Possible sets of autocorrelations and the Simplex algorithm

2006 J. Phys. A: Math. Gen. 39 4161

(http://iopscience.iop.org/0305-4470/39/16/004)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **39** (2006) 4161–4171 doi:10.1088/0305-4470/39/16/004

Possible sets of autocorrelations and the Simplex algorithm

Shahar Keren, Haggai Kfir and Ido Kanter

Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan, 52900, Israel

E-mail: shahar@gteko.com, hkfir@iai.co.il and kanter@mail.biu.ac.il

Received 30 October 2005, in final form 7 March 2006
Published 31 March 2006
Online at stacks.iop.org/JPhysA/39/4161

Abstract

The problem of imposing a set of correlations, $\mathcal{C}$, of any order, $\mathcal{C} = \{C_1, C_{23}, C_{145}, \ldots\}$, on binary sequences is addressed. The entropy of infinitely long sequences obeying such a given set was calculated in previous works using the saddle-point method, and it was observed that a finite fraction of sets are characterized by a non-extensive entropy. In this paper, the region of finite entropy, the *allowed* region of sets of correlations, is found to be a convex hyper-polygon in the space of correlation-sets, using the Simplex algorithm. Outside of this region the Simplex solution indicates that sequences obeying the correlations cannot be found; therefore, the entropy is $-\infty$. In particular, the boundaries of the allowed region for $\{C_1, C_m\}$ are presented. At the boundaries, the entropy drops in a first-order phase transition fashion, and this drop can be explained from a combinatorial point of view. Finally, we observe that the fraction of the volume occupied by allowed correlation-sets drops exponentially with the number of correlations imposed, and a qualitative explanation of this scaling phenomenon is provided.

PACS numbers: 65.40.Gr, 05.20.−y, 87.10.+e

1. Introduction: correlation-sets

Autocorrelation, a cross-correlation of a sequence with itself, and correlated sequences are of great interest in a variety of fields, for example economics [1], biology [2], physiology [3] and digital communication [4]. Efforts usually focus on identifying the correlations typical of a process, such as correlations in the quotes of a stock in the stock market [1], or autocorrelated noise over a communication channel [4]. Alternatively, one may attempt to relate a certain correlation type to a behaviour of a system, such as long-range correlations in heartbeats of an ill individual [3], or autocorrelations within coding or non-coding DNA sections [2]. In this paper, we address the subject from a different perspective: we analyse the consequence

of imposing a *set* of constraints, in the form of autocorrelations, on a binary sequence. We calculate the entropy of the *ensemble* of all sequences obeying a given set of correlations. We show that choosing a specific set of correlations is not arbitrarily free, but the space of *correlation-sets* can actually be divided into *allowed* and *restricted* regions. In the *allowed* region, the entropy of sequences obeying the correlation-set is an extensive quantity, while in the surrounding *restricted* regions the entropy is $-\infty$; therefore, no sequence obeying the constraints can be found. We map the boundaries of the allowed region, and show that it is a *convex* region, forming a hyper-polygon in the correlation-set space.

Autocorrelations can appear in the simple form of two-point correlations [5]

$$C_m = \frac{1}{L}\sum_{i=1}^{L} x_i x_{(i+m) \bmod L}, \tag{1}$$

where L is the length of the binary vector, $x_i \in \pm 1$, m is the correlation distance and we assume periodic boundary conditions. In the general form, high-order correlations are given by

$$C_{m_1,m_2,\ldots,m} = \frac{1}{L}\sum_{i=1}^{L} x_i \prod_{j=m_1,m_2,\ldots,m} x_{(i+j) \bmod L} \tag{2}$$

which is a measure of the correlation of an element with successive elements located at distances $m_1, m_2, \ldots, m$ apart from it, and m being the maximal correlation distance taken $(m_1, m_2, \ldots \leqslant m)$. For a given m, the total number of possible different autocorrelations is 2^m. For $m = 2$, for instance, there are only four possible correlations, C_0, C_1, C_2 and C_{12}, and for $m = 3$ there are eight possible different correlations, $C_0, C_1, C_2, C_3, C_{12}, C_{13}, C_{23}, C_{123}$ (C_0 is simply the bias of the sequence: $C_0 = \frac{1}{L}\sum_1^L x_i$).

