

Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests

Simona Cocco, Rémi Monasson

▶ **To cite this version:**

Simona Cocco, Rémi Monasson. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. 2011. hal-00634921

HAL Id: hal-00634921
https://hal.science/hal-00634921

Preprint submitted on 24 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests

S. Cocco[1,2], R. Monasson[1,3]

[1] *Simons Center for Systems Biology, Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540*
[2] *Laboratoire de Physique Statistique de l'ENS, CNRS & UPMC, 24 rue Lhomond, 75005 Paris, France*
[3] *Laboratoire de Physique Théorique de l'ENS, CNRS & UPMC, 24 rue Lhomond, 75005 Paris, France*

We present a procedure to solve the inverse Ising problem, that is to find the interactions between a set of binary variables from the measure of their equilibrium correlations. The method consists in constructing and selecting specific clusters of variables, based on their contributions to the cross-entropy of the Ising model. Small contributions are discarded to avoid overfitting and to make the computation tractable. The properties of the cluster expansion and its performances on synthetic data are studied. To make the implementation easier we give the pseudo-code of the algorithm.

I. INTRODUCTION

The Ising model is a paradigm of statistical physics, and has been extensively studied to understand the equilibrium properties and the nature of the phase transitions in various systems in condensed matter [1]. In its usual formulation, the Ising model is defined over a set of N binary variables σ_i, with $i = 1, 2, \ldots, N$. The variables, called spins, are submitted to a set of N local fields, h_i, and of $\frac{1}{2}N(N-1)$ pairwise couplings, J_{ij}. The observables of the model, such as the average values of the spins or of the spin-spin correlations over the Gibbs measure,

$$\langle \sigma_i \rangle \ , \langle \sigma_k \sigma_l \rangle \ , \tag{1}$$

are well-defined and can be calculated from the knowledge of those interaction parameters. We will refer to the task of calculating (1) given the interaction parameters as to the direct Ising problem.

In many experimental cases, the interaction parameters are unknown, while the values of observables can be estimated from measurements. A natural question is to know if and how the interaction parameters can be deduced from the data ([2–7]). When the coupling matrix is known *a priori* to have a specific and simple structure, this question can be answered with an ordinary fit. For instance, in a two-dimensional and uniform ferromagnet, all couplings vanish but between neighbors on the lattice, and $J_{ij} = J$ for contiguous sites i and j. In such a case, the observable such as the average correlation between neighboring spins, c, depends on a single parameter, J. The measurement of c gives a direct access to a value of J. However, data coming from complex systems arising in biology, sociology, finance, ... can generally not be interpreted with such a simple Ising model, and the fit procedure is much more complicated for two reasons. First, in the absence of any *prior* knowledge about the interaction network, the number of interaction parameters J_{ij} to be inferred scales quadratically with the system size N, and can be very large. Secondly, the quality of the data is a crucial issue. Experimental data are plagued by noise, coming either from the measurement apparatus or from imperfect sampling. The task of fitting a very large number of interaction parameters from 'noisy' data has received much attention in the statistics community, under the name of high-dimensional inference [13].

To be more specific, the inverse Ising problem is defined as follows. Assume that a set of B configurations $\boldsymbol{\sigma}^\tau = \{\sigma_1^\tau, \sigma_2^\tau, \ldots, \sigma_N^\tau\}$, with $\tau = 1, 2, \ldots, B$ are available from measurements. We compute the empirical 1- and 2-point averages through

$$p_i = \frac{1}{B}\sum_{\tau=1}^{B} \sigma_i^\tau \ , \quad p_{kl} = \frac{1}{B}\sum_{\tau=1}^{B} \sigma_k^\tau \, \sigma_l^\tau \ . \tag{2}$$

The inverse Ising problem consists in finding the values of the N local fields, h_i, and of the $\frac{1}{2}N(N-1)$ interactions, J_{ij}, such that the individual and pairwise frequencies of the spins (1) defined from the Gibbs measure coincide with their empirical counterparts, p_i and p_{kl}. While the Gibbs measure corresponding to the Ising model is by no means the unique measure allowing one to reproduce the data p_i and p_{kl}, it is the distribution with the largest entropy doing so [9]. In other words, the Ising model is the least constrained model capable of matching the empirical values of the 1- and 2-point observables. This property explains the recent surge of interest in defining and solving the inverse Ising problem in the context of the analysis of biological, *e.g.* neurobiological [2–4, 10] and proteomic [5, 6] data.

As a result of its generality, the inverse Ising problem has been studied in various fields under different names, such as Boltzmann machine learning in learning theory [11, 12] or graphical model selection in statistical inference [13, 15, 16]. While the research field is currently very active, the diversity of the tools and, sometimes, of the goals make somewhat difficult to compare the results obtained across the disciplines. Several variants of the inverse Ising problem can be defined:

- A: find the interaction network from a set of spin configurations $\boldsymbol{\sigma}^\tau$. It is generally assumed in the graphical model community that the Ising model is exact, that is, that the underlying distribution of the data is truly an Ising model with unknown interaction parameters $\mathbf{J}$. The question is to find which interactions J_{ij} are non zero (or larger than some J_{min} is absolute value), and how many configurations (value of B) should be sampled to achieve this goal with acceptable probability.

- B: find the interactions J_{ij} and the fields h_i from the frequencies p_i, p_{ij} only. Those frequencies should be reproduced within a prescribed accuracy, ϵ, not too small (compared to the error on the data) to avoid overfitting. Note that in general the Ising model is not the true underlying model for the data here; it is only the model with maximal entropy given the constraints on 1- and 2-point correlations.

- C: same as B, but in addition we want to know the entropy (at fixed individual and pairwise frequencies), which measures how many configurations $\boldsymbol{\sigma}$ really contribute to the Gibbs distribution of the Ising model. Computing the entropy is generally intractable for the direct Ising problem, unless correlations decay fast enough [17].

Variants B and C are harder than A: full spin configurations give access to all K-spin correlations, a knowledge which can be used to design fast network structure inference algorithm. Recently, a procedure to solve problem C was proposed, based on ideas and techniques coming from statistical physics [8]. The purpose of the present paper is to discuss its performances and limitations.

It is essential to be aware of the presence of noise in the data, *e.g.* due to the imperfect sampling (finite number B of configurations). A potential risk is overfitting: the network of interactions we find at the end of the inference process could reproduce the mere noisy data, rather than the 'true' interactions. How can one disentangle noise from signal in the data? A popular approach in the statistics community is to require that the inferred interaction network be sparse. The rationale for imposing sparsity is two-fold. First, physical lattices are very sparse, and connect only close sites in the space; it is possible but not at all obvious that networks modeling other *e.g.* biological data enjoy a similar property. Secondly, an Ising model with a sparse interaction network reproducing a set of correlations is a sparing representation of the statistics of the data, and, in much the same spirit as the minimal message length approach [14], should be preferred to models with denser networks. The appeal of the approach is largely due to the fact that imposing sparsity is computationally tractable.

The criterion required by our procedure is not that the interaction network should be sparse, but that the inverse Ising problem should be well-conditioned. To illustrate this notion, consider a set of data, *i.e.* of frequencies p_i, p_{kl}, and assume one has found the solution h_i, J_{kl} to the corresponding inverse Ising problem. Let us now slightly modify one or a few frequencies, say, $p_{12} \to p'_{12} = p_{12} + \delta p_{12}$, and solve again the corresponding inverse Ising problem, with the results h'_i, J'_{kl}. Let $\delta J_{kl} = J'_{kl} - J_{kl}$ and $\delta h_i = h'_i - h_i$ measure the response of the interaction parameters to the small modification of p_{12} alone. Two extreme cases are:

- *Localized response:* the response is restricted to the parameters involving spins 1 and 2 only, *i.e.* $\delta h_1, \delta h_2, \delta J_{12} \neq 0$; it vanishes for all the other parameters.

- *Extended response:* the response spreads all over the spin system, and all the quantities $\delta h_i, \delta J_{kl}$ are non-zero.

Intermediate cases will generically be encountered, and are symbolized in Fig. 1(a)&(b). For instance, if the response is non-zero over a small number of parameters only, which define a 'neighborhood' of the spins $1, 2$, we will consider it is localized. Obviously, the notion of 'smallness' cannot be rigorously defined here, unless the system size N can be made arbitrarily large and sent to infinity.

Drawing our inspiration from the vocabulary of numerical analysis, we will say that the inverse Ising problem is well-conditioned if the response is localized. For a well-conditioned problem, a small change of one or a few variables essentially affects one or a few interaction parameters. On the contrary, most if not all interaction parameters of a ill-conditioned inverse Ising problem are affected by an elementary modification of the data. This notion must be distinguished from the concept of ill-posed problem. As we will see in Section II, the inverse Ising problem is always well-posed, once an appropriate regularization is introduced: given the frequencies, there exists a unique set of interaction parameters reproducing those data, regardless of how hard it is to compute.

Not all inverse Ising problems are well-conditioned. However, it is our opinion that only those ones should be solved. The reason is that, in generic experimental situations, only a (small) region of the system is accessible. Solving the inverse problem attached to this sub-system makes sense only if the problem is well-conditioned. If it is ill-conditioned, extending even by a bit the sub-system would considerably affect the values of most of the inferred parameters (Fig. 1(c)). Hence, the interaction parameters would be very much dependent on the part of the system which is not measured! Such a possibility simply means that the inverse problem, though mathematically well-posed, is not meaningful.

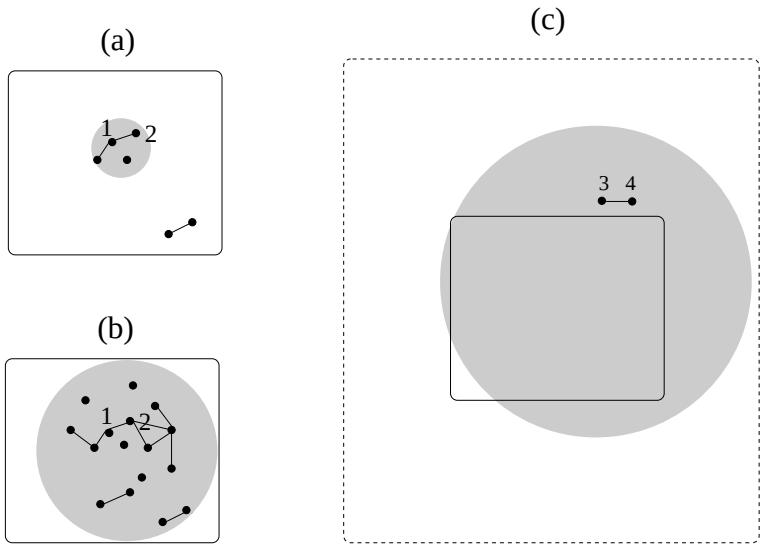


FIG. 1: Schematic representation of a well-conditioned **(a)** and an ill-conditioned **(b)** inverse Ising problems. The gray areas symbolize the set of spins (full dots) whose interactions (links) and fields are affected by a change of the frequency p_{12} of spins 1 and 2. The response is localized in **(a)** and extended in **(b)**. Experiments usually measure a restricted part of the system (dashed contour) only **(c)**. Increasing the size of the measured sub-system, *e.g.* by including the frequencies of the extra-variables 3 and 4, will modify most of the inferred interaction parameters if the problem is ill-conditioned.

Interestingly, the response of the interactions to a change of a few correlations can be localized, while the response of the correlations to a change of a few interactions is extended. An example is given by 'critical' Ising models, where correlations extend over the whole system. However, the corresponding inverse Ising problem may be well-conditioned.

The presence of noise in the data considerably affects the status of the inverse Ising problem. As we will see later, even well-conditioned problems in the limit of perfect sampling ($B \to \infty$) become ill-conditioned as soon as sampling is imperfect (finite B). The same statement holds for the sparsity-based criterion mentioned above: when data are generated by a sparse interaction network, the solution to the inverse Ising model is not sparse as a consequence of imperfect sampling. Only the presence of an explicit and additional regularization forces the solution to be sparse. In much the same way, the procedure we present hereafter builds a well-conditioned inverse Ising problem, which prevents overfitting of the noise. This procedure is based on the expansion of the entropy at fixed frequencies in clusters of spins, a notion closely related to the neighborhoods appearing in the localized responses.

The plan of the article is as follows. In Section II we give the notations and precise definitions of the inverse Ising problem, and briefly review some of the resolution procedures in the literature. In Section III, we explain how the entropy can be expanded as a sum of contributions, one for each cluster (or sub-set) of spins, and review the properties of those entropic contributions. The procedure to truncate the expansion and keep only relevant clusters is discussed in Section IV. The pseudo-codes and details necessary for the implementation of the algorithm can be found in Section V. Applications to artificial data are discussed at length in Section VI. Finally, Section VII presents some perspectives and conclusions. To improve the readability of the paper most technical details have been relegated to technical appendices.

II. THE INVERSE ISING PROBLEM: FORMULATIONS AND ISSUES

A. Maximum Entropy Principle formulation

We consider a system of N binary variables, $\sigma_i = 0, 1$, where $i = 1, 2, \ldots, N$. The average values of the variables, p_i, and of their correlations, p_{kl}, are measured, for instance through the empirical average over B sampled configurations of the system, see equations (2). As the correlations p_{kl} are obtained from the empirical measure, the problem is realizable [18]. Let $\mathbf{p} = \{p_i, p_{kl}\}$ denote the data. The Maximum Entropy Principle (MEP) [9] postulates that the

probabilistic model $P(\boldsymbol{\sigma})$ should maximize the entropy S of the distribution P under the constraints

$$\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) = 1 \ , \quad \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \ \sigma_i = p_i \ , \quad \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \ \sigma_k \, \sigma_l = p_{kl} \ . \tag{3}$$

In practice these constraints are enforced by the Lagrange multipliers λ and $\mathbf{J} = \{h_i, J_{kl}\}$. The maximal entropy is [50]

$$\begin{aligned} S(\mathbf{p}) = \min_{\lambda,\mathbf{J}} \max_{P(\boldsymbol{\sigma})} \Bigg[\ - \ \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \log P(\boldsymbol{\sigma}) + \lambda \left(\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) - 1 \right) \\ + \ \sum_i h_i \left(\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \, \sigma_i - p_i \right) + \sum_{k<l} J_{kl} \left(\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \, \sigma_k \, \sigma_l - p_{kl} \right) \Bigg] \ . \end{aligned} \tag{4}$$

The maximization condition over P shows that the MEP probability corresponds to the Gibbs measure $P_{\mathbf{J}}$ of the celebrated Ising model,

$$P_{\mathbf{J}}[\boldsymbol{\sigma}] = \frac{e^{-H_{Ising}[\boldsymbol{\sigma}|\mathbf{J}]}}{Z[\mathbf{J}]} \tag{5}$$

where the energy function is

$$H_{Ising}[\boldsymbol{\sigma}|\mathbf{J}] = - \sum_i \ h_i \, \sigma_i - \sum_{k<l} \ J_{kl} \, \sigma_k \, \sigma_l \tag{6}$$

and $Z[\mathbf{J}] = \sum_{\boldsymbol{\sigma}} \exp(-H_{Ising}[\boldsymbol{\sigma}|\mathbf{J}])$ denotes the partition function. The values of the couplings and fields [51] are then found through the minimization of

$$S_{Ising}[\mathbf{J}|\mathbf{p}] = \log Z[\mathbf{J}] - \sum_i h_i \ p_i - \sum_{k<l} J_{kl} \ p_{kl} \ . \tag{7}$$

over $\mathbf{J}$. The minimal value of S_{Ising} coincides with S defined in (4).

