eqs. (2) are ''cavity equations'' [10]. Assuming independence of $\nu_{i-1,j},\mu_{i,j-1}$ and $\epsilon_{ij}$ in Eq. (2) we get

$$p_{ij}=1-p'_{i-1,j}-(1-1/S)(1-p_{i,j-1})(1-p'_{i-1,j}),$$

$$p'_{ij}=1-p_{i,j-1}-(1-1/S)(1-p_{i,j-1})(1-p'_{i-1,j}). \tag{3}$$

These equations can be solved in a continuum limit [10], leading to

$$p(r)=\frac{\sqrt{rS}-1}{S-1},\quad p'(r)=\frac{\sqrt{S/r}-1}{S-1}, \tag{4}$$

where $p(r)=\lim_{i\to\infty}p_{i,ri}$ and $p'(r)=\lim_{i\to\infty}p'_{i,ri}$, and

$$\gamma_S^B(r)=\lim_{i\to\infty}\frac{L_{i,ri}}{i}=p(r)+rp'(r)=\frac{\sqrt{rS}-r-1}{S-1}. \tag{5}$$

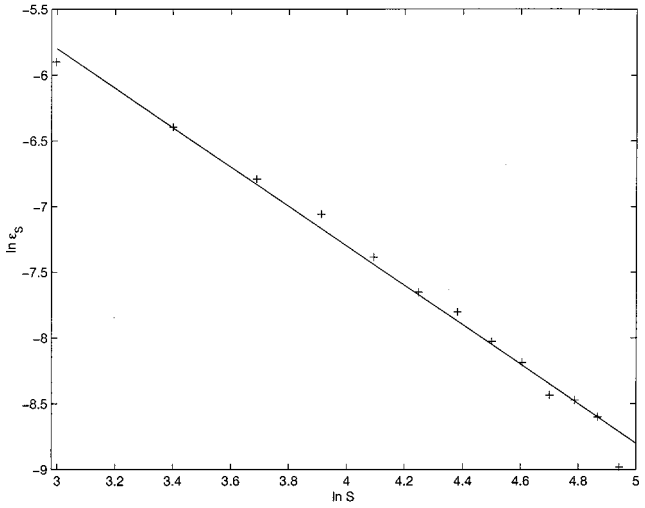


FIG. 1. Scaling of $\epsilon_S=\gamma_S^B-\gamma_S$ with $S$. Log-log plot for $20\leqslant S\leqslant 130$ (error bars not reproduced), together with a reference line of slope $-3/2$.

Note that Eqs. (4) and (5) are only valid for $1/S\leqslant r\leqslant S$. If $r>S$ $(r<1/S)$, the process evolves toward the state $(p,p')=(1,0)$ $[(p,p')=(0,1)]$ (this is a ''percolation transition'' of the LCS problem [10]).

## III. BERNOULLI MATCHING MODEL VERSUS RANDOM STRING MODEL

Let us briefly compare Eq. (5) to the numerical estimates obtained for the random string model. For simplicity we shall restrict ourselves to the case $r=1$ (random strings of equal sizes). Using Monte Carlo simulations and a finite-size scaling analysis, it was found [10] that the relative error $(\gamma_S^B-\gamma_S)/\gamma_S$ [with $\gamma_S^B=\gamma_S^B(r=1)=2/(1+\sqrt{S})$] is about $+2\%$ for $S=2$ and $S=3$, and decreases for $4\leqslant S\leqslant 15$ (it is about $+0.9\%$ for $S=15$). Figure 1 reproduces the behavior of the difference $\epsilon_S=\gamma_S^B-\gamma_S$ in a log-log plot for $S$ up to 130. Numerically, $\epsilon_S=\gamma_S^B-\gamma_S$ decreases rather fast at large $S$, showing a $1/S^\alpha$ dependence for a value of $\alpha$ compatible with 3/2. We remark that a simple expansion holds for the Bernoulli matching model, as we have $S\gamma_S^B/(2\sqrt{S}-2)=1/(1-1/S)=1+1/S+1/S^2+\cdots$. Anticipating on a similar expansion for the random string model, we would expect corrections in the left-hand side of this relation to occur in the $1/S$ term.