Let a *correlation-set*, $\mathcal{C}$, be a collection of some high-order correlations. Think of each element of $\mathcal{C}$ as a base vector in a vector space, the *correlation-set space*. The dimensionality d of the correlation-set space is determined by the cardinality of $\mathcal{C}$, and every specific set (e.g., $\{C_1 = -0.3, C_{14} = 0.2, \ldots\}$) is a point in this space. Since by definition $-1 \leqslant C_{m_1,m_2,\ldots,m} \leqslant 1$, the *volume* of this space is simply $V = 2^d$.

We raise the following questions regarding the correlation-set space: is there a way to estimate the number of sequences obeying the constraints of a given point in this space? And are there correlation-sets representing contradicting constraints that cannot co-exist? In other words, what is the entropy of a point in a correlation-set space, and do correlation-sets with entropy equal to $-\infty$ exist[1].

An answer to the first question is given in section 2. Using the statistical-physics technique of the transfer matrix, the entropy of a point in a given correlation-set space is calculated. However, we observe that this method does not provide a solution for every point in the space, corresponding to points with zero entropy, or may be a consequence of numerical precision problems. In sections 3 and 4, the problem is reformulated in terms of constraining marginal probabilities on short blocks, and the Simplex algorithm [6] is applied for solving this problem. The Simplex method answers the second question in the affirmative. All points of non-zero entropy are located within an *allowed* region, which is a polygon-like convex region in the correlation-set space. Outside this region, no sequences obeying the constraints exist, and the entropy is $-\infty$. Section 5 provides detailed results for the cases

[1] We remark that two main differences distinct the correlation-set problem from the well-known K-SAT problem [15, 16]: (1) K-SAT is mapped onto highly diluted *infinite*-ranged frustrated systems, while our system is a one-dimensional system with *finite*-range interactions. (2) In our system there is a finite number of *global* constraints, while in K-SAT, there is an extensive number of *local* constrains.

$\mathcal{C} = \{C_1, C_m\}$ and $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, indicating that the volume of the allowed region exponentially decays with the number of constraints imposed, and we support it by a qualitative explanation. Concluding remarks are given in section 6, followed by the appendices, describing a combinatorial method for constructing the boundaries of the allowed region and a proof of convexity.

2. Calculating the entropy of correlated sequences

In this section, the method of finding the entropy (per bit) of the ensemble of sequences obeying a set of autocorrelations is briefly described. We repeat our calculations from recent papers [7] and references therein. For the sake of clarity of notation, we focus on two-point correlations only, considering the set of all such correlations up to the order m : $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ (in practice we can only deal with $m < 10$ due to computational limitations). Extending to high-order correlations is straightforward, and is presented in [8]. Let $\Omega(\mathcal{C})$ denote the number of sequences of length L obeying the set $\mathcal{C}$. Assuming periodic boundary conditions, and representing each two-point correlation constraint as a delta function, $\delta\left(\sum_{i=1}^{L} x_i x_{i+k} - LC_k\right)$, $(1 \leqslant k \leqslant m)$, one obtains

$$\Omega(\mathcal{C}) = \mathrm{Tr}_{\{x_i=\pm 1\}} \prod_{k=1}^{m} \delta\left(\sum_{i=1}^{L} x_i x_{i+k} - LC_k\right). \tag{3}$$

Using the integral representation of the delta functions, equation (3) can be written as

$$\Omega(\mathcal{C}) = \mathrm{Tr}_{\{x_i=\pm 1\}} \prod_{k=1}^{m} \int_{-i\infty}^{i\infty} \exp\left\{ y_k \left(\sum_{i=1}^{L} x_i x_{i+k} - LC_k\right)\right\} \mathrm{d}y_k. \tag{4}$$

Since the Trace is over x_i and the integrations are over y_k, equation (4) can be rearranged to

$$\Omega(\mathcal{C}) = \int \ldots \int \mathrm{e}^{-L\sum_{k=1}^{m} C_k \cdot y_k} \times \mathrm{Tr}_{\{x_i=\pm 1\}} \prod_{i=1}^{L} \exp\left(\sum_{k=1}^{m} y_k \cdot x_i x_{i+k}\right) \mathrm{d}y_1 \ldots \mathrm{d}y_m. \tag{5}$$