The cross-entropy S_{Ising} has a simple interpretation in terms of the Kullback-Leibler divergence between the Ising distribution $P_{\mathbf{J}}[\boldsymbol{\sigma}]$ and the empirical measure over the observed configurations, $P_{\rm obs}[\boldsymbol{\sigma}]$. Assume B configurations of the N variables, $\boldsymbol{\sigma}^\tau$, with $\tau = 1, 2, \ldots, B$, are sampled. We define the empirical distribution through

$$P_{\rm obs}[\boldsymbol{\sigma}] = \frac{1}{B} \sum_{\tau=1}^{B} \delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}^\tau} \ , \tag{8}$$

where δ denotes the N-dimensional Kronecker delta function. It is easy to check from (7) that

$$S_{Ising}[\mathbf{J}|\mathbf{p}] = - \sum_{\boldsymbol{\sigma}} P_{\rm obs}[\boldsymbol{\sigma}] \log P_{\mathbf{J}}[\boldsymbol{\sigma}] = - \sum_{\boldsymbol{\sigma}} P_{\rm obs}[\boldsymbol{\sigma}] \log P_{\rm obs}[\boldsymbol{\sigma}] + D\big(P_{\rm obs} || P_{\mathbf{J}}\big) \ , \tag{9}$$

where D denotes the KL-divergence. Hence, the minimization procedure over $\mathbf{J}$ ensures that the 'best' Ising measure (as close as possible to the empirical measure) is found.

B. Regularization and Bayesian formulation

We consider the Hessian of the cross-entropy S_{Ising}, also called Fisher information matrix, which is a matrix of dimension $\frac{1}{2}N(N+1)$, defined through

$$\boldsymbol{\chi} = \frac{\partial^2 S_{Ising}}{\partial \mathbf{J} \partial \mathbf{J}} = \begin{pmatrix} \chi_{i,i'} & \chi_{i,k'l'} \\ \chi_{kl,i'} & \chi_{kl,k'l'} \end{pmatrix} \ . \tag{10}$$

The entries of $\boldsymbol{\chi}$ are obtained upon repeated differentiations of the partition function $Z[\mathbf{J}]$, and can be expressed in terms of averages over the Ising Gibbs measure $\langle \cdot \rangle_{\mathbf{J}}$,

$$\begin{aligned} \chi_{i,i'} &= \langle \sigma_i \sigma_{i'} \rangle_{\mathbf{J}} - \langle \sigma_i \rangle_{\mathbf{J}} \langle \sigma_{i'} \rangle_{\mathbf{J}} \ , \\ \chi_{i,k'l'} &= \langle \sigma_i \sigma_{k'} \sigma_{l'} \rangle_{\mathbf{J}} - \langle \sigma_i \rangle_{\mathbf{J}} \langle \sigma_{k'} \sigma_{l'} \rangle_{\mathbf{J}} \ , \\ \chi_{kl,k'l'} &= \langle \sigma_k \sigma_l \sigma_{k'} \sigma_{l'} \rangle_{\mathbf{J}} - \langle \sigma_k \sigma_l \rangle_{\mathbf{J}} \langle \sigma_{k'} \sigma_{l'} \rangle_{\mathbf{J}} \ . \end{aligned} \tag{11}$$

Consider now an arbitrary $\frac{1}{2}N(N+1)$-dimensional vector $\mathbf{x} = \{x_i, x_{kl}\}$. The quadratic form

$$\mathbf{x}^\dagger \cdot \boldsymbol{\chi} \cdot \mathbf{x} = \left\langle \left(\sum_i x_i \big(\sigma_i - \langle \sigma_i \rangle_{\mathbf{J}}\big) + \sum_{k<l} x_{kl} \big(\sigma_k \sigma_l - \langle \sigma_k \sigma_l \rangle_{\mathbf{J}}\big) \right)^2 \right\rangle_{\mathbf{J}} \tag{12}$$

is semi-definite positive. Hence, S_{Ising} is a convex function.

However the minimum is not guaranteed to be unique if $\boldsymbol{\chi}$ has zero modes, nor to be finite. To circumvent those difficulties, one can 'regularize' the cross-entropy S_{Ising} by adding a quadratic term in the interaction parameters, which forces $\boldsymbol{\chi}$ to become definite positive, and ensures the uniqueness and finiteness of the minimum of S_{Ising}. In many applications, no regularization is needed for the fields h_i. The reason can be understood intuitively as follows. Consider a data set where all variables are independent, with small but strictly positive means p_i. Then, the empirical average products, p_{kl}, may vanish if the number B of sampled configurations is not much larger than $(p_k p_l)^{-1}$. This condition is often violated in practical applications, *e.g.* the analysis of neurobiological or protein data [2, 5, 10]. Hence, poor sampling may produce infinite negative couplings. We therefore add the following regularization term to S_{Ising},

$$\gamma \sum_{k<l} J_{kl}^2 \, p_k(1-p_k)p_l(1-p_l) \ . \tag{13}$$

The precise expression of the regularization term is somewhat arbitrary, and is a matter of convenience. The dependence on the p_i's in (13) will be explained in Section III B. Other regularization schemes, based on the L_1 norm rather than on the L_2 norm are possible, such as

$$\gamma \sum_{k<l} |J_{kl}| \, \sqrt{p_k(1-p_k)p_l(1-p_l)} \ . \tag{14}$$

The above regularization is especially popular among the graphical model selection community [15], and favors sparse coupling networks, *i.e.* with many zero interactions.

The introduction of a regularization is natural in the context of Bayesian inference. The Gibbs probability $P_{\mathbf{J}}[\boldsymbol{\sigma}]$ defines the likelihood of a configuration $\boldsymbol{\sigma}$. The likelihood of a set of B independently drawn configurations $\boldsymbol{\sigma}^\tau$ is given by the product of the likelihoods of each configuration. The posterior probability of the parameters (fields and couplings) $\mathbf{J}$ given the configurations $\boldsymbol{\sigma}^\tau$, $\tau = 1, 2, \ldots, B$, is, according to Bayes' rule,

$$P_{post}[\mathbf{J}|\{\boldsymbol{\sigma}^\tau\}] \propto \prod_{\tau=1}^{B} P_{\mathbf{J}}[\boldsymbol{\sigma}^\tau] \ P_0[\mathbf{J}] \ , \tag{15}$$

up to an irrelevant $\mathbf{J}$-independent multiplicative factor. In the equation above, P_0 is a prior probability over the couplings and fields, encoding the knowledge about their values in the absence of any data. Taking the logarithm of (15), we obtain, up to an additive $\mathbf{J}$-independent constant,

$$\log P_{post}[\mathbf{J}|\{\boldsymbol{\sigma}^\tau\}] = -B \ S_{Ising}[\mathbf{J}|\mathbf{p}] + \log P_0[\mathbf{J}] \ . \tag{16}$$

Hence, the most likely value for the parameters $\mathbf{J}$ is the one minimizing $S_{Ising}[\mathbf{J}|\mathbf{p}] - \frac{1}{B}\log P_0[\mathbf{J}]$. The regularization terms (13) and (14) then correspond to, respectively, Gaussian and exponential priors over the parameters. In addition, as the prior is independent of the number B of configurations, we expect the strength γ to scale as $\frac{1}{B}$. The optimal value of γ can be also determined based on Bayesian criteria [10, 39] (Appendix A).

We emphasize that the Bayesian framework changes the scope of the inference. While the MEP aims to reproduce the data, the presence of a regularization term leads to a compromise between two different objectives: finding an Ising model whose observables (one- and two-point functions) are close to the empirical values and ensuring that the interaction parameters $\mathbf{J}$ have a large prior probability P_0. In other words, a compromise is sought between the faithfulness to the data and the prior knowledge about the solution. The latter is especially important in the case of poor sampling (small value of B or data corrupted by noise). For instance, the regularization term based on the L_1–norm (14) generally produces more couplings equal to zero than its L_2–norm counterpart (13). This property is desirable if one a priori knows that the interaction graph is sparse. Hence, the introduction of a regularization term can be interpreted as an attempt to approximately solve the inverse Ising problem while fulfilling an important constraint about the structure of the solution. We will discuss the nature of the structural constraints corresponding to our adaptive cluster algorithm in Section IV C.

Knowledge of the inverse of the Fisher information matrix, $\boldsymbol{\chi}^{-1}$, allows for the computation of the statistical fluctuations of the inferred fields and couplings due to a finite number B of sampled configurations. According to

the asymptotic theory of inference, the posterior probability $P_{post}[\mathbf{J}|\{\boldsymbol{\sigma}^\tau\}]$ over the fields and couplings becomes, as B gets very large, a normal law centered in the minimum of $S_{Ising}[\mathbf{J}|\mathbf{p}]$. The covariance matrix of this normal law is simply given by $\frac{1}{B}\,\boldsymbol{\chi}^{-1}$. Consequently the standard deviations of the fields h_i and of the couplings J_{kl} are, respectively,

$$\delta h_i = \sqrt{\frac{1}{B}\,(\chi^{-1})_{i,i}}\ , \quad \delta J_{kl} = \sqrt{\frac{1}{B}\,(\chi^{-1})_{kl,kl}}\ . \tag{17}$$

In order to remove the zero modes of $\boldsymbol{\chi}$ and have a well-defined inverse matrix $\boldsymbol{\chi}^{-1}$, the Ising model entropy S_{Ising} (7) can be added a regularization term, *e.g.* (13), which guarantees that $\boldsymbol{\chi}$ is positively defined.

The Fisher information matrix, $\boldsymbol{\chi}$, can also be used to estimate the statistical deviations of the observables coming from the finite sampling. If the data were generated by an Ising model with parameters $\mathbf{J}$, we would expect, again in the large B setting, that the frequencies p_i, p_{kl} would be normally distributed with a covariance matrix equal to $\frac{1}{B}\,\boldsymbol{\chi}$. Hence, the typical uncertainties over the 1- and 2-point frequencies are given by

$$\delta p_i = \sqrt{\frac{1}{B}\,\chi_{i,i}} = \sqrt{\frac{\langle\sigma_i\rangle_{\mathbf{J}}(1-\langle\sigma_i\rangle_{\mathbf{J}})}{B}}\ , \qquad \delta p_{kl} = \sqrt{\frac{1}{B}\,\chi_{kl,kl}} = \sqrt{\frac{\langle\sigma_k\sigma_l\rangle_{\mathbf{J}}(1-\langle\sigma_k\sigma_l\rangle_{\mathbf{J}})}{B}}\ . \tag{18}$$

In practice, we can replace the Gibbs averages above with the empirical averages p_i and p_{kl} to obtain estimates for the expected deviations. These estimates will be used to decide whether the inference procedure is reliable, or leads to an overfitting of the data in Section VI.

C. Methods

The inverse Ising problem has been studied in statistics, under the name of graphical model selection, in the machine learning community under the name of (inverse) Boltzmann machine learning, and in the statistical physics literature. Different methods have been developed, with various applications. Some of the methods are briefly discussed below.

A direct calculation of the partition function $Z[\mathbf{J}]$ generally requires a time growing exponentially with the number N of variables, and is not feasible when N exceeds a few tens. Inference procedures therefore tend to avoid the computation of $Z[\mathbf{J}]$:

- A popular algorithm is the *Boltzmann learning* procedure, where the fields and couplings are iteratively updated until the averages $\langle\sigma_i\rangle_{\mathbf{J}}$'s and $\langle\sigma_k\sigma_l\rangle_{\mathbf{J}}$'s, calculated from Monte Carlo simulations, match the imposed values [12]. The number of updatings can be very large in the absence of a good initial guess for the parameters $\mathbf{J}$. Furthermore, for each set of parameters, thermalization may require prohibitive computational efforts for large system sizes N, and problems with more than a few tens of spins can hardly be tackled. Finally, learning data exactly leads to overfitting in the case of poor sampling.

- the *Pseudo-Likelihood*-based algorithm by Ravikumar *et al.* [15, 16] is an extension to the binary variable case of Meinshausen and Bühlmann's algorithm [20] and is related to a renormalisation approach introduced by Swendsen [19]. The procedure requires the complete knowledge of the configurations $\{\boldsymbol{\sigma}^\tau\}$ (and not only of the one- and two-point functions $\mathbf{p}$). The starting point is given by well-known Callen's identities for the Ising model,

$$\langle\sigma_i\rangle_{\mathbf{J}} = \left\langle \frac{1}{1+\exp\left(-\sum_j J_{ij}\sigma_j - h_i\right)} \right\rangle_{\mathbf{J}} \simeq \frac{1}{B}\sum_{\tau=1}^{B} \frac{1}{1+\exp\left(-\sum_j J_{ij}\sigma_j^\tau - h_i\right)} \tag{19}$$

  where the last approximation consists in replacing the Gibbs average with the empirical average over the sampled configurations. Imposing that the Gibbs average $\langle\sigma_i\rangle_{\mathbf{J}}$ coincides with p_i is equivalent to minimizing the following pseudo-likelihood over the field h_i,

$$S_{i,PL}[h_i, \{J_{ij}, j\neq i\}] = \frac{1}{B}\sum_{\tau=1}^{B} \log\left[1+\exp\left(\sum_j J_{ij}\sigma_j^\tau + h_i\right)\right] - h_i\, p_i - \sum_{j(\neq i)} J_{ij} p_{ij}\ . \tag{20}$$

  The minimization equations over the couplings J_{ij}, with $j\neq i$ (and fixed i), correspond to Callen identities for two-point functions. Informally speaking, the pseudo-likelihood approach simplifies the original N-body problem into N independent 1-body problem, each one in a bath of $N-1$ quenched variables. Note that the

couplings J_{ij} and J_{ji} (found by minimizing $S_{j,PL}$) will generally not be equal. However, as far as graphical model selection is concerned, what matters is whether J_{ij} and J_{ji} are both different from zero.

The pseudo-entropy $S_{i,PL}$ is convex, and can be minimized after addition of a L_1–norm regularization term [13, 15, 21]. The procedure is guaranteed to find strong enough couplings [52] in a polynomial time in N, provided that the data were generated by an Ising model (which is usually not the case in practical applications) and that a quantity closely related to the susceptibility $\boldsymbol{\chi}$ (10) is small enough. The latter condition holds for weak couplings and may break down for strong couplings [16]. For a review of the literature in the statistics community, see [13].