## IV. INCORPORATION OF CORRELATIONS

We now come to the question of computing corrections to the above approximation, by incorporating some of the correlations of the random string model in our calculation. This can be done in a systematic way as follows. We iterate relations (2) a certain number, say $k$ of times. The resulting equations are averaged, taking into account correlations among the $\epsilon_{ij}$'s, to build up the transition probabilities of a Markov process that we use as a refined approximation to the LCS problem. This approach is similar to the $n$-tree approximations that were used by Cook and Derrida to obtain a $1/d$ expansion for the directed polymer problem on finite-

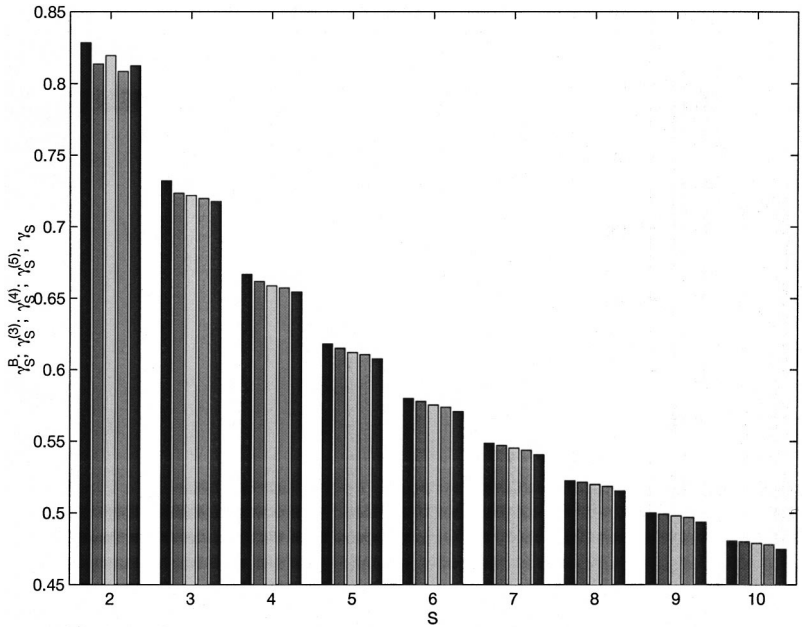


FIG. 2. Perturbative approximations to $\gamma_S$. This is a bar graph: For each $2 \leqslant S \leqslant 10$, the first to fifth bars from left to right give, respectively, the values of $\gamma_S^B = \gamma_S^{(1)}$, $\gamma_S^{(3)}$, $\gamma_S^{(4)}$, $\gamma_S^{(5)}$, and our numerical estimate of $\gamma_S$.

dimensional lattices [13]. We note however that the Bernoulli matching model is very different from a model of directed polymers on a hierarchical lattice, and the word ''tree'' would be somewhat misleading here. In order to analyze the above new process, we use a perturbative approach, *assuming* that the variables $\nu_{ij}$ and $\mu_{ij}$ for $i+j=t$ are independent in the stationary distribution, as they are in the Bernoulli matching model. It is then straightforward to obtain a self-consistent equation for $p = \lim_{i\to\infty} p_{ii}$ in the form $f_S^{(k)}(p) = p$, where $f_S^{(k)}(p)$ is an $S$-dependent polynomial of degree $2k$ in $p$. The positive solution $p_S^{(k)}$ to this equation provides us with a new approximation $\gamma_S^{(k)} = 2p_S^{(k)}$ to $\gamma_S$. Since there are no three-term correlations in the random string model (correlations among the $\epsilon_{ij}$'s occur only for configurations forming loops on the square lattice, e.g., in the four corners of a rectangle), it follows that no correlation in the disorder occur at level $k=2$, so $\gamma_S^{(k)}$ differ from $\gamma_S^B$ only for $k \geqslant 3$. An explicit computation shows that the equation $f_S^{(k)}(p) = p$ has only one positive root, at least up to $k=5$. The corresponding values of $\gamma_S^{(k)}$ thus provide sensible perturbative approximations to $\gamma_S$, which are reproduced in Fig. 2. Note that the estimates are improving, at least up to $k=5$ for $S \geqslant 3$. The successive values of $\gamma_2^{(k)}$ are not incompatible with a nonmonotonous approach to $\gamma_2$. The relative error $(\gamma_S^{(k)} - \gamma_S)/\gamma_S$ at $k=5$ is of $-0.48\%$ for $S=2$ and $+0.28\%$ for $S=3$, a significant improvement compared to the error committed with the Bernoulli matching estimate $\gamma_S^B$. This approximation scheme would be perfectly consistent if a decorrelation property occurred at every levels $k$. This is in fact *not* the case, for example, one can show that in the invariant distribution of the process at level $k=3$, the variables $\nu_{ij}$ and $\mu_{ij}$ are necessarily correlated. In the KPZ picture we may say that for $k \geqslant 3$, there remains as $t \to \infty$ short-ranged correlations along the $x$ direction in the local gradients of the growing interface's height. However, these correlations turn out to be numerically very small, which explains why our perturbative approach gives already a pretty accurate result at $S=2,3$. Moreover, when $S$ becomes large this approach becomes more and more accurate, as the exact invariant distribution resembles more and more that of the Bernoulli matching model, and we expect that the leading corrections introduced *at a given level k* are captured by this approximation.