The term inside the Trace represents the interactions (according to $\mathcal{C}$) of each element, x_i, in the sequence. Since the maximal correlation length is bounded, $k \leqslant m$, one can group the sequence into L/m blocks of size m, and apply the *transfer matrix* method [9] (of dimensions $2^m \times 2^m$) for representing all the interactions among the $2m$ elements in two successive blocks. In the leading order, one finds

$$\Omega(\mathcal{C}) \approx \int \ldots \int \exp\left(-L\left[\sum_{k=1}^{m} C_k y_k - \frac{1}{m} \ln \lambda_{\max}(y_1 \ldots y_m)\right]\right) \mathrm{d}y_1 \ldots \mathrm{d}y_m, \tag{6}$$

where $\lambda_{\max}(y_1 \ldots y_m)$ is the maximal eigenvalue of the corresponding transfer matrix. For large L, Ω can be found using the saddle-point method. Denoting by $y_1^*, \ldots, y_m^*$ the solutions of the saddle point (i.e., y's of the extremum, minimizing the exponent term in (6)), the binary entropy[2], $H_2(\mathcal{C})$, is given in the leading order by[3]:

$$H_2(\mathcal{C}) = \frac{1}{\ln 2}\left[\frac{1}{m} \ln \lambda_{\max}(y_1^* \ldots y_m^*) - \sum_{k=1}^{m} y_k^* C_k\right]. \tag{7}$$

[2] We choose to report the binary entropy, $H_2 = \log_2(\Omega(\mathcal{C}))$, rather than the natural logarithm in order to comply with the information theory literature notation.

[3] The numerical solution of the saddle-point equation was preformed using Powell's algorithm for minimization in high dimensions, available from numerical recipes at http://www.library.cornell.edu/nr/bookcpdf/c10-5.pdf.

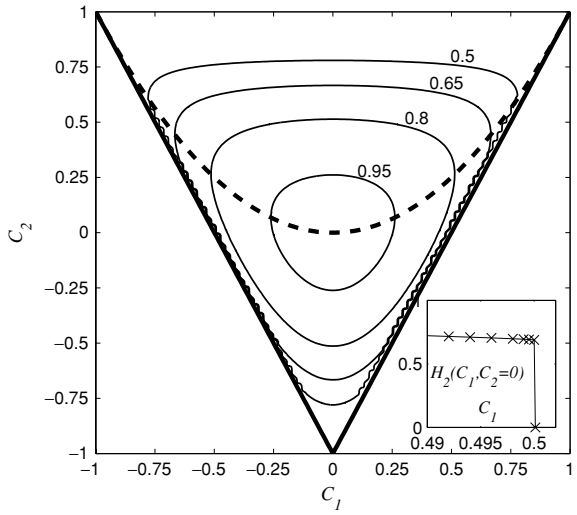

Figure 1. The correlation-set space for $\mathcal{C} = \{C_1, C_2\}$. The allowed region is $C_2 \geqslant 2|C_1| - 1$ (thick line). Iso-entropy lines (thin line), and the $C_2 = C_1^2$ line (dashed line) are presented. Inset: the entropy as a function of C_1 for $C_2 = 0$. This cross-section demonstrates the sharp drop in the entropy at the point $C_1 = 0.5$.

Table 1. Entropy, H_2, and saddle point solutions, y^*, near the boundary of the allowed region for $\mathcal{C} = \{C_1, C_2\}$. Note the divergence of $|y_1|$ and $|y_2|$ at the boundary.

c_1	c_2	y_1	y_2	H_2
0.6	0.22	1.705	−0.633	0.675
0.58	0.2	1.376	−0.479	0.703
0.6	0.2	8.801	−4.198	0.64
0.62	0.2	–	–	No solution
0.6	0.18	–	–	No solution

By surveying, for instance, the correlation-set space $\mathcal{C} = \{C_1, C_2\}$, it was observed [7] (and references therein) that a solution for the saddle point, equation (7), can be found only in the region

$$C_2 \geqslant 2|C_1| - 1. \tag{8}$$

At the boundary of this region, $C_2 = 2|C_1| - 1$, the parameters $\{|y^*|\}$ diverge, and the entropy drops to zero in a first-order phase transition fashion. In table 1 results around the point $C_1 = 0.6, C_2 = 0.2$ are reported. In figure 1, the correlation space for $\mathcal{C} = \{C_1, C_2\}$ is presented, together with iso-entropy lines, and the line $C_2 = C_1^2$, representing the most probable C_2 for a given C_1.