In specific cases, however, the partition function can be obtained in polynomial time. Two tractable examples are:

- *Mean-field models*, which are characterized by dense but weak interactions. An example is the Sherrington-Kirkpatrick model where every pair of spins interact through couplings of the order of $N^{-1/2}$ [27]. The entropy $S[\mathbf{p}]$ coincides asymptotically with

$$S_{MF}(\mathbf{p}) = \frac{1}{2}\log\det M(\mathbf{p}), \text{where} \quad M_{ij}(\mathbf{p}) = \frac{p_{ij}-p_ip_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}, \tag{21}$$

  which can be calculated in $O(N^3)$ time [11, 30]. Expression (21) has been obtained from the high temperature expansion [23–25] of the Legendre transform of the free energy , and is consistent with the so-called TAP equations [26]. The derivative of S_{MF} with respect to $\mathbf{p}$ gives the value of the couplings and the fields,

$$\begin{aligned}(J_{MF})_{kl} &= -\frac{\partial S_{MF}}{\partial p_{kl}} = -\frac{(M^{-1})_{kl}}{\sqrt{p_k(1-p_k)p_l(1-p_l)}}\ ,\\ (h_{MF})_i &= -\frac{\partial S_{MF}}{\partial p_i} = \sum_{j(\neq i)}(J_{MF})_{ij}\left(c_{ij}\frac{p_i-\frac{1}{2}}{p_i(1-p_i)}-p_j\right)\ ,\end{aligned} \tag{22}$$

  where $c_{ij} = p_{ij} - p_ip_j$ is the connected correlation. From a practical point of view, expression (21) is a good approximation for solving the inverse Ising problem [28–30] on dense and weak interaction networks , but fails to reproduce dilute graphs with strong interactions.

- Ising models on tree-like structures, *i.e.* with no or few interaction loops. *Message passing methods* are guaranteed to solve the associated inverse Ising problems. For trees, the partition functions can be calculated in a time linear in N. Sparse networks of strong interactions with long-range loops, such as Erdös-Renyi random graphs, can also be successfully treated in polynomial time by message-passing procedures [5, 31, 32]. However, these methods generally break down in the presence of strongly interacting groups (clusters) of spins.

When an exact calculation of the partition function is out-of-reach, accurate estimates can be obtained through cluster expansions. Expansions have a rich history in statistical mechanics, *e.g.* the virial expansion in the theory of liquids [33, 34]. However, cluster expansions suffer from several drawbacks. First, in cluster variational methods [31, 35], the calculation of the contributions coming from each cluster generally involves the resolution of non trivial and self-consistent equations for the local fields, which seriously limits the maximal size of clusters considered in the expansion. Secondly, the composition and the size of the clusters is usually fixed *a priori*, and does not adapt to the specificity of the data [10]. The combinatorial growth of the number of clusters with their size entails strong limits upon the maximal sizes of the network, N, and of the clusters, K. Last of all, cluster expansions generally ignore the issue of overfitting.

Recently, we have proposed a new cluster expansion, where clusters are built recursively, and are selected or discarded, according to their contribution to the cross-entropy S [8]. This selection procedure allows us to fully account for the complex interaction patterns present in experimental systems, while preventing a blow-up of the computational time. The purpose of this paper is to illustrate this method and discuss its advantages and limitations.

III. CLUSTER EXPANSION OF THE CROSS-ENTROPY

A. Principle of the expansion

In this Section, we propose a cluster expansion for the entropy $S(\mathbf{p})$. A cluster, Γ, is defined here as a non-empty subset of $(1, 2, \ldots, N)$. To illustrate how the expansion is built we start with the simple cases of systems with a few variables ($N = 1, 2$), in the absence of the regularization term (13).

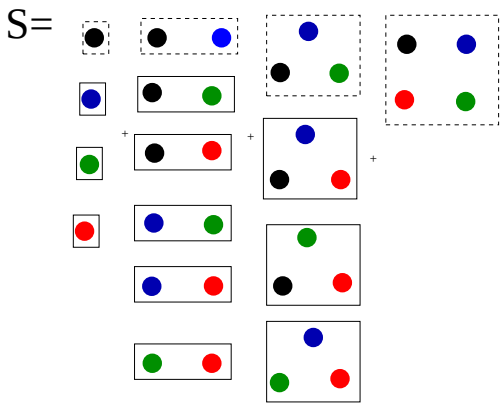


FIG. 2: Decomposition of the cross-entropy $S(\mathbf{p})$ for a system of 4 spins, indicated with different colors, as the sum of cluster contributions. Each cluster-entropy $\Delta S_\Gamma(\mathbf{p})$ depends only on the one- and two-point frequencies of the variables in the cluster: it can be calculated in a recursive way, see main text. Dotted clusters are decomposed into a diagrammatic expansion in Fig. 3.

Consider first the case of a single variable, $N = 1$, with average value p_1. The entropy $S(p_1)$ can be easily computed according to the definitions given in Section II, with the result

$$\begin{aligned} S_{(1)}(p_1) &= \min_{h_1} S_{Ising}[h_1|p_1] = \min_{h_1}\left[\log\left(1+e^{h_1}\right) - h_1 p_1\right] \\ &= -p_1 \log p_1 - (1-p_1)\log(1-p_1)\ . \end{aligned} \tag{23}$$

We recognize the well-known expression for the entropy of a 0-1 variable with mean value p_1. For reasons which will be obvious in the next paragraph, we will hereafter use the notation $\Delta S_{(1)}(p_1)$ to denote the same quantity as $S_{(1)}(p_1)$. The subscript (1) refers to the index of the (unique) variable in the system.

Consider next a system with two variables, with mean values p_1, p_2 and two-point average p_{12}. The entropy $S_{(1,2)}(p_1, p_2, p_{12})$ can be explicitly computed:

$$\begin{aligned} S_{(1,2)}(p_1,p_2,p_{12}) &= \min_{h_1,h_2,J_{12}} S_{Ising}[h_1,h_2,J_{12}|p_1,p_2,p_{12}] \\ &= \min_{h_1,h_2,J_{12}}\left[\log\left(1+e^{h_1}+e^{h_2}+e^{h_1+h_2+J_{12}}\right) - h_1p_1 - h_2p_2 - J_{12}p_{12}\right] \\ &= -(p_1-p_{12})\,\log\left(p_1-p_{12}\right) - (p_2-p_{12})\log(p_2-p_{12}) \\ &- p_{12}\log p_{12} - (1-p_1-p_2+p_{12})\,\log\left(1-p_1-p_2+p_{12}\right)\ . \end{aligned} \tag{24}$$

We now define the entropy $\Delta S_{(1,2)}$ of the cluster of the two variables $1, 2$ as the difference between the entropy $S_{(1,2)}(p_1, p_2, p_{12})$ calculated above and the two single-variable contributions $\Delta S_{(1)}(p_1)$ and $\Delta S_{(1)}(p_2)$ coming from the variables 1 and 2 taken separately:

$$\Delta S_{(1,2)}(p_1,p_2,p_{12}) = S_{(1,2)}(p_1,p_2,p_{12}) - \Delta S_{(1)}(p_1) - \Delta S_{(2)}(p_2)\ . \tag{25}$$

In other words, $\Delta S_{(1,2)}$ measures the loss of entropy between the system of two isolated variables, constrained to have means equal to, respectively, p_1 and p_2, and the same system when, in addition, the average product of the variables is constrained to take value p_{12}. Using expressions (23) and (24), we find

$$\begin{aligned} \Delta S_{(1,2)}(p_1,p_2,p_{12}) &= -(p_1-p_{12})\log\left(\frac{p_1-p_{12}}{p_1-p_1p_2}\right) - (p_2-p_{12})\log\left(\frac{p_2-p_{12}}{p_2-p_1p_2}\right) \\ &- p_{12}\log\left(\frac{p_{12}}{p_1p_2}\right) - (1-p_1-p_2+p_{12})\log\left(\frac{1-p_1-p_2+p_{12}}{1-p_1-p_2+p_1p_2}\right)\ . \end{aligned} \tag{26}$$

The entropy of the cluster $(1, 2)$ is therefore equal to the Kullback-Leibler divergence between the true distribution of probability over the two spins and the one corresponding to two independent spins with averages p_1 and p_2. It vanishes for $p_{12} = p_1p_2$.

Formula (25) can be generalized to define the entropies of clusters with larger sizes $N \geq 3$. Again $\mathbf{p} = \{p_i, p_{kl}\}$ denotes the data. For any non-empty subset Γ including $1 \leq K \leq N$ variables, we define two entropies:

- the subset-entropy $S_\Gamma(\mathbf{p})$, which is the entropy of the subset of the K variables for fixed data. It is defined as the right hand side of (4), when the variable indices, i, k, l are restricted to Γ. Note that, when the subset Γ includes all N variables, $S_\Gamma(\mathbf{p})$ coincides with $S(\mathbf{p})$.

- the cluster-entropy $\Delta S_\Gamma(\mathbf{p})$, which is the remaining contribution to the subset-entropy $S_\Gamma(\mathbf{p})$, once all other cluster-entropies of smaller clusters have been substracted. The cluster entropies are then implicitly defined through the identity

$$S_\Gamma(\mathbf{p}) = \sum_{\Gamma' \subset \Gamma} \Delta S_{\Gamma'}(\mathbf{p}) \ , \tag{27}$$

  where the sums runs over all $2^K - 1$ non-empty clusters Γ' of variables in Γ.

Identity (27) states that the entropy of a system (for fixed data) is equal to the sum of the entropies of all its clusters. Figure 2 sketches the cluster decomposition of the entropy for a system of $N = 4$ variables.

For $\Gamma = (1)$, equation (27) simply expresses that $S_{(1)}(p_1) = \Delta S_{(1)}(p_1)$. For $\Gamma = (1, 2)$, equation (27) coincides with (25). For $\Gamma = (1, 2, 3)$, we obtain the definition of the entropy of a cluster made of a triplet of variables:

$$\begin{aligned} \Delta S_{(1,2,3)}(p_1, p_2, p_3, p_{12}, p_{13}, p_{23}) &= S_{(1,2,3)})(p_1, p_2, p_3, p_{12}, p_{13}, p_{23}) - \Delta S_{(1)}(p_1) - \Delta S_{(2)}(p_2) - \Delta S_{(3)}(p_3) \\ &- \Delta S_{(1,2)}(p_1, p_2, p_{12}) - \Delta S_{(1,3)}(p_1, p_3, p_{13}) - \Delta S_{(2,3)}(p_2, p_3, p_{23}) \ . \end{aligned} \tag{28}$$

The analytical expression of the cluster-entropy $\Delta S_{(1,2,3)}$ is given in Appendix B.

The examples above illustrate three general properties of cluster-entropies:

- the entropy of the cluster Γ, ΔS_Γ, depends only on the frequencies p_i, p_{ij} of the variables i, j in the cluster Γ (and not on all the data in $\mathbf{p}$).

- the entropy of a cluster with, say, K variables, can be recursively calculated from the knowledge of the subset-entropies $S_{\Gamma'}(\mathbf{p})$ of all the subsets $\Gamma' \in \Gamma$ with $K' \leq K$ variables. According to Möbius inversion formula,

$$\Delta S_\Gamma(\mathbf{p}) = \sum_{\Gamma' \subset \Gamma} (-1)^{K'-K} \ S_{\Gamma'}(\mathbf{p}) \ . \tag{29}$$

- the sum of the entropies of all $2^N - 1$ clusters of a system of N spins is the exact entropy of the system, see (27) with $\Gamma = (1, 2, \ldots, N)$.

In practice, to calculate $S(\mathbf{p})$, one first computes the partition function $Z[\mathbf{J}]$ by summing over the 2^K configurations $\boldsymbol{\sigma}$ and, then, minimizes $S_{Ising}[\mathbf{J}|\mathbf{p}]$ (7) over the interaction parameters $\mathbf{J}$. The minimization of a convex function of $\frac{1}{2}K(K+1)$ variables can be done in time growing polynomially with K. Moreover the addition of the regularization term (13) can be easily handled. The limiting step is therefore the calculation of Z, which can be done exactly for clusters with less than, say, $K = 20$ spins.

Hence, only a small number of the $2^N - 1$ terms in (27) can be calculated. In the present work we claim that, in a wide set of circumstances, a good approximation to the entropy $S(\mathbf{p})$ can be already obtained from the contributions of well-chosen clusters of small sizes,

$$S(\mathbf{p}) \simeq \sum_{\Gamma \in L} \Delta S_\Gamma(\mathbf{p}) \ , \tag{30}$$

We will explain in Section IV how the list of selected clusters, L, is established.

B. The reference entropy S_0

So far we have explained how the entropy $S(\mathbf{p})$ can be expanded as a sum of contributions $\Delta S_\Gamma(\mathbf{p})$ attached to the clusters Γ. In this Section we present the expansion against a reference entropy, $S_0(\mathbf{p})$, and two possible choices for the reference entropy.

The idea underlying the introduction of a reference entropy is the following. Assume one can calculate a (rough) approximation $S_0(\mathbf{p})$ to the true entropy $S(\mathbf{p})$. Then, the difference $S(\mathbf{p}) - S_0(\mathbf{p})$ is smaller than $S(\mathbf{p})$, and it makes sense to expand the former rather than the latter. We expect, indeed, the cluster-entropies to be smaller when the

reference entropy $S_0(\mathbf{p})$ is substracted from the true entropy. We substitute the original definition (27) with the new definition

$$S(\mathbf{p}) = S_0(\mathbf{p}) + \sum_{\Gamma \subset (1,2,\ldots,N)} \Delta S_\Gamma(\mathbf{p}) \ . \tag{31}$$

With this new definition, the values of the cluster-entropies ΔS_Γ depend on the choice of S_0; the previous definition (27) is found back when $S_0 = 0$. The procedure for the calculation of the cluster-entropies $\Delta S_\Gamma(\mathbf{p})$ is the same as in Section III A, upon replacement of $S(\mathbf{p})$ with $S(\mathbf{p}) - S_0(\mathbf{p})$. The three properties of the cluster expansion listed above still hold.

Our final estimate for the entropy will be, compare to (30),

$$S(\mathbf{p}) \simeq S_0(\mathbf{p}) + \sum_{\Gamma \in L} \Delta S_\Gamma(\mathbf{p}) \ . \tag{32}$$

Hence, the cluster expansion is a way to calculate a correction to the approximation S_0 to the true entropy S. Obviously, the introduction of a reference entropy is useful in practice only if $S_0(\mathbf{p})$ can be quickly calculated for the entire system of size N. In other words, the computational effort required to obtain S_0 should scale only polynomially with N. A natural choice for the reference entropy is $S_0 = S_{MF}$ (21), the mean-field entropy discussed in Section II C. As the calculation of S_{MF} requires the one of the determinant of the matrix $M(\mathbf{p})$, it can be performed in a time scaling as N^3 only. In addition, we expect S_{MF} to be a sensible approximation to S for systems with rather weak interactions. Corrections coming from the strongest interactions will be taken care of by the cluster expansion.