We now evaluate the behavior of $\gamma_S^{(k)}$ as $S \to \infty$. This evaluation involves comparing $f_S^{(k)}(p)$ to the analogous polynomial $f_S^{B(k)}(p)$ computed within the Bernoulli matching model. The coefficients of $\delta f_S^{(k)} = f_S^{B(k)} - f_S^{(k)}$ are directly related to correlations among the $\epsilon_{ij}$'s. For example, the computation of $\delta f_S^{(3)}$ involves the four-correlation term $\langle \overline{\epsilon}_{i_1 j_1} \overline{\epsilon}_{i_1 j_2} \overline{\epsilon}_{i_2 j_1} \overline{\epsilon}_{i_2 j_2} \rangle = (1 + 1/(S-1)^3)(1 - 1/S)^4$, and we have

$$\delta f_S^{(3)}(p) = \frac{1}{(S-1)^3}\left(1 - \frac{1}{S}\right)^4 (1-p)^2 [1 - f_S^{(1)}(p)]^2, \quad (6)$$

with $f_S^{(1)}(p) = f_S^{B(1)}(p) = 1 - p - (1 - 1/S)(1-p)^2$. The coefficients of $\delta f_S^{(k)}$ all turn out to be of order $O(1/S^3)$ or smaller. For completeness we also give the expression of the polynomial $f_S^{B(3)}(p)$, which reads

$$\begin{aligned} f_S^{B(3)}(p) = 1 - f_S^{(2)}(p) &- \left(1 - \frac{1}{S}\right)\frac{1}{S}(1-p)^2 \\ &- 2\left(1 - \frac{1}{S}\right)^2 p(1-p)[1 - f_S^{(1)}(p)] \\ &- \left(1 - \frac{1}{S}\right)^3 [1 - f_S^{(1)}(p)]^2 \left[p^2 + \left(1 - \frac{1}{S}\right)(1-p)^2\right], \end{aligned} \quad (7)$$

where $f_S^{(2)}(p) = f_S^{B(2)}(p) = f_S^{(1)} \circ f_S^{(1)}(p)$. If we now let $\delta p_S^{(k)} = p_S^{B(k)} - p_S^{(k)}$ where $p_S^{B(k)}$ is the positive solution to $f_S^{B(k)}(p) = p$, i.e., $p_S^{B(k)} = p_S^B = 1/(1 + \sqrt{S})$, a standard computation leads to

$$\delta p_S^{(k)} = \frac{\delta f_S^{(k)}(p_S^B)}{1 - \frac{d}{dp} f_S^{B(k)}(p_S^B)} \tag{8}$$

up to negligible terms at large $S$. It can be checked that $(d/dp) f_S^{B(k)}(p_S^B) = 1 - 2k/\sqrt{S} + O(1/S)$. It follows then from Eq. (8) that for fixed $k$, the correction $\delta p_S^{(k)}$ is of order $O(1/S^{5/2})$, which cannot account for the observed $1/S^{3/2}$ behavior of $\epsilon_S = \gamma_S^B - \gamma_S$. The computation gives $\delta p_S^{(3)} \sim 1/6S^{5/2}$, $\delta p_S^{(4)} \sim 1/2S^{5/2}$, and $\delta p_S^{(5)} \sim 1/S^{5/2}$, together with correcting terms in the form of series in powers of $1/\sqrt{S}$. We could not extract the general terms of these series for arbitrary $k$, but we strongly suspect that (at least) the coefficient $A_k$ in front of $1/S^{5/2}$ *diverges* at large $k$. The argument goes roughly as follows. The correlation terms involved in the computation of $\delta f_S^{(k)}(p)$ can all be put into the form

$$\langle \overline{\epsilon}_{i_1 j_1} \cdots \overline{\epsilon}_{i_l j_l} \rangle = \langle (1 - \epsilon_{i_1 j_1}) \cdots (1 - \epsilon_{i_l j_l}) \rangle. \tag{9}$$