3. From ensembles of sequences to marginal probabilities of short blocks

The limited results presented in the previous section, obtained from the numerical solutions of the saddle point equations, suffer from the following limitations: (a) finding the boundaries of the allowed region for a set $\mathcal{C} = \{C_{m_1,m_2,\ldots,m}\}$ is very sensitive to the numerical precision, since on the boundary the interactions diverge; (b) the extension of the saddle-point method to many dimensions (i.e., increasing the number of correlations in the set), or even to large

m with only two-point correlations, is a very heavy numerical task because the saddle-point method requires repeatedly finding the maximal eigenvalue of a $2^m \times 2^m$ transfer matrix; (c) the question of whether out of the space with a finite entropy, there are a finite or infinite number of sequences (for instance $e^{\sqrt{L}}$) obeying the set of correlations cannot be answered using the saddle-point method; (d) it is unclear whether the available space consists of a connected region.

To overcome these difficulties in the following section, we show how the allowed region of a correlation-set can be mapped using the Simplex algorithm, but first we draw a conceptual connection between the ensemble of all the (finite or infinitely long) sequences obeying a given set of correlations $\mathcal{C}$, and marginal probabilities of short blocks.

We term a short section of the sequence that contains N binary elements, a *block* of length N. Let $P(\pm\cdots\pm)$ be the marginal probability of a certain internal representation of a block. For $N = 4$, for instance, $P(++-+)$ is the probability of finding the subsequence $+, +, -, +$ in the ensemble of all the binary sequences obeying $\mathcal{C}$. Let i be a running index over the 2^N configurations and P_i be the marginal probability of the configuration i. Since the correlations are defined with periodic boundary conditions, if a given sequence is included in the ensemble, then all the cyclic permutations of that sequence will also be included. P_i is therefore independent of the specific location of the block along the sequence. In other words, sampling blocks from different locations yield the same marginal distribution of P_i. Hence, instead of treating long sequences obeying a set $\mathcal{C}$, we can discuss short blocks and the marginal probabilities that induce the desired correlations. In the following section we show how this problem can be solved by the Simplex algorithm.

4. Determine the allowed region using the Simplex algorithm

The Simplex algorithm [6] is a method for solving problems in linear programming. Linear programming (LP) problems are optimization problems in which the objective function (the function to be minimized or maximized, also called the *target* function) and the constraints (equalities and inequalities) are all linear. This method, invented by G B Dantzig in 1947 [10], runs along polytope edges of the visualization solid to find the best answer. The algorithm's complexity is considered to be polynomial in the number of parameters or constraints for every practical use (although rare cases with exponential complexity were demonstrated [11]). Further information on the LP and the Simplex algorithm can be found in [12–14].

Using the marginal probabilities as variables, we can write linear equalities and inequalities as constraints and an additional linear target function. The target function is the correlation we want to maximize (or minimize). The method is first demonstrated for the simple case of $\mathcal{C} = \{C_1, C_2\}$.

4.1. Imposing $\mathcal{C} = \{C_1, C_2\}$ constraints

Following the methodology of the transfer matrix, having $m = 2$, let us concentrate on a block of $N = 4$ successive binary variables S_1, S_2, S_3, S_4, where $S_k = \pm 1$.

The number of possible configurations for this block is $2^N = 16$.

4.1.1. Constraining P_i to be legitimate probabilities. In order to ensure that the variables P_i have proper values as probabilities, they should obey

$$\forall i \qquad 0 \leqslant P_i \leqslant 1 \tag{9}$$

and

$$\sum_{i=1}^{16} P_i = 1. \tag{10}$$

However, the Simplex method deals only with non-negative variables therefore the condition in equation (10) immediately implies that $\forall i\, P_i \leqslant 1$, rendering equation (9) unnecessary.

4.1.2. Implying the correlation constraints. Following the discussion in section 3, we demand that the average over every two successive binary elements is equal to C_1. We therefore sum the probabilities of all configurations where $S_1 = S_2$ and subtract the probabilities of all configurations where $S_1 \neq S_2$.

Implying the constraint C_1 on the binary elements S_1 and S_2 yields the following equation[4],

$$\sum_{i=1}^{16} P_i f_{1,2}(i) = C_1, \tag{11}$$

where $f_{k,k+l}(i)$ is an indicator function; $f_{k,k+l}(i) = 1$ when i is a configuration with both elements (S_k and S_{k+l}) having the same sign, and $f_{k,k+l}(i) = -1$ when the elements have opposite signs.