Regularized versions of the Mean Field entropy can be derived as follows. First, we use the MF expression for the cross-entropy at fixed couplings J_{kl} and frequencies p_i, see (7) and [24], to rewrite

$$S_{Ising}(\{p_i\},\{J_{kl}\}) = -\frac{1}{2}\log\det\big(\mathrm{Id} - J'\big) - \sum_{k<l} J_{kl}\,(p_{kl} - p_k\,p_l) \ , \quad \text{where} \quad J'_{kl} = J_{kl}\sqrt{p_k(1-p_k)p_l(1-p_l)} \ , \tag{33}$$

and Id denotes the N-dimensional identity matrix. We consider the L_2-norm regularization (13). The entropy at fixed data $\mathbf{p}$ is

$$\begin{aligned} S_0(\mathbf{p}) &= \min_{\{J_{kl}\}} \left[S_{Ising}(\{p_i\},\{J_{kl}\}) + \gamma \sum_{k<l} J_{kl}^2\, p_k(1-p_k)p_l(1-p_l) \right] \\ &= \min_{\{J'_{kl}\}} \left[-\frac{1}{2}\log\det\big(\mathrm{Id} - J'\big) - \frac{1}{2}\mathrm{Trace}\big(J'\cdot M(\mathbf{p})\big) + \frac{\gamma}{2}\mathrm{Trace}\big(J'\big)^2 \right] , \end{aligned} \tag{34}$$

where $M(\mathbf{p})$ is defined in (21). The optimal interaction matrix J' is the root of the equation

$$\big(\mathrm{Id} - J'\big)^{-1} - M(\mathbf{p}) + \gamma\, J' = 0 \ . \tag{35}$$

Hence, J' has the same eigenvectors as $M(\mathbf{p})$, a consequence of the dependence on p_i we have chosen for the quadratic regularization term in (13). Let j_q denote its q^{th} eigenvalue, and $\hat{m}_q = (1-j_q)^{-1}$. Then,

$$S_0(\mathbf{p},\gamma) = \frac{1}{2}\sum_{q=1}^{N} \big(\log\hat{m}_q + 1 - \hat{m}_q\big) \ , \tag{36}$$

where $\hat{m}_q$ is the largest root of $\hat{m}_q^2 - \hat{m}_q(m_q - \gamma) = \gamma$, and m_q is the q^{th} eigenvalue of $M(\mathbf{p})$. Note that $\hat{m}_q = m_q$ when $\gamma = 0$, as expected.

C. Properties of the cluster entropies ΔS_Γ

1. *Diagrammatic expansion in powers of the connected correlations*

A better understanding of the cluster expansion and of the role of the reference entropy S_0 can be gained through the diagrammatic expansion of the entropy $S(\mathbf{p})$ in powers of the connected correlations (high-temperature expansion),

$$c_{ij} = p_{ij} - p_i p_j \ . \tag{37}$$

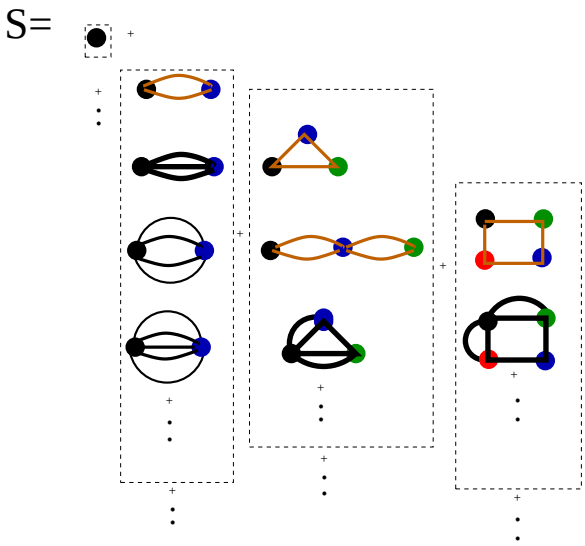


FIG. 3: Diagrammatic expansion of the cross-entropy $S(\mathbf{p})$. A cluster-entropy (see Fig. 2) is the infinite sum of all the diagrams in a box (dashed contour), linking the K sites in the cluster. Each link in a diagram carries M_{ij}, and each site p_i; in addition, each diagram carries a multiplicative factor, which is a function of the p_i's. In the Figure only one cluster among all $\binom{N}{K}$ clusters is represented. Only the first diagrams with non–zero coefficients are drawn. Loop diagrams are analytically summed up and removed from the expansion through the reference entropy $S_0 = S_{MF}$; Eulerian circuit diagrams (brown/gray) are partly removed, see main text. Diagrams giving the largest contributions to the universal central peak of the cluster-entropy distribution (Appendix C) are shown in bold.

Note that the entry M_{ij} of the matrix M defined in (21) vanishes linearly with c_{ij}. Thus, an expansion in powers of c_{ij} is equivalent to an expansion in powers of M_{ij}. A procedure to derive in a systematic way the diagrammatic expansion of $S(\mathbf{p})$ is proposed in [30]. The diagrammatic expansion provides a simple representation of the cluster-entropies, in which the entropy $S(\mathbf{p})$ can be represented as a sum of connected diagrams (Fig. 3). Each diagram is made of sites, connected or not by one or more edges. Each point symbolizes a variable, and carries a factor p_i. The presence of $n(\geq 0)$ edges between the sites k and l results in a multiplicative factor $(c_{kl})^n$. The contribution of a diagram to the entropy is the product of the previous factors, times a function of the p_i specific to the topology of the diagram, see [30]. Diagrams of interest include (Fig. 3):

- the N single-point diagrams, whose contributions are $\Delta S_{(i)}(p_i)$;
- the 'loop' diagrams, which consist of a circuit with K edges going through K sites $i_1 \to i_2 \to \ldots \to i_K \to i_1$, whose contributions to the entropy are

$$S_{loop}(\mathbf{p}|i_1, i_2, \ldots, i_K) = (-1)^{K-1} M_{i_1,i_2} M_{i_2,i_3} \ldots M_{i_{K-1},i_K} M_{i_K,i_1} \ ; \tag{38}$$

- the Eulerian circuit diagrams, for which there exists a closed path visiting each edge exactly once;
- the non-Eulerian diagrams, with the lowest number of links (smallest power in M).

The entropy for two variables i, j, $S(p_i, p_j, p_{ij})$ (24), is the sum of the two single-point diagrams i and j, plus the sum of all connected diagrams made of the two sites i and j with an arbirtrary large number of edges ($n \geq 2$) in between (first two columns in Fig. 3). According to (25), the cluster-entropy $\Delta S_{(i,j)}(p_i, p_j, p_{ij})$ is equal to the latter sum (second column in Fig. 3). More generally, the entropy of a cluster $\Delta S_\Gamma(\mathbf{p})$ is the infinite sum of all diagrams whose sites are the indices in Γ.

We now interpret the Mean Field expression for the entropy, S_{MF}, in the diagrammatic framework. We start from identity (21), and rewrite,

$$S_{MF}(\mathbf{p}) = \frac{1}{2}\text{Trace}\log M = \frac{1}{2}\text{Trace}\log\left[\text{Id} - \left(\text{Id} - M\right)\right] = \sum_{K\geq 1} \frac{-\text{Trace}\left[\left(\text{Id} - M\right)^K\right]}{2K} \ . \tag{39}$$

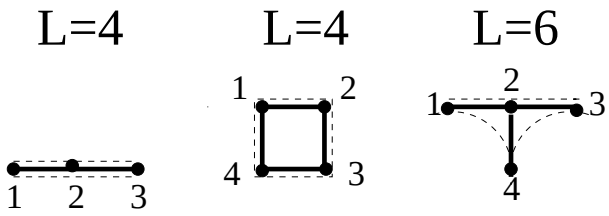


FIG. 4: Examples of contour paths for three different graphs. Spins are labelled by $1, 2, 3, 4$ and first-neighbor interactions are represented by bold lines. The contour path is depicted with a dotted line. The contour length L, which can be calculated as the sum of distances along the contour path (dotted arcs have length 2) is indicated above each graph. Different clusters may have the same contour path and contour length. Left: $(1, 3)$ and $(1, 2, 3)$ have contour length $L = 4$, while $(1, 2)$ has $L = 2$. Middle: clusters $(1, 3)$, $(1, 2, 3)$, $(1, 2, 3, 4)$, $(1, 3, 4)$ have the same contour path and contour length. Right: $(1, 2, 3, 4)$ has length $L = 6$.

Using the fact that the diagonal elements of M are equal to unity, the term corresponding to $K = 1$ above vanishes. For $K \geq 2$, we have

$$\begin{aligned} -\mathrm{Trace}\big[\big(M - \mathrm{Id}\big)^K\big] &= -\sum_{i_1,i_2,\ldots,i_K} \big(\delta_{i_1,i_2} - M_{i_1,i_2}\big)\big(\delta_{i_2,i_3} - M_{i_2,i_3}\big)\ldots\big(\delta_{i_{K-1},i_K} - M_{i_{K-1},i_K}\big)\big(\delta_{i_K,i_1} - M_{i_K,i_1}\big) \\ &= \sum_{i_1,i_2,\ldots,i_K} (-1)^{K-1}\, \hat{M}_{i_1,i_2}\hat{M}_{i_2,i_3}\ldots\hat{M}_{i_{K-1},i_K}\hat{M}_{i_K,i_1}\ , \end{aligned} \tag{40}$$

where the matrix $\hat{M}$ has the same off-diagonal elements as M, and has zero diagonal elements. Each term in the above sum corresponds to an Eulerian circuit over $K' \leq K$ sites, where K' is the number of distinct indices in $(i_1, i_2, \ldots, i_K)$. Note that the same circuit can be obtained from different K-uplets of indices. Consider for instance the longest circuits, obtained for $K' = K$, *i.e.* all distinct indices. $2K$ different K–uplets $(i_1, i_2, \ldots, i_K)$ correspond to the same circuit, as neither the starting site nor the orientation of the loop matter. For instance, $i_1 \to i_2 \to i_3 \to i_1$, $i_2 \to i_3 \to i_1 \to i_2$, $i_1 \to i_3 \to i_2 \to i_1$, ... are all equivalent. This multiplicity factor $2K$ precisely cancels the $2K$ at the denominator in (39). The contribution corresponding to a circuit therefore coincides with expression (38) for the loop entropy. We conclude that

- $S_{MF}(\mathbf{p})$ sums up all loop diagrams exactly;
- $S_{MF}(\mathbf{p})$, in addition, sums up Eulerian circuit diagrams, but with weights *a priori* different from their values in the cross-entropy $S(\mathbf{p})$ [53]. An exception is the three-variable Eulerian diagram shown in Fig. 3, whose weights in S_{MF} and S coincide.
- no non-Eulerian diagram is taken into account in $S_{MF}(\mathbf{p})$.

As a conclusion, the diagrammatic expansion provides a natural justification for the choice of the reference entropy $S_0(\mathbf{p}) = S_{MF}(\mathbf{p})$. In addition, it provides us with the dominant contribution to the cluster-entropies once the Mean-Field entropy is substracted, see Fig. 3. A detailed study of those dominant contributions is presented in Appendix C.

2. Dependence on the cluster size and on the interaction path length

To reach a better understanding of what the cluster-entropy means, we consider the case of finite-dimensional Ising model, *e.g.* with coupling $J > 0$ between nearest-neighbors on a D-dimensional lattice. We call ξ the correlation length: the connected correlation c between two sites at large distance d decays as $\sim \exp(-d/\xi)$. We want to characterize the behavior of the cluster-entropy ΔS_Γ when the K sites in the cluster Γ are far apart on the lattice. We first choose no reference entropy ($S_0 = 0$). According to the above diagrammatic expansion, the lowest order diagram (in powers of c) with K sites has the loop topology. We look for the shortest closed path joining all the sites in Γ; let $L(\Gamma)$ be this contour length, that is, the sum of the distances between neighboring sites along the path (Fig. 4). Then, according to (38), the largest contribution (in absolute value) to the cluster entropy is

$$\Delta S_\Gamma \simeq A(p, K)\ (-1)^{K-1}\ \exp\big(-L(\Gamma)/\xi\big)\ , \tag{41}$$

where A is a positive function K and of p, the representative value of the frequencies p_i of the variables in Γ. We conclude that the sign of the cluster-entropy depends on the parity of the number of sites. Furthermore, ΔS_Γ decreases

exponentially fast (in absolute value) with the length of the shortest path joining the sites in the cluster. As soon as one site is very far away from the remaining $K-1$ ones, the cluster-entropy is small.

As a consequence, the sum (27) is alternate, and we expect cancellation between contributions coming from clusters sharing the same shortest path, but with different sizes. This crucial point is perfectly illustrated by the one-dimensional Ising model. The correlation between two sites at distance $d_{ij} = j - i$ is, in one dimension, $c_{ij} = \sqrt{p_i(1-p_i)p_j(1-p_j)}\exp(-d_{ij}/\xi)$ (Appendix F). The matrix $\boldsymbol{M}$ defined in (21) has elements

$$M_{ij} = e^{-d_{ij}/\xi} \tag{42}$$

Then, according to (38), the largest contribution (in absolute value) to the cluster entropy of a cluster containing the K spins $i_1 < i_2 < \ldots < i_K$ is given by (41) with

$$L\big(\Gamma = (i_1, i_2, \ldots, i_k)\big) = 2(i_k - i_1) \ , \tag{43}$$

and $A(p,K) = \frac{1}{2}$. An exact calculation, reported in Appendix F, shows that

$$\Delta S_{(i_1,i_2,\ldots,i_K)} = (-1)^K \, F\left(\exp\Big(-\frac{i_k - i_1}{\xi}\Big)\right) \ , \tag{44}$$

where F is a smooth function given in (F11), such that $F(0) = F'(0) = 0, F''(0) = -1$. This identity is in agreement with (41), since the shortest path encircling all sites has length $L(\Gamma) = 2(i_K - i_1)$. Hence, all clusters sharing the same 'extremities, i.e. the same values of i_1 and i_K, have the same entropies in absolute value. The sign is determined by the parity of K as mentioned above. Let $i_K - i_1 \equiv d$. $\Gamma = (i_1, i_K)$ is the unique cluster of size $K = 2$ having its 'extremities' equal to i_1 and i_K; its entropy is $\Delta S^*_{(i_1,i_K)} = F(\exp(-d/\xi))$. There is $(d-1)$ clusters of size $K = 3$ with the same extremities, each having an entropy equal to $-\Delta S^*_{(i_1,i_K)}$. More generally, there are $\binom{d-1}{K-2}$ clusters of size K with the same extremities, each having an entropy equal to $(-1)^{K-2}\Delta S^*_{(i_1,i_K)}$. The total contribution to the entropy of all those clusters (at fixed extremities i_1, i_k) is

$$\Delta S_{\text{fixed } i_1,i_k} = \sum_{K=2}^{d+1} (-1)^{K-2} \binom{d-1}{K-2} \Delta S^*_{(i_1,i_K)} = (1-1)^{d-1} \, \Delta S^*_{(i_1,i_K)} = \begin{cases} \Delta S^*_{(i_1,i_K)} & \text{if } \ d = 1 \ , \\ 0 & \text{if } \ d \geq 2 \ . \end{cases} \tag{45}$$

The above calculation nicely exemplifies the cancellation of cluster-entropies. The contributions of all clusters sharing the same extremities exactly compensate each other, unless those extremities are nearest-neighbors on the lattice. We show in Appendix F that this exact cancellation is a consequence of the existence of a unique interaction path along the unidimensional chain. As a result, in dimension $D = 1$, the cross-entropy S is simply the sum of the entropies of the clusters made of nearest neighbours.