Expanding the product and taking averages, it is not difficult to see that any such term behaves as

$$\begin{aligned} &\langle (1 - \epsilon_{i_1 j_1}) \cdots (1 - \epsilon_{i_l j_l}) \rangle \\ &\quad = 1 - \binom{l}{1} \frac{1}{S} + \binom{l}{2} \frac{1}{S^2} - \binom{l}{3} \frac{1}{S^3} + \frac{n_{i_1 j_1 \cdots i_l j_l}}{S^3} + O\left(\frac{1}{S^4}\right) \\ &\quad = \left(1 - \frac{1}{S}\right)^l \left[ 1 + \frac{n_{i_1 j_1 \cdots i_l j_l}}{S^3} + O\left(\frac{1}{S^4}\right) \right], \end{aligned} \tag{10}$$

where $n_{i_1 j_1 \cdots i_l j_l}$ is the number of rectangles that can be formed with four corners on the graph made up by the lattice points $(i_1 j_1), \ldots, (i_l j_l)$. At level $k$ we have to consider rectangles formed on the triangular lattice $\Delta_k$ made up by the points $(i,j)$ such that $0 \leqslant i,j \leqslant k$ and $i + j \geqslant k$. The number $n_k$ of these rectangles satisfies the recursion relation $n_k = 2n_{k-1} - n_{k-2} + k(k+1)/2$, which, in the large-$k$ limit, gives $(d^2/dk^2) n_k \sim k^2/2$, leading to $n_k \sim \frac{1}{24} k^4$ (from a more precise computation, taking account of $n_0 = 0$ and $n_1 = 1$, one finds that $n_k = \frac{1}{24} k^4 + \frac{1}{4} k^3 + \frac{11}{24} k^2 + \frac{1}{4}$). All these rectangles are involved in $A_k$, as, for example, the polynomial $\delta f_S^{(k)}(p)$ always contains a term of the form

$$\left[ \left(1 - \frac{1}{S}\right)^{k(k+1)/2} - \left\langle \prod_{(ij) \in \Delta_k} \overline{\epsilon}_{ij} \right\rangle \right] (1-p)^{2k}, \tag{11}$$

which gives a contribution $n_k/S^3 + O(1/S^{7/2})$ to $\delta f_S^{(k)}(p_S)$. Unless some special cancelation occurs between the different correlation terms, we thus expect that the behavior of $A_k$ will be approximately given by $A_k \propto n_k/2k$, i.e., we find that it diverges like $k^3$ at large $k$. We conclude that the leading corrections to the Bernoulli matching model in a large $S$ expansion depend on long-ranged correlations among the matches in the random string model. In order to capture them we should look at arbitrarily large values of $k$, for which the above perturbative approach is no longer valid.

## V. CONCLUSION

The main point of this paper is that, while the Bernoulli matching model provides a natural and accurate mean-field-like approximation to the LCS problem valid in the limit of a large alphabet, the corresponding large-$S$ expansion is nonperturbative: Inclusion of finite-ranged correlations leads to a series with diverging coefficients, while the overall behavior of the expansion at large $S$ does not reproduce the observed gap between the two models. This contrasts with the results of [13], where the $n$-tree approximation led to a consistent $1/d$ expansion for the directed polymer problem. As already pointed out, we are dealing here with a different kind of mean-field approximation. The Bernoulli matching model is not an infinite-dimensional model, and replica symmetry is not broken in this model [10]. Note that the $1/d$ expansion for the directed polymer problem is also known to be singular, but in a more subtle way: Replica symmetry is restored at finite dimensions, leading to important ''tunneling'' effects between the energy valleys of the mean-field picture [14]. An interesting feature of the LCS problem is that the corrections induced by finite-ranged correlations, while singular, remain within a perturbative series in powers of $1/\sqrt{S}$. In this respect, the situation for the LCS problem seems more favorable than in other combinatorial problems where the correlations in the disorder induce nonperturbative corrections in an expansion around the mean-field approximation [15]. This makes the LCS problem an interesting model for investigating this kind of singularity.

## ACKNOWLEDGMENTS

The main part of this work was done at the BiBoS center of Bielefeld. J.BdM. is grateful to Professor P. Blanchard for his kindness and for several discussions that were very stimulating in this research, and he thanks O.C. Martin for many useful comments and suggestions on the manuscript. He also thanks R. Bundschuh for pointing him to Ref. [12]. This work was supported by the EU-TMR project ''Stochastic Analysis and its Applications.''