Similarly, for elements S_2 and S_3, and for elements S_3 and S_4,

$$\sum_{i=1}^{16} P_i f_{2,3}(i) = C_1 \tag{12}$$

$$\sum_{i=1}^{16} P_i f_{3,4}(i) = C_1. \tag{13}$$

Our target parameter is C_2 (our goal is to find the min/max C_2 for a given C_1). By adding the probabilities of all configurations where $S_1 = S_3$ and subtracting the probabilities of all configurations where $S_1 \neq S_3$, we find for the elements S_1 and S_3

$$\sum_{i=1}^{16} P_i f_{1,3}(i) = C_2. \tag{14}$$

Similarly, for elements S_2 and S_4

$$\sum_{i=1}^{16} P_i f_{2,4}(i) = C_2. \tag{15}$$

Since the Simplex algorithm only treats one target function, we use one of these last two equations as a target function (e.g., equation (14)), and the other equation as an additional constraint, which demands that C_2 between elements S_1 and S_3 is equal to C_2 between elements S_2 and S_4. Solving these equations for a given C_1 with the Simplex method results in the maximal/minimal C_2, which complies with all the constraints.

4.2. Imposing $\{C_1, C_m\}$ constraints

Implementing the previous example to the more general case of the set $\{C_1, C_m\}$ is fairly straightforward. The equations are very similar to the former $\{C_1, C_2\}$ case. Again we use blocks with size $N = 2m$, but as m is larger, the length of the block increases. The number of equations for C_1 becomes $N - 1$, and the number of equations for C_m becomes $(N - m)$, out of which one becomes the target function and the rest are constraints.

[4] The explicit form of equation (11): $[P_{(--++)} + P_{(--+-)} + P_{(---+)} + P_{(----)} + P_{(++++)} + P_{(+++-)} + P_{(++-+)} + P_{(++--)}] - [P_{(-+++)} + P_{(-++-)} + P_{(-+-+)} + P_{(-+--)} + P_{(+-++)} + P_{(+-+-)} + P_{(+--+)} + P_{(+---)}] = C_1$.

4.3. The general case of two-point correlations: $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$

The most general case of two-point correlations is the case of the set $\{C_1, C_2, \ldots, C_m\}$. Assuming that the allowed region of $\{C_1, C_2, \ldots, C_{m-1}\}$ is already mapped, our next step is to find the max/min C_m for each of the allowed points in the $(m-1)$ space. Again, following the methodology of the transfer matrix, we take as block size $N = 2m$.

For every $l < m$, for all element-pairs l sites apart, we enforce a constraint C_l

$$\forall l < m \quad \forall k \leqslant N - l \qquad \sum_{i=1}^{2^N} P_i f_{k,k+l}(i) = C_l. \tag{16}$$

For each correlation C_l we have $(N - l)$ pairs, and therefore $(N - l)$ equations. Summing over all $l < m$, we find that the number of equations for correlation constraints is $[(N-1) + (N-(m-1))](m-1)/2$.

The C_m constraint is represented by $(N - m)$ equations

$$\forall k \leqslant N - m \qquad \sum_{i=1}^{2^N} P_i f_{k,k+m}(i) = C_m. \tag{17}$$

As before, one of the $(N - m)$ equations becomes the target function, and we set the rest as additional constraints. In total there are 2^N variables and the number of equations is

$$(2N - m)(m-1)/2 + N - m - 1 \tag{18}$$

plus one target function.

5. Results

5.1. A pair of two-point correlations, $\mathcal{C} = \{C_1, C_m\}$

The allowed region in the $\mathcal{C} = \{C_1, C_2\}$ space is a head-down isosceles triangle (equation (8), figure 1) defined by

$$C_2 \geqslant 2|C_1| - 1.$$

For $\mathcal{C} = \{C_1, C_3\}$, we found that the allowed region is a parallelogram (figure 2), formed by the lines

$$3C_1 - 2 \leqslant C_3 \leqslant 3C_1 + 2.$$

From the solutions up to $m \leqslant 7$ and from the argument given in appendix A, we discovered a simple generic behaviour for a pair of two correlations C_1 and C_m

$$\begin{aligned} &C_m \geqslant m|C_1| + (1-m), && \text{for even } m \\ &mC_1 - (m-1) \leqslant C_m \leqslant mC_1 + (m-1), && \text{for odd } m \end{aligned} \tag{19}$$