In the presence of a reference entropy, $S_0 = S_{MF}$, the asymptotic scaling of the cluster-entropy with its contour length L changes, as the dominant contribution coming from loop diagrams is removed from the cluster expansion and absorbed into S_0. The subleading contribution to the cluster-entropies is depicted in bold in Fig. 3 and derived in Appendix C. In dimension $D = 1$, formula (41) is replaced with

$$\Delta S_{(i_1,i_2,\ldots,i_K)} = A'(p,K) \ (-1)^{K-1} \ \exp\big(-3(i_k - i_1)/\xi\big) \ . \tag{46}$$

Note the sharper asymptotics decay with the distance between the extremities of Γ than in the absence of reference entropy. As expected, the terms in the expansion of $S - S_0$ are smaller than the one in the expansion of S alone. Remarkably, the exact cancellation property studied above also holds when the reference entropy is non-zero, as proven in Appendix F.

In dimension $D = 2$ or higher, more than one interaction path connect any two spins, and cluster-entropies with the same contour path do not cancel exactly as in the $D = 1$ case. However, partial cancellations are present. Figure 5 shows the values of the cluster-entropies versus the length of the shortest path, $L(\Gamma)$, for a small bidimensional 3×3 grid. For such a small system all data $\mathbf{p}$ and cluster-entropies ΔS_Γ (with up to $K = 9$ spins) can be calculated by exact enumeration methods. We observe that:

- $|\Delta S_\Gamma|$ is sensitive to the value of L_Γ more than to the size K of the cluster;
- $|\Delta S_\Gamma|$ rapidly decreases with L_Γ;
- the values of the cluster-entropies reflect the structural properties of the lattice, *e.g.* clusters made of central sites, such as 4-5, have a larger entropy than the clusters including pairs of edge spins, such as 1-2;

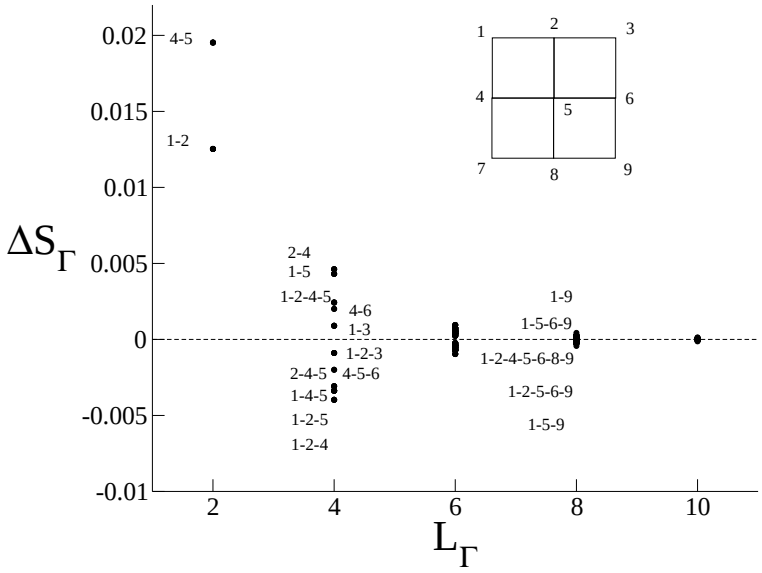


FIG. 5: Cluster entropy contribution ΔS_Γ for a 3×3 grid (top-right) with nearest-neighbour couplings $J = 1.777$ (in units of $k_B T$) as a function of the contour length L_Γ of the shortest closed path on the lattice joining the spins in Γ. To illustrate cancellation effects, some ΔS_Γ are labelled with the indices in Γ, see main text. The values of J and of the fields $h_i = -\frac{1}{2}\sum_{j(\neq i)} J_{ij}$ [49] are chosen to make the system critical in the infinite grid size limit, see Section VI C.

- the sign of ΔS_Γ changes with the parity of the size of the cluster.

As a result, the contributions to the entropies coming from the clusters sharing the same path, of length L, partially cancel each other. Consider for example the path 1-2-4-5 of length $L = 4$; all 7 clusters that share this path have similar $|\Delta S|$, ranging between 0.0024 and 0.0046, and so does their sum, $\Delta S_{(2,4)} + \Delta S_{(1,5)} + \Delta S_{(2,4,5)} + \Delta S_{(1,4,5)} + \Delta S_{(1,2,5)} + \Delta S_{(1,2,4)} + \Delta S_{(1,2,4,5)} = -0.00242$ [54]. The sum of the entropies of the clusters sharing the same path is generally of the same order of magnitude as, or even smaller than the single contributions. Figure 6 shows that the sum of the 12 clusters of contour length $L = 2$ and of the 4 square–path contributions ($|\Delta S| \geq .0024$) approximates the entropy within 10^{-6}.

IV. TRUNCATION OF THE CLUSTER EXPANSION

In this Section we present a truncation scheme for the cluster expansion, which consists in discarding all clusters with entropies smaller than a threshold Θ. We explain why this scheme is efficient, in particular in the presence of sampling noise, and robust against strong correlations in the data (large correlation length). The behavior of the expansion as a function of the threshold is discussed.

A. Schemes for truncating the expansion in the noiseless case

Expansion (27) for $S(\mathbf{p})$ includes $2^N - 1$ terms, and is useless unless an accurate truncation scheme is available. A naive truncation consists in keeping the contributions from the clusters with $\leq K$ spins, where K is an arbitrary size. This procedure was applied to neurobiological data (with $N \leq 40$, $K = 7$) in [10], which are characterized by large negative fields. However it suffers from two drawbacks. First, the combinatorial growth of the number of clusters with N and K impedes its application to very large systems. Secondly, the truncation does not converge properly with increasing K if the correlation length of the system is large.

As an illustration, consider again the 1D-ferromagnetic Ising model, with correlation length ξ. The sign of ΔS_Γ alternates with the parity of the size K of Γ; its modulus decays asymptotically as $\exp(-\Omega d/\xi)$, where d is the maximal distance between any two spins in Γ (Section III C 2), and $\Omega = 2$ if there is no reference entropy ($S_0 = 0$), $\Omega = 3$ if $S_0 = S_{MF}$. Let $\Delta S(K)$ be the sum of ΔS_Γ over all the clusters Γ with K spins. In the thermodynamic limit

$(N \to \infty)$,

$$\frac{1}{N}\,\Delta S(K) \sim (-1)^{K-1} \sum_{d \geq K-1} \binom{d-1}{K-2} \exp\big(-\,\Omega\, d/\xi\big) = \frac{(-1)^{K-1}}{\big(\exp(\Omega/\xi)-1\big)^{K-1}}\ . \tag{47}$$

Consider then the series summing all $\frac{1}{N}\Delta S(K)$ with $K \geq 2$. The series is convergent if $\xi < \xi_c = \frac{\Omega}{\log 2}$, and divergent when $\xi > \xi_c$. In the latter case, for a finite–N system, the maximum of $|\Delta S(K)|$ is exponentially large in N, and is reached in $K \simeq \frac{N}{2}$. As a consequence, for $\xi > \xi_c$, the sum (27) can not be truncated according to the size of the clusters. This result is not specific to the dimension unity, and holds for other interaction networks. The expansion of $S(\mathbf{p})$ defines an alternate series, and the order of its terms matters for its convergence in the $N \to \infty$ limit. For an Ising model on a generic lattice with fixed degree (number of neighbours) v, the largest value of ξ such that the series (27) (after division by N) is absolutely convergent in the $N \to \infty$ limit is $\xi_c = \frac{\Omega}{\log v}$ (Appendix D).

A better truncation scheme consists in keeping cluster-entropies larger than a threshold Θ only. Let us define

$$S(\mathbf{p},\Theta) = \sum_{\substack{\Gamma \subset (1,2,\ldots,N) \\ |\Delta S_\Gamma(\mathbf{p})| > \Theta}} \Delta S_\Gamma(\mathbf{p})\ . \tag{48}$$

The rationale is that, due to the properties of the cluster entropies and to the cancellation mechanism exposed in Section III C 2, summing large cluster-entropies may provide a good approximation to the true value of $S(\mathbf{p})$. In the $D = 1$ Ising model case, the exact value of $S(\mathbf{p})$ is, indeed, obtained as soon as $\Theta < \Delta S_{(1,2)}$. We show in Fig. 6 the residual error in the cross-entropy due to the truncation as a function of the threshold Θ for the same small $D = 2$ grid as in Fig. 5. The error $S(\mathbf{p},\Theta) - S(\mathbf{p})$ is very small, and equal to 10^{-6} when all clusters with contour length smaller than 4 are taken into account. As Θ is made smaller, clusters with larger contour lengths are summed up, and the error reaches the numerical accuracy $\sim 10^{-14}$. On top of this trend, positive fluctuations, corresponding to larger errors, arise when not all the clusters with the same interaction path (and length L) are summed up, and the cancellation of those contributions is not effective (Fig. 6 and caption). We will study in more details this phenomenon in Section IV D.

We now explain why the presence of noise in the data provides a compelling argument supporting the introduction of the cut-off Θ.

B. Distribution of small cluster-entropies in the presence of noisy data

In this Section, we investigate how limited sampling affects the values of the cluster-entropies. We assume that B configurations $\boldsymbol{\sigma}^\tau$ are sampled from the Gibbs distribution of an Ising model with interaction parameters $\mathbf{J}$ using a Monte Carlo procedure to generate the data $\mathbf{p}$.

1. *Universality at small $|\Delta S|$: numerical evidence*

The empirical correlations, $c_{ij} = p_{ij} - p_i p_j$, differ from the Gibbs correlations, $\langle\sigma_i\sigma_j\rangle_{\mathbf{J}} - \langle\sigma_i\rangle_{\mathbf{J}}\langle\sigma_j\rangle_{\mathbf{J}}$, by random fluctuations of amplitude

$$c_B \simeq \frac{p(1-p)}{\sqrt{B}}\ , \tag{49}$$

where p is the typical value of the p_i. For pairs i,j with weak Gibbs correlations ($< c_B$ in absolute value), the experimental correlations are dominated by the noise. As a consequence, the distribution of the cluster-entropies is universal for small ΔS. Its structure is a consequence of the noise in the data, and not of the interaction network of the model used to generate the data.

Figure 7 shows the histograms H (full distributions) of the entropies $\Delta S_{(i,j,k)}$ for the $K = 3$–clusters for a one-dimensional Ising model, for three values of the numbers B of sampled configurations. The histograms are made of two components: a bell-shaped distribution at small $|\Delta S|$, and isolated peaks at larger $|\Delta S|$. The cluster-entropies corresponding to the isolated peaks have the same values as in the perfect sampling case ($B = \infty$, impulses). When B increases, the bell shapes move towards smaller entropies (in the log-scale of Fig. 7), and more peaks are unveiled in H.

We show also in Fig. 7 the histograms H_{IS} for a system of Independent Spins (IS), with the same p_i's as the original system, and the same number B of sampled configurations. Contrary to H, H_{IS} does not exhibit isolated

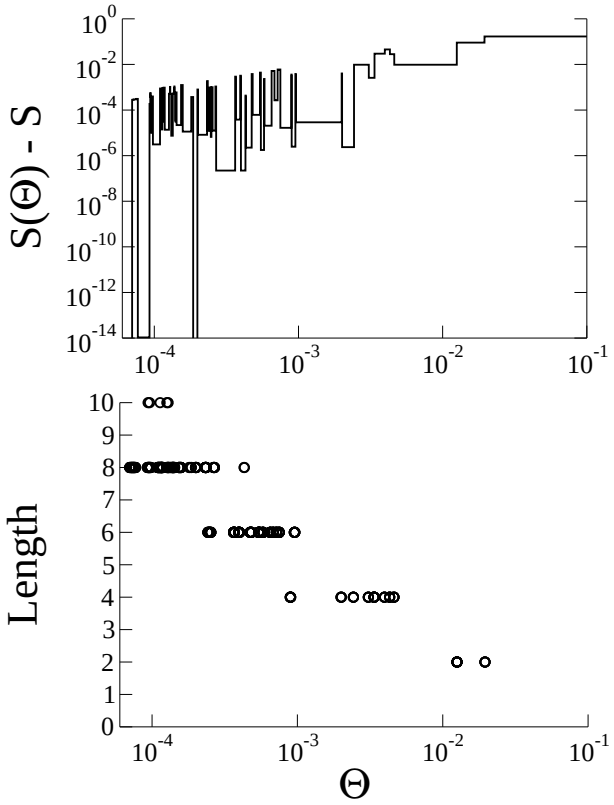


FIG. 6: Effect of truncation on the bidimensional 3×3 grid with $J = 1.777$ (units of k_BT) and fields $h_i = -\sum_{j(\neq i)} J_{ij}$ [49], for these parameters values the model has a phase transition between the paramagnetic and ferromagnetic phase and therefore the correlation length is proportional to the linear size of the system. Top: difference between the truncated and the true cross-entropies as a function of the cut-off on the absolute cluster-entropies, Θ. Bottom: contour lengths $L(\Gamma)$ vs. Θ. The fluctuations of $S(\Theta) - S$ reflect the cancellation phenomenon. Summation of the 12 clusters of nearest-neighbours with $K = 2$ and $L = 2$ gives $S(\Theta = 0.1) - S \simeq 0.01$, of the 21 clusters contributions corresponding to squared paths, *e.g.* 1-2-4-5), with $K = 2, 3, 4$ and $L = 4$ gives $S(\Theta = 0.002) - S \simeq 10^{-6}$. Fluctuations arise if only a part of the clusters that share the same interaction path are summed up, and cancellation is incomplete. For instance, fixing $\Theta = 0.0025$ discards $(1, 2, 4, 5)$, which has the same interaction path as $(2, 4)$.

peaks at well-defined, B–independent cluster-entropies. The histograms H_{IS} concentrate around smaller $|\Delta S|$ as the number B of configurations increases. Note that the histograms H_{IS} roughly correspond to the bell-shape parts of the distributions H for the same value of B. We have checked that these features are largely independent of the particular sample and of the cluster size, K.