## APPENDIX: MARKOV CHAIN APPROACH TO THE BERNOULLI MATCHING MODEL

Let us denote by $(\nu\mu)^t = \{\nu_{ij}, \mu_{ij}, i + j = t\}$ the *state* of the process defined by Eq. (2), with $t$ interpreted as time. In the Bernoulli matching model the evolution is Markovian, i.e., the transition from a given state at time $t$ to another state at time $t+1$ is not affected by the states at times $t' < t$ (this is not the case in the random string model, where the transition from time $t$ to time $t+1$ is affected by the whole history of the process). We will first show that, as a Markov process, the Bernoulli matching model admits invariant distributions in which the components of $(\nu\mu)^t$ are completely decorrelated (we mention that the same result has been found in a different way in [16], in the case $N = M$). Consider the relations (2) restricted to a given cell of the lattice $\mathcal{C}_{NM}$. The corresponding ''one-cell'' transition probability $P_1(\nu, \mu | \nu', \mu')$ is given by

$$P_1(\nu,\mu|\nu',\mu') = \frac{\nu\mu\overline{\nu'}\,\overline{\mu'}}{S} + \overline{\nu}\mu\overline{\nu'}\mu' + \nu\overline{\mu}\nu'\overline{\mu'} + \overline{\nu\mu} \times [\nu'\mu' + (1-1/S)\overline{\nu'}\,\overline{\mu'}]. \quad \text{(A1)}$$

A simple computation shows that the one-cell Perron-Frobenius equation $P_1\pi_1=\pi_1$ (with matrix notations) has a solution of the form $\pi_1(\nu,\mu)=[p\nu+(1-p)\overline{\nu}][p'\mu+(1-p')\overline{\mu}]$ provided the probabilities $p$ and $p'$ satisfy

$$1=p+p'+(S-1)pp'. \quad \text{(A2)}$$

Suppose now that we let the bonds on the lower corner of any given rectangle be occupied independently with probability $p$ for horizontal bonds and $p'$ for vertical bonds. A moment's thought shows that the *same* distribution will hold for the upper corner bonds if we let the occupation numbers for the bonds inside the rectangle evolve according to Eq. (2), as long as $p$ and $p'$ satisfy Eq. (A2). Hence any solution of Eq. (A2) provides us with a decorrelated invariant distribution as was claimed. In a continuum limit, these invariant distributions can be identified locally with the ''pure'' invariant distributions of the process, i.e., those invariant distributions evolved from a single initial state of the variables $\nu,\mu$. More precisely let us impose periodic boundary conditions along the $x=i-j$ direction (this is a way of working ''locally''). We let $\nu^t=(\nu_1^t,\ldots,\nu_L^t)$ and $\mu^t=(\mu_1^t,\ldots,\mu_L^t)$ be the state variables at time $t$ on a band of ''width'' $L$, and we adopt a numerotation such that Eq. (2) reads

$$\nu_i^{t+1}=(1-\overline{\epsilon}_i^{t+1}\overline{\nu}_{i+1}^t)\overline{\mu}_i^t,$$

$$\mu_i^{t+1}=(1-\overline{\epsilon}_i^{t+1}\overline{\mu}_i^t)\overline{\nu}_{i+1}^t, \quad \text{(A3)}$$

for $i=1,\ldots,L$ ($L+1$ being identified with 1). From the remarks made above, the Perron-Frobenius equation $P_L\pi_L=\pi_L$ of this process admits solutions of the form $\pi_L^{(p,p')}(\nu_1,\mu_1\ldots,\nu_L,\mu_L)=\Pi_{i=1}^L\pi_1(\nu_i,\mu_i)$, where again $(p,p')$ is any solution of Eq. (A2). For finite $L$, however, $\pi_L^{(p,p')}$ is not pure. To get an understanding of the pure distributions, we adopt a lattice gas point of view, remarking that the quantity

$$C=\sum_{i=1}^{L}\nu_i-\mu_i \quad \text{(A4)}$$

is a conserved charge of the evolution (this conservation law is exact only under the above periodic boundary conditions). It can also be checked that the Markov process defined by Eq. (A3) connects any two states $(\nu_i\mu_i),(\nu_i'\mu_i')$ having the same charge $C$. It follows that there are exactly $2L+1$ pure distributions, in correspondence with the possible values $-L\leqslant C\leqslant L$. We can extract a formal expression for the pure invariant distribution $\pi_C(\nu_i,\mu_i)$ evolved from an arbitrary state of charge $C$, from the ''mixed'' invariant distribution $\pi_L^{(p,p')}(\nu_i,\mu_i)$. Namely, we have