(bearing in mind that by definition $|C_l| \leqslant 1$ $\forall l$). In figure 2, the allowed regions of $\mathcal{C} = \{C_1, C_m\}$ for $m = 2, \ldots, 7$ are presented. An explanation for this generic behaviour from combinatorial considerations is given in appendix A. Note that for the either case, the *area* of the allowed region covers a fraction $\left(1 - \frac{1}{m}\right)$ of the entire 2×2 space.

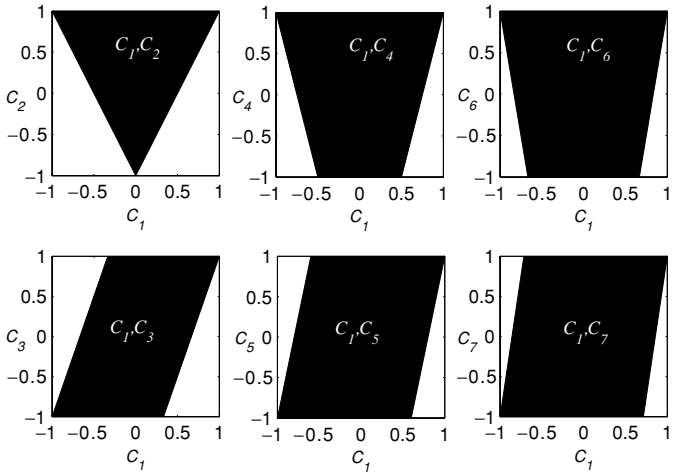

Figure 2. Allowed region in sets of $\mathcal{C} = \{C_1, C_m\}$, for $m = 2, \ldots, 7$.

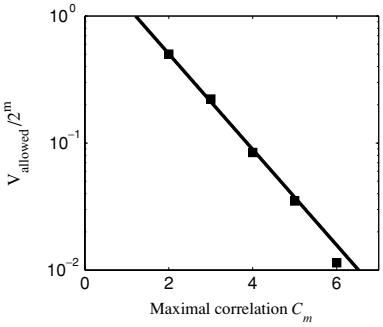

Figure 3. The fraction of the allowed volume, $V_{\text{allowed}}/2^m$ versus the number of two-point constraints imposed. The line represents the best fit $V_{\text{allowed}}/2^m = 2.849\exp\{-0.8667m\}$.

5.2. The allowed region for $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$

We find that imposing a series of ever growing constraints in the form of two-point correlations $\mathcal{C} = \{C_1, C_2\}, \mathcal{C} = \{C_1, C_2, C_3\} \cdots \mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ results in a complicated-shaped, convex polytope of the allowed region in the m-dimensional space. The proof of convexity is given in appendix B. The fraction of volume occupied by the allowed region, $V_{\text{allowed}}/2^m$, decays as m increases (constraints are added). In figure 3, the fraction of the allowed volume is plotted versus the maximal correlation length taken, m, and it appears to fit very well with an exponential decay.

This observation can be *qualitatively* supported by the following theoretical argument[5]. As reported above, the allowed region for a pair of two-point correlations, $\mathcal{C} = \{C_1, C_m\}$ occupies $1 - (1/m)$ of the volume. In the limit of large m we make two assumptions: first, we treat the overall constraint as a combination of *independent* pair constraints, taking into

[5] The calculated allowed fraction (figure 3) slightly differs from the theoretical argument ($e^{-0.8667m/2}$ instead of $e^{-m/2}$). Note that the theoretical argument relies both on the results for $\{C_1, C_m\}$, and some general position assumptions, and therefore, reveals the *qualitative scaling* behaviour. However, the prefactors cannot be precisely calculated based on those assumptions.