The histograms H_{IS} depend on B through their standard deviation, $\sigma_{IS}(B)$. The calculation of $\sigma_{IS}(B)$ from the dominant contribution (C7) in the diagrammatic expansion of the cluster entropies (Section III C 1) is presented in Appendix E. We obtain that, for clusters of size K and in the case of uniform averages $p_i = p$ different from 0, $\frac{1}{2}$, and 1 [55],

$$\sigma_{IS}(B) \simeq \sqrt{\frac{3^K K!}{8}} \frac{(2p-1)^2}{p(1-p)} \Big(\frac{1}{B}\Big)^{K-\frac{1}{2}} \ . \tag{50}$$

Figure 8 shows how the small-entropy regions of the histograms H obtained for different B collapse onto each other once rescaled by $\sigma_{IS}(B)$. As expected, the rescaled H coincide with H_{IS} in the small $|\Delta S| \leq \sigma_{IS}(B)$ region, which concentrates most of the distribution (Fig. 8). The universality of the distribution at small ΔS is not specific to the one-dimensional Ising model, but holds, in the thermodynamic limit, for all interacting spin systems when the measured connected correlations are corrupted by noise. For a finite system in dimension D with correlation length ξ, we expect that the small-ΔS is universal when $N > \ell^D$, where $\ell = \xi \ \log(1/c_B)$. Indeed, the number of large c_{ij} coming out of the noisy background is $\approx N\,\ell^D$, while, for most of the $\binom{N}{2}$ pairs of spins i, j, the connected correlations have random values of amplitude c_B.

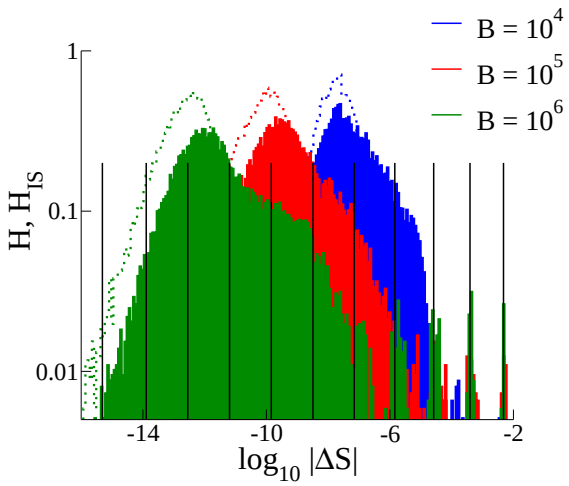


FIG. 7: Histograms of $\Delta S_{(i,j,k)}$ for the 1D-Ising (full distributions, $p = .02$, $\xi = 1$) and Independent Spin (dotted distribution) models, with $N = 50$ spins. Each histogram correspond to one random sample of B configurations. Impulses show the histogram for perfect sampling ($B = \infty$), with Dirac peaks located at $\log|\Delta S_\Gamma| = -\frac{3d_\Gamma}{\xi} + \text{Cst}$. The cut-off at small entropies comes from the finite value of N.

The full distribution H_{IS} can be characterized analytically in the $N \to \infty$ limit. Details can be found in Appendix E. We find the following scalings, depending on the value of the cluster size, K:

$$\begin{aligned} H_{IS}(\Delta S) \;&\sim\; \frac{1}{|\Delta S|^{2/3}} \text{ for } K = 2\,, \; \frac{(-\log \Delta S)^{K-2}}{\sqrt{|\Delta S|}} \text{ for } K \geq 3 \qquad (|\Delta S| \to 0)\,, \\ &\sim\; \exp\Big(-C_2(B)\,|\Delta S|^{2/(2K-1)}(1+o(1))\Big) \text{ for every } K \geq 2 \qquad (\text{large } |\Delta S|)\,, \end{aligned} \tag{51}$$

where $C_2(B) = 2{\times}3^{(K-1)/(2K-1)}\; K^{(2K-3)/(2K-1)}/(2K-1)^2\; (\sigma_{\Delta S})^{-2/(2K-1)}$ is proportional to B, see Appendix E and equation (E21). The distribution is therefore characterized by a divergence at the origin, and stretched exponential tails. The scalings above were derived with the choice $S_0 = S_{MF}$; in the absence of the reference entropy, the stretched exponential has exponent $\frac{2}{K}$ instead of $\frac{2}{2K-1}$.

2. *Finite–N effects and lower bound to the threshold* Θ

The discussion about the localized peaks and the bell-shape distribution in H_{IS} in the previous Section is an oversimplification. In reality, for finite systems, large fluctuations of the sampled correlations take place, and no clear-cut boundary exist between cluster-entropies due to the noise and the ones deriving from the interaction network. From extreme value theory [38], the largest value of the correlations are of the order of $c_{ij}^{MAX} = c_B\sqrt{4\log N}$. Therefore, the largest cluster-entropy is, according to (C7), of the order of

$$\Delta S^{max} \approx (4\log N)^{(2K-1)/2}\sigma_{IS}\,. \tag{52}$$

A more detailed calculation to estimate where this fuzzy boundary between the signal and the noise in the entropy distribution takes place is presented below. Let $M_K(\Theta)$ be the average number of clusters of size K with entropies $|\Delta S| > \Theta$. According to (51),

$$M(\Theta) = \binom{N}{K}\int_\Theta d(\Delta S)\, H_{IS}(\Delta S) \simeq \exp\Big(-C_2(B)\,\Theta^{2/(2K-1)}\Big)\,\frac{(2K-1)\,N^K\,\Theta^{(2K-3)/(2K-1)}}{2\,K!\,C_2(B)}\,, \tag{53}$$

for large Θ and N (compared to K). The value of the threshold Θ such that $M(\Theta) = N^\alpha$, with $\alpha < K$, is, to the leading order in N,

$$\Theta(\alpha) \simeq \left(\frac{K-\alpha}{C_2(B)}\,\log N\right)^{K-\frac{1}{2}}\,. \tag{54}$$

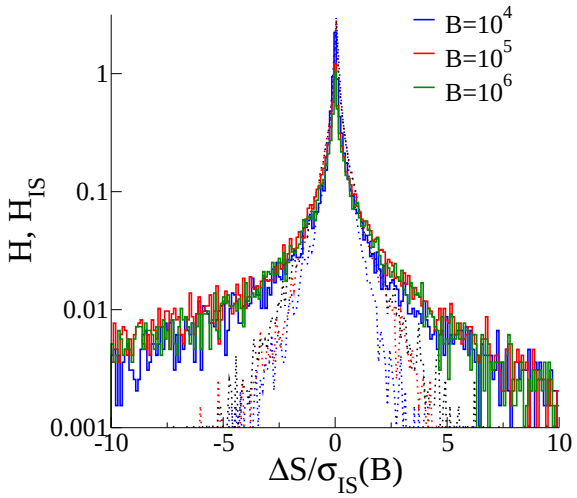


FIG. 8: Same as Fig. 7, after rescaling by the standard deviation $\sigma_{IS}(B)(50)$ of the Independent Spin model. Note the linear scale of the x-axis. As a result of the presence of the interaction network, the Ising histograms H are asymmetric in $\Delta S \to -\Delta S$ for large values of ΔS, while the IS distributions H_{IS} are obviously symmetric when averaged over the realizations of B configurations (not shown).

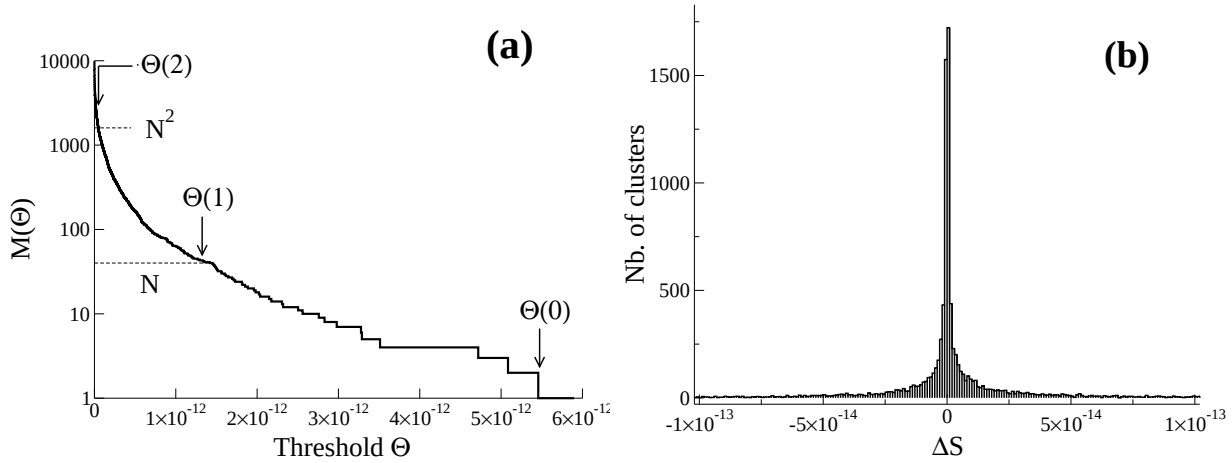


FIG. 9: **(a)** Number M of clusters (i,j,k) with $|\Delta S_{(i,j,k}| > \Theta$ as a function of the threshold Θ, for one realization of $B = 10^6$ configurations ($N = 40$ independent spins with $p = .0248$). The theoretical values for the threshold, $\Theta(2) = 5\,10^{-14}, \Theta(1) = 1.3\,10^{-12}, \Theta(0) = 5.5\,10^{-12}$, corresponding to, respectively, $M = N^2, N, 1$, are shown. **(b)** Number of clusters as a function of their entropy ΔS. Same data as in **(a)**, on a smaller entropy scale.

In particular, using the formula above for $\alpha = 0$, it is likely that no cluster have entropy larger than $\Theta(0)$, in agreement with (52).

We have tested formula (54) through a computation based of a system of $N = 40$ Independent Spins, with uniform mean $p = .0248$; these parameters were chosen to mimick real data described in [2]. Figure 9(a) shows the number of clusters with entropies larger than Θ in absolute value, for a random set of $B = 10^6$ configurations ($K = 3$). The theoretical predictions based on (54) are in very good agreement with the simulations. The vast majority of clusters have entropies smaller than, say, $\Theta(2)$. On a smaller entropy scale, the histogram H_{IS} of the small cluster entropies is strongly concentratred around zero as predicted in 51 (Fig. 9(b)).

As a conclusion, due to the sampling noise, most small cluster-entropies are random quantities, and provide no information about the underlying interactions parameters. Imposing a threshold Θ allows one to remove these artifact contributions. A lower bound to the value of Θ is given by (54), with, say $\alpha = 1$ or 2. In practice, we will see that higher values of Θ may be sufficient for an accurate solution of the inverse Ising problem.

C. Properties of the susceptibility matrix and of its inverse

We now present a theoretical argument suggesting that the truncation scheme we have introduced is robust against an increase of the correlation length of the system. More precisely, the maximal size of the clusters to be summed up to reach an accurate solution of the inverse problem is not directly related to the correlation length, but rather depends on the structure of the interaction graph.

The susceptibility matrix $\boldsymbol{\chi}$ (10) characterizes how the observables of the Ising model, such as the averages and correlations in $\mathbf{p}$, are modified in response to an infinitesimal change in one or more interaction parameters in $\mathbf{J}$. As far as the inverse Ising problem is concerned, it is more natural to ask the following question. Assume the inverse problem has been solved for a set of data $\mathbf{p}$ and the corresponding interations $\mathbf{J}$ have been found. Now imagine that the data are slightly changed, $\mathbf{p} \to \mathbf{p} + \boldsymbol{\delta}\mathbf{p}$. How large will be the resulting change $\boldsymbol{\delta}\mathbf{J}$ in the interactions? The response function characterizing the inverse problem,

$$\frac{\boldsymbol{\delta}\mathbf{J}}{\boldsymbol{\delta}\mathbf{p}} = -\frac{\partial^2 S(\mathbf{p})}{\partial \mathbf{p} \partial \mathbf{p}} = \boldsymbol{\chi}^{-1} \ , \tag{55}$$

is simply the inverse of the susceptibility matrix $\boldsymbol{\chi}$. Whether the inverse problem is well-behaved or not will therefore depend on the properties of $\boldsymbol{\chi}^{-1}$. In particular, it will depend on the largest eigenvalues of $\boldsymbol{\chi}^{-1}$ and on the structure of the corresponding eigenvectors.

A quantity which is closely related to (55) in liquid theory is the Ornstein-Zernike direct correlation function. The direct correlation is widely believed to be short–ranged, as the interaction potential [46]. This property is used in closure schemes such as the Percus-Yevick scheme to obtain the equation of state [42]. We discuss below in details the property of the inverse susceptibility matrix in the case of the spherical model and of the unidimensional Ising model.

1. *Case of perfect sampling: properties of $\boldsymbol{\chi}$ and $\boldsymbol{\chi}^{-1}$*

Consider first the $O(m)$ model, where each site $i = 1, 2, \ldots, N$ carries a m-dimensional real-valued spin vector $\boldsymbol{\sigma}_i = (\sigma_i^1, \sigma_i^2, \ldots, \sigma_i^m)$, of norm $\sqrt{m}$. As usual, two spins, say, i and j, are coupled through the dot product of their spin vectors, $-J_{ij}\, \boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_i$ (units of $k_B T$). Hence the interaction J_{ij} couples the same component (α) of the spins in the pair i, j. The fields h_i^α, with $\alpha = 1, 2, \ldots, m$, are chosen to vanish for simplicity. In the large–m limit the model can be exactly solved [40]. The cross-entropy is equal to

$$S(\mathbf{p}) = \frac{m}{2}\ \log\det\ \hat{\mathbf{p}} + O\big(\log m\big) \ , \tag{56}$$

where $\hat{\mathbf{p}}$ is the $N \times N$ matrix with diagonal elements $\hat{p}_{ii} = 1$ and off-diagonal elements $\hat{p}_{ij}$, equal to the average of the product of the components α of spins i and j. The elements of the inverse susceptibility matrix are obtained by differentiating $S(\mathbf{p})$ twice with respect to $\hat{\mathbf{p}}$,

$$\big(\chi^{-1}\big)_{kl,k'l'} = \frac{1}{2}\big(J_{k,k'} J_{l,l'} + J_{k,l'} J_{l,k'}\big) \ . \tag{57}$$

Hence, the inverse susceptibility has the same structure as the interaction graph. In particular, if the coupling matrix $\mathbf{J}$ is sparse (has many zero elements), so is $\boldsymbol{\chi}^{-1}$. On the contrary, the susceptibility matrix $\boldsymbol{\chi}$ is generally not sparse.