$$\pi_C(\nu_i,\mu_i)=\frac{\pi_L^{(p,p')}(\nu_i,\mu_i)\,\delta\left(C-\sum_i\nu_i+\sum_i\mu_i\right)}{Z(C)} \quad \text{(A5)}$$

where $Z(C)$ is a normalization factor. Note that, contrary to the appearances, the right-hand side of Eq. (A5) does *not* depend on $(p,p')$ [this can be seen directly from the expression of $\pi_L^{(p,p')}$ by making use of Eq. (A2)]. In the limit $L\to\infty$, the fluctuations of $C$ about its mean value $L(p'-p)$ with respect to $\pi_L^{(p,p')}$ become negligible, and we expect that the differences between the pure distributions $\pi_C$ (more precisely the differences between their finite correlation functions) for which $C/L$ is close to $p'-p$ will become insignificant. This can be checked directly from Eq. (A5) using a saddle-point evaluation. Hence the pure distributions on a periodic band of infinite width can be identified with a continuum of decorrelated distributions $\pi^{(p,p')}$, parametrized by the solutions of Eq. (A2) for which $0\leqslant p,p'\leqslant 1$. Returning to the original lattice $\mathcal{C}_{NM}$, one must take care of the boundary conditions imposed along the axes $i=0$ and $j=0$. The process will now develop only locally (i.e., along any given direction) according to a distribution of the form $\pi^{(p,p')}$, with $p$ and $p'$ being functions of $r=i/j$. The cavity approach of Sec. II allows us to treat different boundary conditions in a simple way. For example, if the horizontal bonds along $j=0$ (the vertical bonds along $i=0$) are supposed to be occupied independently with probability $p_1$ ($p_2$), where $0\leqslant p_1,\ p_2\leqslant 1/(1+\sqrt{S})$ (the original problem corresponds to the case $p_1=p_2=0$), one finds that

$$p(r)=p_1, \quad 0\leqslant r\leqslant r_1,$$

$$p(r)=\frac{\sqrt{rS}-1}{S-1}, \quad r_1\leqslant r\leqslant r_2,$$

$$p(r)=\frac{1-p_2}{1+(S-1)p_2}, \quad r\geqslant r_2, \quad \text{(A6)}$$

where $r_1$ and $r_2$ are such that $p_1=(\sqrt{r_1S}-1)/(S-1)$, $p_2=(\sqrt{S/r_2}-1)/(S-1)$, and $p'(r)$ is such that Eq. (A2) is satisfied.


[1] D. Sankoff and J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison Wesley, Reading, MA, 1983).

[2] *Mathematical Methods in DNA Sequences*, edited by M. Waterman (CRC Press, Boca Raton, FL, 1989).

[3] M. Steele, *Probability Theory and Combinatorial Optimisation* (SIAM, Philadelphia, 1997).

[4] V. Chvatal and D. Sankoff, J. Appl. Probab. **12**, 306 (1975).

[5] V. Dancik and M. Paterson, in *STACS94. Lecture Notes in Computer Science 775* (Springer, New York, 1994), pp. 306–315.

[6] K. Alexander, Ann. Appl. Prob. **4-4**, 1074 (1994).

[7] T. Hwa and M. Lässig, Phys. Rev. Lett. **76**, 2591 (1996).
[8] S. Needleman and C. Wunsch, J. Mol. Biol. **48**, 443 (1970).
[9] E. Ukkonen, Inf. Control. **64**, 100 (1985).
[10] J. Boutet de Monvel, Eur. Phys. J. B **7**, 293 (1999).
[11] M. Kardar, G. Parisi, and Y. Zhang, Phys. Rev. Lett. **56**, 889 (1986).
[12] J. Krug and H. Spohn, in *Solids Far From Equilibrium: Growth, Morphology, and Defects*, edited by C. Godrèche (Cambridge University Press, Cambridge, England, 1991), pp. 412–525.
[13] J. Cook and B. Derrida, J. Phys. A **23**, 1523 (1990).
[14] G. Parisi and F. Slanina, Eur. Phys. J. B **8**, 603 (1999).
[15] J. Houdayer, J. Boutet de Monvel, and O. Martin, Eur. Phys. J. B **6**, 383 (1999).
[16] R. Bundschuh and T. Hwa, in *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, edited by S. Istrail, P. Pevzner, and M. Waterman (ACM Press, New York, 1999), pp. 70–76.