account all $m(m-1)/2$ possible pairs. This assumption can rely on the general position argument in high dimensions. Second, we assume that the allowed fraction of volume of a *typical* pair is $1-\mathcal{O}\left(\frac{1}{m}\right)$ in the large m limit. Under these assumptions, the allowed region should roughly be

$$\frac{V_{\text{allowed}}}{2^m} \approx (1-(1/m))^{m\cdot(m-1)/2}$$

and in the $m \to \infty$ limit

$$\frac{V_{\text{allowed}}}{2^m} \approx \mathrm{e}^{-m/2}. \tag{20}$$

6. Conclusions

In this work we addressed the problem of imposing a set of correlations on binary sequences. A correlation, say C_m, is a macroscopic consequence of microscopic interaction among lattice sites located m sites apart. These interactions, as well as the entropy of the system, can be evaluated by the transfer matrix method. We observed that choosing a correlation-set is not arbitrarily free, but limited to an *allowed region* in the correlation-set space, out of which the saddle-point equations cannot be solved. The transfer matrix method suffers from numerical inaccuracy near the boundaries, and becomes a heavy computational task when the correlation length increases. Using the Simplex algorithm, we managed to circumvent these difficulties. The Simplex solution has the following benefits over the statistical mechanics approach:

- Using the Simplex method, exact expressions for the boundaries of the region can be found with no numerical imprecision, and with affordable complexity.
- It is obvious from the Simplex solution that outside the allowed region *no* sequence exists (which is not clear from the saddle-point solution).
- The Simplex solution indicates that the allowed region is *convex* (any point on a line between two allowed points is an allowed point). A proof of convexity is given in appendix B.
- The allowed region found by the Simplex algorithm is valid also for finite sequences.

The volume of the allowed region decays exponentially with the number of correlations imposed, an observation that is supported by an argument in high dimensions.

The findings reported here have practical application in the design of a digital communication system for correlated sequences over noisy channels [7, 8] (and references therein), and we anticipate further applications in various fields.

Appendix A. The allowed region for the set $\mathcal{C} = \{C_1, C_m\}$-combinatorial perspective

For the simple case of a 2D correlation-set $\mathcal{C} = \{C_1, C_m\}$ described in section (5.1), the boundary line can be found from combinatorial arguments by considering a long (but finite!) sequence. Note that for finite sequences of length L, correlations are discrete rather than continuous, with $\frac{4}{L}$ intervals. Consequently, the allowed region becomes a dotted grid of allowed points (smeared to a continuous volume in the limit $L \to \infty$). As an example, we will explore the right-hand boundary $C_4 = 4C_1 - 3$ of the set $\mathcal{C} = \{C_1, C_4\}$ (see figure 4). Application to the other boundary or to different m is straightforward.

At the top right of the region, $C_1 = C_4 = 1$, only the homogenous sequences $(++\cdots++)$ or $(--\cdots--)$ comply with the constraint (the entropy is $\log_2(2)$). By flipping any single element, say $(++\cdots+-+\cdots+)$, one obtains the set $\left\{C_1 = 1-\frac{4}{L}, C_4 = 1-\frac{4}{L}\right\}$ (with the

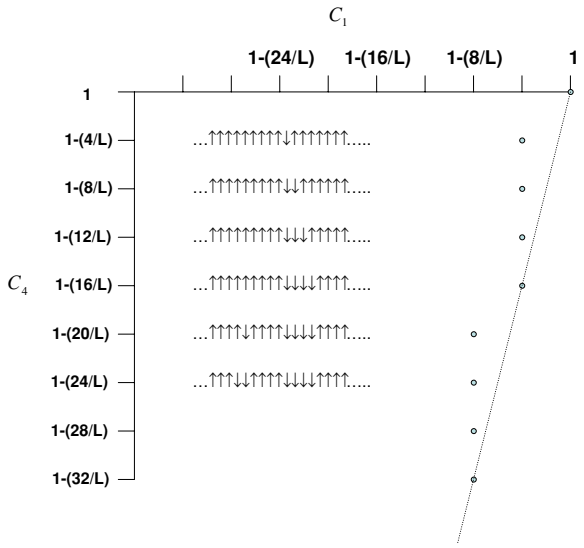

Figure 4. Illustration of possible sequences on the boundary of the allowed region for the correlation-set $\mathcal{C} = \{C_1, C_4\}$. Each dot marks a possible set, the dotted line marks the boundary of the allowed region for infinite sequences, $C_4 = 4C_1 - 3$.