The observation above is not specific to spherical spins. Consider now the D–dimensional Ising model with $\sigma_i = 0, 1$ spins on a hypercubic lattice, with nearest neighbour interactions J_{ij}. In the $D = 1$ case the susceptibility matrix (top right corner in matrix (10)) is non zero for all i, i': $\chi_{i,i} \propto x^{|i-i'|}$, where the proportionality constant does not depend on i, i', and $x = \exp(-1/\xi)$. The inverse susceptibility matrix is a tridiagonal matrix [43]: the only non-zero elements are

$$(\chi^{-1})_{ii} = \frac{1+x^2}{1-x^2} \text{ and } (\chi^{-1})_{i,i\pm1} = -\frac{x}{1-x^2} \ . \tag{58}$$

As in the spherical model, the structure of the inverse susceptibility matrix is the same as the one of the interaction matrix. In dimension $D \geq 2$ the inverse susceptibility matrix is not, strictly speaking, sparse. However it exhibits a much faster decay with the distance $r = |i - i'|$ than the susceptibility itself [56]. At the critical point, the latter decays as $\chi(r) \sim r^{-(D-2+\eta)}$, where the critical exponent attached to the decay of the spin-spin correlation, η, vanishes in dimension $D \geq 4$, and is positive and small in dimension $D \leq 3$, *i.e.* $\eta = \frac{1}{4}$ for $D = 2$. The inverse susceptibility is

the Laplacian in dimension $D \geq 4$, a purely local operator, and decays as $\chi^{-1}(r) \sim r^{-(D+2-\eta)}$ for $D \leq 3$. While both quantities decrease as power laws in r, the inverse susceptibility has a much sharper decay than the susceptibility itself. In particular, the integrated contribution to the susceptibility coming from distances larger than R,

$$\int_R^\infty d^D r\ \chi(r) = R^{1-\eta}\ , \tag{59}$$

diverges when $R \to \infty$, while the same quantity calculated for the inverse susceptibility,

$$\int_R^\infty d^D r\ \chi^{-1}(r) = \frac{1}{R^{3-\eta}}\ , \tag{60}$$

tends to zero as $R \to \infty$. This fact is a good news for the inverse problem. According to (55) the error on the field h_i done when discarding all the spins at distance $R > \epsilon^{-1/(3-\eta)}$ is of the order of ϵ only. In this regard, the inverse Ising problem remains local even at the critical point.

While the discussion above is related to the response of a field h_i to a change in the average $p_{i'}$ of spin i', the response of a coupling J_{kl} following a modification of the 2-point average $p_{k'l'}$, see (10), is also of interest. Unfortunately, to our best knowledge, this quantity has not been studied in the case of the Ising model so far. As a first step, we focus here on the $D = 1$-Ising model with uniform nearest-neighbour interactions, and in the thermodynamical limit ($N \to \infty$). The four-spin connected correlation function is, up to a p-dependent multiplicative constant, equal to

$$\chi_{ij,kl} = x^{i_4 - i_3 + i_2 - i_1} - x^{j-i+k-l}\ , \tag{61}$$

where $i_1 \leq i_2 \leq i_3 \leq i_4$ are the same indices as i, j, k, l but sorted in increasing order, and $x = \exp(-1/\xi) < 1$. We show in Appendix G that the inverse susceptibility matrix is given by

$$\left(\chi^{-1}\right)_{ij,kl} = \begin{cases} \frac{(1+x^2)^2}{(1-x^2)^2} & \text{if} \quad i = k, j = l \text{ and } j \geq i+2\ , \\ \frac{1+x^2+x^4}{(1-x^2)^2} & \text{if} \quad i = k, j = l \text{ and } j = i+1\ , \\ -\frac{x(1+x^2)}{(1-x^2)^2} & \text{if} \quad i = k \pm 1, j = l \text{ or } i = k, j = l \pm 1\ , \\ \frac{x^2}{(1-x^2)^2} & \text{if} \quad i = k \pm 1, j = l \pm 1\ , \\ 0 & \text{otherwise} \quad . \end{cases} \tag{62}$$

Hence, the inverse susceptibility matrix is sparse, with at most 9 non-zero elements per line, while the dimension of the matrix is $\frac{1}{2}N(N-1) \to \infty$. In dimension $D \geq 2$, we do not expect $\boldsymbol{\chi}^{-1}$ to be sparse. However we conjecture that $\left(\chi^{-1}\right)_{ij,kl}$ decays quickly with the minimal distance between the four points i, j, k, l (each index, *e.g.* j, is now a D-dimensional vector).

2. *Influence of the sampling noise on the norms of $\boldsymbol{\chi}$ and $\boldsymbol{\chi}^{-1}$*

To corroborate this statement we have carried out exact numerical analysis of small bidimensional grids (Section III C 2). We show in Fig. 10(a) the fraction of elements $\chi_{ij,kl}$ of the susceptibility matrix larger than $\epsilon = 10^{-7}$ in absolute value (the largest elements have magnitude ~ 1). This fraction is closed to 1 for all the values J of the coupling we have studied. As expected, the inverse susceptibility matrix has many more small elements (Fig. 10(b)). In addition, the fraction of entries in $\boldsymbol{\chi}^{-1}$ smaller than ϵ seem to increase with the size N of the grid.

In the presence of noise in the sampling process the inverse matrix $\boldsymbol{\chi}^{-1}$ loses its quasi-sparse structure. More precisely, for the number B of sampled configurations chosen in Fig. 10(b), all the elements $(\chi^{-1})_{ij,kl}$ are larger than ϵ in absolute value. Indeed, the quasi-sparsity $\boldsymbol{\chi}^{-1}$ in the perfect sampling case reflects the sparse structure of the underlying interaction matrix. When data are corrupted by noise, the Ising model (over)fitting the data has no reason to be sparse anymore, and neither has the inverse susceptibility.

The influence of the sampling noise on the susceptibility matrix and on its inverse can be measured through the largest and smallest eigenvalue of $\boldsymbol{\chi}$, denoted by, respectively, λ_{max} and λ_{min}. According to Figs. 11(a,b), we have that:

- λ_{max} increases with the size of the system (we expect λ_{max} to diverge at the critical coupling $J \simeq 1.778$ in the thermodynamical limit), but is not affected by the sampling noise (the black and red/gray curves associated to the same size are nearly indistinguishable in Fig. 11(a)).

- λ_{min} is not strongly affected by the system size in the case of perfect sampling. In case of noisy sampling, λ_{min} acquires a smaller value. The effect of the noise increases with the system size (Fig. 11(b)).

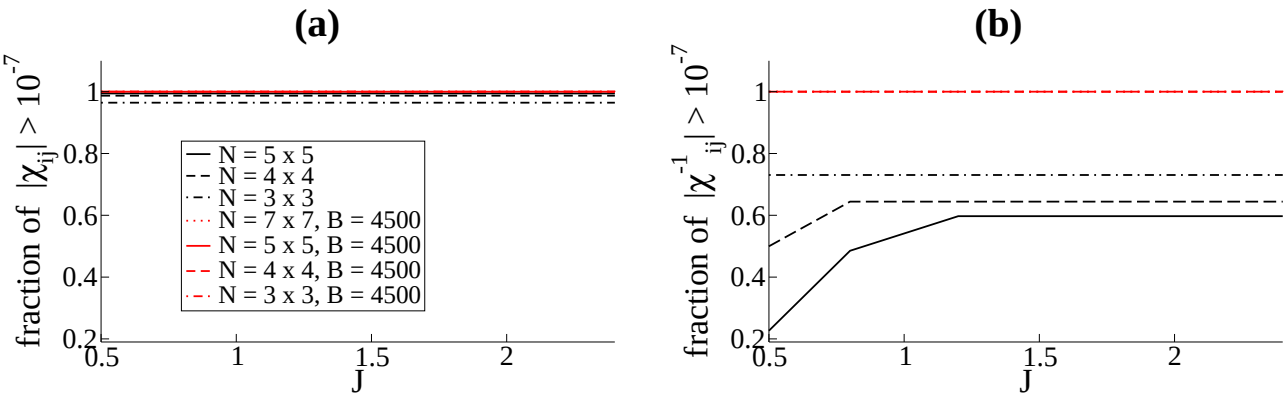


FIG. 10: Fraction of elements larger than 10^{-7} (in absolute value) for the susceptibility $\boldsymbol{\chi}$ **(a)** and the inverse susceptibility $\boldsymbol{\chi}^{-1}$ **(b)** matrices vs. strength J of the nearest-neighbour coupling. The sizes N of the grids are indicated. Data were obtained from exact numerations for sizes 3×3, 4×4, 5×5 (perfect sampling, black) and from Monte Carlo simulations for all sizes (one realization of $B = 4500$ configurations, red/gray). Periodic boundary conditions were used.

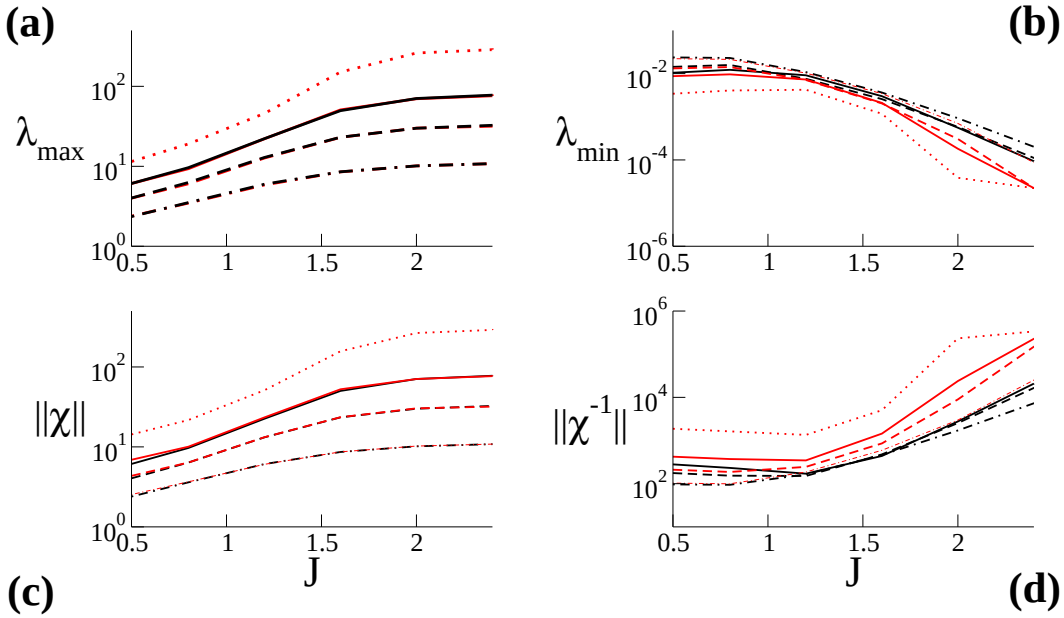


FIG. 11: Largest **(a)** and smallest **(b)** eigenvalues of the susceptibility matrix, and norms of $\boldsymbol{\chi}$ **(c)** and of its inverse $\boldsymbol{\chi}^{-1}$ **(d)** as functions of the coupling strength J for the 3×3 grid. See Fig. 10 for explanations regarding the color code and the line styles.

Those facts are observed from the study of the norms of the two matrices $\boldsymbol{\chi}$ and $\boldsymbol{\chi}^{-1}$. Here, we define the norm of the matrix A through

$$||A|| = \max_i \sum_j |A_{i,j}| \ . \tag{63}$$

Figures 11(c,d) show that the behaviours of the norms $||\boldsymbol{\chi}||$ and $||\boldsymbol{\chi}^{-1}||$ are, from a qualitative point of view, similar to the ones of, respectively, λ_{max} and $1/\lambda_{min}$. However the norms are directly related to the magnitudes of the elements of the matrices, according to (63). The independence of $||\boldsymbol{\chi}^{-1}||$ from the size N, contrary to the strong increase of $||\boldsymbol{\chi}||$, supports the notion that most elements of $\boldsymbol{\chi}^{-1}$ are very small (or even zero) in the case of perfect sampling. This property is lost when the sampling is not perfect: the presence of noise in the correlation makes the norm $||\boldsymbol{\chi}^{-1}||$ increases with N (Fig. 11(d)).

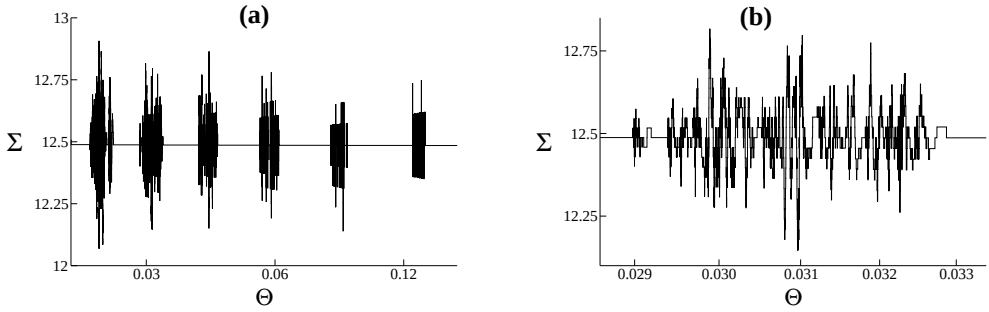


FIG. 12: **(a)** Sum Σ of the cluster-entropies larger than Θ (in absolute value) for the nearest neighbour one-dimensional Ising model with $\xi \simeq 8.97 \gg \xi_c \simeq 4.33$, and $B = 10^5$ configurations. The initial increase from zero, taking place at small $\Theta \simeq \Delta S(1) \simeq 0.41$, is not shown. **(b)** magnification of **(a)** in the range $.029 < \Theta < .033$. Within the random sign model, the behavior of Σ within a packet is similar to a Brownian bridge, see main text.

D. Dependence of the truncated entropy on the threshold

Hereafter, we study how the error on the entropy $S(\Theta)$ resulting from the truncation varies with the threshold Θ and we discuss the fluctuations of $S(\Theta) - S$ observed in Fig. 6. We start by sorting the absolute values of the cluster-entropies $|\Delta S_\Gamma|$ in decreasing order:

$$\Delta S_1 \geq \Delta S_2 \geq \Delta S_3 \geq \ldots \geq \Delta S_n \geq \ldots \geq 0 \ . \tag{64}$$

We call $\eta_n = \pm 1$ the sign of the cluster-entropy ΔS_Γ attached (equal in absolute value) to ΔS_n. Given the threshold Θ, we define $n^*(\Theta)$ as the index of the smallest cluster-entropy larger than Θ: $\Delta S_{n^*(\Theta)} \geq \Theta > \Delta S_{n^*(\Theta)+1}$. The truncated entropy (48) can be rewritten as $S(\mathbf{p}, \Theta) = \Sigma(\Theta)$, where

$$\Sigma(\Theta) = \sum_{n=1}^{n^*(\Theta)} \eta_n \ \Delta S_n \ . \tag{65}$$

We want to study how $\Sigma(\Theta)$ behaves when Θ is made small. In particular, how does the difference $\epsilon_s(\Theta) = \Sigma(\Theta) - \Sigma(0)$ behave as a function of Θ? Is it a smooth function, or does it exhibit large and irregular fluctuations? From a mathematical point of view, it is convenient to imagine that $N \to \infty$. The above question can be formalized as whether $\frac{1}{N}\Sigma(\Theta)$ converges to some limit value; the normalization factor comes from the fact that we expect the cross-entropy to be extensive in the system size N. Depending on the system under consideration, different situations can be encountered.