entropy $\log_2(L)$). In an attempt to reconstruct the boundary line, we further flip elements in a manner that will reduce C_4 and retain the *maximal* C_1. This can be achieved by flipping one of the neighbouring elements: $(++\cdots+--+\cdots+)$, resulting in unchanged C_1 and $C_4 = 1 - \frac{8}{L}$. This process can continue for two more elements, resulting in $(+ + \cdots + - - - - + \cdots +)$ with $C_1 = 1 - \frac{4}{L}$, $C_4 = 1 - \frac{16}{L}$, for which the boundary relation, $C_4 = 4C_1 - 3$ is recovered, as demonstrated in figure 4. However, flipping another adjacent element will not decrease C_4 any further (in fact, all the sequences with two 'walls' and the cluster size larger than 3 belong to the same $\{C_1, C_4\}$ set). In order to further decrease C_4, one should flip an element at least four sites apart from the inverted cluster: $(+ + \cdots + - + + + + - - - - + \cdots +)$. Eventually, the sequence $(+ + + + - - - - + + + + - - - - \cdots)$, with $C_1 = \frac{3}{4}$, $C_4 = -1$ represents the bottom-right point of the region.

Appendix B. Proving convexity for the allowed region

The main idea is that one can verify that the set of equalities (such as equations (16) and (17)) can be written in a matrix representation in the following form:

$$\mathbf{M}P = C \tag{B.1}$$

where $\mathbf{M}$ is a matrix with elements ± 1, P represents the marginal probabilities P_i, and C represents the desired correlations and the normalization constant. The inequalities force the probabilities into the range $[0 : 1]$. In order to show convexity we should prove that any point which lies on the line between two 'allowed' points is also 'allowed'. Let C_α and C_β be two 'allowed' points obeying the following equations:

$$\mathbf{M}P_\alpha = C_\alpha \qquad \mathbf{M}P_\beta = C_\beta \tag{B.2}$$

where P_α and P_β are two sets of probabilities solving equation (B.2). A point between C_α and C_β can be written as $C_\gamma = \lambda C_\alpha + (1 - \lambda)C_\beta$, where $0 \leqslant \lambda \leqslant 1$.

Clearly

$$\mathbf{M}P_\gamma = C_\gamma, \tag{B.3}$$

where $P_\gamma \equiv \lambda P_\alpha + (1-\lambda)P_\beta$. Since $0 \leqslant \lambda \leqslant 1$, P_γ is a proper set of probabilities yielding the desired set of autocorrelations. Hence, the 'allowed' region is convex.

References

[1] Bouchaud J, Gefen Y, Potters M and Wyart M 2004 *Quant. Finance* **4** 176–90
[2] Dokholyan N V, Buldyrev S V, Havlin S and Stanley H E 1997 *Phys. Rev. Lett.* **79** 5182–85
[3] Peng C K, Mietus J, Hausdorff J M, Havlin S, Stanley H E and Goldberger A L 1993 *Phys. Rev. Lett.* **70** 1343–46
[4] Dittes F 1996 *Phys. Rev. Lett.* **76** 4651–5
[5] Ein-Dor L, Kanter I and Kinzel W 2002 *Phys. Rev.* E **65** 020102
[6] Dantzig G B, Orden A and Wolfe P 1955 The generalized simplex method for minimizing a linear form under *Pac. J. Math.* **5** 183–95
[7] Kanter I and Kfir H 2003 *Europhys. Lett.* **63** 310–16 (and its cited references)
[8] Kanter I, Kfir H and Keren S 2005 *Prog. Theor. Phys. Suppl.* **157** 184–96
[9] Baxter R J 1982 *Exactly Solved Models in Statistical Mechanics* (New York: Academic)
[10] Dantzig G B 1951 Maximization of a linear function of variables subject to linear inequalities *Activity Analysis of Production and Allocation* ed T C Koopmans (New York: Wiley)
[11] Klee V and Minty G J 1972 How good is the simplex algorithm? *Inequalities III* ed O Shisha (New York: Academic)
[12] Dantzig G B 1963 *Linear Programming and Extensions* (Princeton, NJ: Princeton University Press)
[13] Wu N and Coppins R 1981 *Linear Programming and Extensions* (New York: McGraw-Hill)
[14] Tokhomirov V M 1996 *Am. Math. Mon.* **103** 65–71
[15] Monasson R and Zecchina R 1997 *Phys. Rev.* E **56** 1357–70
[16] Mezard M, Mora T and Zecchina R 2005 *Phys. Rev. Lett.* **94** 197205