The most favorable case is when

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n\geq 1} \Delta S_n < \infty \ . \tag{66}$$

If this condition holds, the difference $\epsilon_s(\Theta)$ can be made arbitrarily small if Θ is small enough. An illustration is provided by the one-dimensional Ising model with small correlation length ξ and perfect sampling ($B = \infty$). For this model, the sequence of ΔS_n is highly degenerate, and its distinct values are in one-to-one correspondence with the integer distances $d \geq 1$ between the extremities of the clusters (Fig. 7). The cluster-entropy $\Delta S(d)$ asymptotically decays as $\exp(-3d/\xi)$, and has multiplity 2^{d-1}, since each point between the extremities may or may not belong to the cluster. We find

$$\frac{1}{N} \sum_{n\geq 1} \Delta S_n \simeq \sum_{d\geq 1} 2^{d-1} \ \exp(-3d/\xi) \ , \tag{67}$$

which converges if $\xi < \xi_c = \frac{3}{\log 2}$. The calculation above is very similar to the one of Section IV A. Indeed, when the series with general term $\Delta S(K)$ is absolutely convergent, any ordering of the cluster-entropies is possible. In particular, one is allowed to sum all the clusters of a given size K as proposed at the beginning of Section IV A.

What happens when condition (66) is violated? Again consider the one-dimensional Ising model. For perfect sampling, the cancellation property discussed in Section III C 2 ensures that $\frac{1}{N}\Sigma(\Theta)$ has reached its limit $\Delta S(d = 1)$

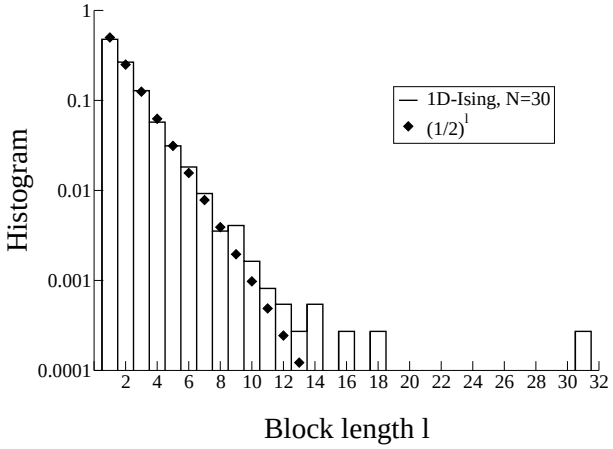


FIG. 13: Frequences of the block length ℓ for the 1D-Ising model, with $N = 30$ spins, $\xi = 8.96$, and one set of $B = 10^5$ sampled configurations. The statistics takes into account the cluster-entropies used to draw Fig. 12(a) only.

as soon as $\Theta < \Delta S(1)$. In the case of noisy sampling (finite B), the situation is more complex. In the presence of noise in the correlations c_{kl} the cluster-entropies with the same distance d between extremities are not degenerate any longer. We show in Fig. 12(a) the value of Σ as function of Θ for a large correlation length ξ compared to ξ_c, and $B = 10^5$ sampled configurations. We observe the appearance of 'packets' of cluster-entropies, located around the noiseless values $\Delta S(d \geq 2)$. The width of a packet depends on the amount of noise due to the sampling, *i.e.* on the number B of sampled configurations. The values of Σ at the two edges of the packet are very close to one another due to the cancellation property. As Θ spans the range of cluster-entropies in the packet, Σ fluctuates. The maximal amplitude of the fluctuations seems to weakly increase as we look at packets with smaller and smaller entropies (Fig. 12(a)).

We have analyzed the statistics of the signs ϵ_n of the cluters -entropies in (65). Writing the sequence of signs $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \ldots)$, we consider the blocks j of contiguous and equal signs, and defines their lengths ℓ_j. For instance, the block lengths corresponding to $\boldsymbol{\eta} = (+,+,-,+,+,+,+,-,-,-,-,-,+,-,....)$ are $\ell_1 = 2, \ell_2 = 1, \ell_3 = 4, \ell_4 = 5, \ell_5 = 1, ...$ The histogram of the block-lengths is shown in Fig. 13. The two main features are:

- The frequence of ℓ decreases exponentially when $\ell \ll N$, and is in very good agreement with the exponential law $\left(\frac{1}{2}\right)^{\ell}$.
- A large 'structural' block of length $\ell \simeq N$ is present. This block corresponds to the N clusters of size $K = 2$ (having all sign +), and the cluster of size $K = 3$ with largest entropy (which has the same sign +).

We have calculated the correlation between successive block lengths, normalized by the variance of the block length,

$$\rho = \frac{\overline{l_j l_{j+1}} - \overline{l_j}^2}{\overline{l_j^2} - \overline{l_j}^2} \ , \tag{68}$$

where $\overline{(\cdot)}$ denotes the average over the blocks j. For the model and the data shown in Figs. 12 and 13, we find $\rho \simeq .012$. Changing the set of sampled configurations does not affect the amplitude of the ratio ρ, which is always found to be about 1%. This ratio coincides with the inverse of the square root of the number of blocks, equal to a few thousands. Hence, the analysis is compatible with the absence of any correlation between the lengths of successive blocks. The same conclusion is reached with experimental data, *e.g.* multi-electrode recordings of the activity of a neural population [2, 37] (not shown).

The simple statistics sets above suggests the following 'random sign' model, allowing us to deepen our theoretical understanding of the behavior of $\Sigma(\Theta)$. In the random sign model, the signs η_n are replaced with random variables, equal to $\pm$ with probabilities $\frac{1}{2}$, and independent from each other. We emphasize that, in $\Sigma(\Theta)$ defined in (65), the signs are deterministic (for given data $\mathbf{p}$). The random sign model is therefore an approximation motivated by the statistical analysis above. Assume now that the value chosen for the threshold Θ falls within a packet p including $\mathcal{N}_p$ clusters. Fluctuations of the order of

$$\Delta\Sigma \sim \pm\Theta\sqrt{\mathcal{N}_p} \tag{69}$$

are expected on the entropy. As Θ decreases, the size of the packets, $\mathcal{N}_p$, tends to be bigger. Loosely speaking, smaller entropies correspond to longer interaction paths, shared by many more clusters. In the case of the one-dimensional Ising model, as Θ decreases, the distance between the extremities of the clusters involved in a packet, d, increases. We have $\mathcal{N}_p = 2^{d-1}$; hence, $\Delta\Sigma \sim \exp(-3d/\xi)\sqrt{2^d}$. We conclude that the error on the entropy tends to zero if $\xi < 2\,\xi_c$.

From the above discussion, it appears that a general, sufficient condition for the amplitude of the fluctuations to vanish as $\Theta \to 0$ is

$$\lim_{N\to\infty} \frac{1}{N} \sum_n (\Delta S_n)^2 < \infty \ . \tag{70}$$

Indeed, if condition (70) is fulfilled, the sum of the fluctuations due to *all* packets corresponding to cluster-entropies smaller than Θ is guaranteed to vanish with Θ. Hence condition (70) not only ensures that the fluctuations $\Delta\Sigma$ attached to the packet 'cut' by Θ vanishes, but also that the error on the entropy, $\epsilon_s(\Theta)$, tends to zero when $\Theta \to 0$. It is important to realize that the guarantee is of probabilistic nature. Arbitrary large fluctuations are possible (in the $N \to \infty$ limit), but are very unlikely. More precisely, within the random sign model, the error is a normal variable,

$$\epsilon_s(\Theta) = \mathcal{N}\left(0, \frac{1}{N} \sum_{n>n^*(\Theta)} (\Delta S_n)^2\right) , \tag{71}$$

with a variance vanishing with Θ according to (70). The true error is expected to be even smaller than the random sign estimate (71). Indeed, packets need not be isolated from each other as in Fig. 12. In the presence of a strong sampling noise, or in higher dimension than $D = 1$, packets will overlap. As a consequence, the number of packets 'cut' by the threshold and their size will determine the amplitude of $\Delta\Sigma$. Further investigations of those points are needed.

V. ADAPTIVE ALGORITHM FOR THE INVERSE ISING PROBLEM

A. Procedure to construct and select clusters

As explained above discarding the cluster-entropies smaller than a threshold Θ is an efficient step against overfitting of the sampling noise. In addition, for systems with dilute and strong interactions, we expect that only the clusters of neighboring sites on the interaction network will have substantial entropies. These arguments provide a heuristic basis for the threshold-based truncation of the expansion (27).

How can we implement the truncation scheme in practice? The combinatorial explosion of the number of clusters of size K among N sites impedes any brute force computation approach, as soon as N is larger than a few tens. Even for small–N system for which it is feasible, computing $\sim 2^N$ cluster-entropies and, then, discarding most of them does not sound like an efficient procedure.

We propose below an alternative approach, based on a recursive and selective construction of relevant clusters. The approach is based on the principle that clusters with large entropies should be compatible with the interaction network to be inferred. Suppose that two clusters Γ and Γ' have both large entropies, and share most of their spins. Then, the union $\Gamma \cup \Gamma'$ is a good candidate for a bigger cluster. If the entropy of the union cluster is large, a new part of the interaction network will be unveiled. Conversely, if it is small, no new interaction path with respect to the one discovered from Γ and Γ' separately exists. Hence, combining strongly overlapping clusters should allow us to progressively deepen our knowledge of the local structure of the interaction graph.

The above heuristics is formalized as follows:

A1. Initial step: build the list of all clusters of size one: $L_1 = \{(i) : i = 1, 2, \ldots, N\}$. All the other lists L_K for $K \geq 2$ are empty.

A2. Iteration: assume the current size of clusters is $K \geq 1$, *i.e.* L_K is not empty while L_{K+1} is empty. For every pair Γ_1, Γ_2 in L_K:

- **A21.** Construction: build $\Gamma = \Gamma_1 \cup \Gamma_2$
- **A22.** Selection: if Γ is of size $K+1$ and if $|\Delta S_\Gamma(\mathbf{p})| \geq \Theta$, then select Γ and add it to L_{K+1}.

A3. Recursion: if at least one cluster has been selected, then add 1 to K, and go to step 2 to pursue the construction process. If no cluster has been selected, the construction process is over.

The first condition in **A22** is about the size of Γ. The union of two clusters of size K has size $K+1$ if and only if they have exactly $K-1$ common spins. $\Gamma_1 = (i_1, i_2 \ldots i_{K-1}, x)$ and $\Gamma_2 = (i_1, i_2, \ldots, i_{K-1}, y)$ can be merged into $\Gamma = (i_1, i_2, \ldots, i_{K-1}, x, y)$; the ordering of x, y, and of the i_l's is irrelevant here.

B. Calculation of $\Delta S_\Gamma(\mathbf{p})$

Step **A22** requires the calculation of the cluster-entropy $\Delta S_\Gamma(\mathbf{p})$ for each selected cluster Γ (of size K). In order to do so we make use of the formula

$$\Delta S_\Gamma(\mathbf{p}) = S_\Gamma(\mathbf{p}) - (S_0)_\Gamma(\mathbf{p}) - \sum_{\substack{\Gamma' \subset \Gamma \\ (\Gamma' \neq \Gamma)}} \Delta S_{\Gamma'}(\mathbf{p}) \ , \tag{72}$$

which can be easily deduced from (27). The procedure is as follows:

B1. calculate the subset-entropy $S_\Gamma(\mathbf{p})$ through the minimization of $S_{Ising}(\mathbf{J}|\mathbf{p})$ (7) with respect to the fields and couplings. The partition function $Z[\mathbf{J}]$ is computed as the sum over the 2^K configurations of the spins in Γ.

B2. substract the reference entropy $(S_0)_\Gamma(\mathbf{p})$. For the mean-field reference entropy, $(S_0)_\Gamma(\mathbf{p}) = \frac{1}{2}\log\det M_\Gamma(\mathbf{p})$, according to formula (21); $M_\Gamma(\mathbf{p})$ is the $K \times K$ restriction of matrix $M(\mathbf{p})$ to the indices $i_1, i_2, \ldots, i_K$ in Γ. In presence of a regularization term (13) equation (36) has to be used instead of (21) to calculate $(S_0)_\Gamma(\mathbf{p})$.

B3. Substract the entropies $\Delta S_{\Gamma'}(\mathbf{p})$ of all the sub-clusters Γ' of size $K' < K$, included in Γ.

The last step (**B3**) assumes that the entropies of all the sub-clusters of Γ are known, *i.e.* have been computed at a previous step in the algorithm. This is true for $K' = 2$, but not necessarily so for $K' \geq 3$. To circumvent this difficulty we maintain at all times during the execution of the algorithm the list L_{all} of all the clusters and of their entropies calculated so far; L_{all} is a larger list than the one of the selected clusters (union of all L_K). The procedure to compute $\Delta S_\Gamma(\mathbf{p})$ is then:

B0. build the list $\hat{L}$ of all the sub-clusters Γ' in Γ *not* already present in L_{all}. For each $\Gamma' \in \hat{L}$, starting from the smallest sub-cluster and ending up with the largest one, run steps **B1**, **B2**, **B3** to obtain $\Delta S_{\Gamma'}(\mathbf{p})$, and add Γ' and its entropy to the list L_{all}.

The ordering of $\hat{L}$ ensures that all the sub-clusters of Γ' required to calculated its entropy are in L_{all} when step **B3** is executed.

C. Calculation of the cross-entropy, couplings and fields

Once the construction process is finished, the list $L_{sel} = L_1 \cup L_2 \cup L_3 \cup \ldots \cup L_{K_{max}}$ of all selected clusters is available. Here, K_{max} is the size of the largest cluster selected by the construction procedure. We then

C1. estimate the cross-entropy through

$$S(\mathbf{p}) = S_0(\mathbf{p}) + \sum_{\Gamma \in L_{sel}} \Delta S_\Gamma(\mathbf{p}) \ . \tag{73}$$

Next we need to estimate the values of the fields and of the couplings, solution to the inverse Ising problem. One possibility would be to use recursion relations similar to (72) for $\Delta h_{i,\Gamma}(\mathbf{p})$ and $\Delta J_{ij,\Gamma}(\mathbf{p})$, that is, the contributions to, respectively, the field h_i and the coupling J_{ij} coming from the cluster Γ. Next we could sum up those contributions over the clusters included in L_{sel}. However, to save memory space, it is possible to resort to the following, alternative procedure:

C2. define the 'multiplicities' m_Γ of the subsets Γ through:

C21. let L_{sub} be the list of all clusters in L_{sel} *and* of all their subsets. Initialize $m_\Gamma = 0$ for every $\Gamma \in L_{sub}$.

C22. For each $\Gamma \in L_{sel}$, and for each $\Gamma' \subset \Gamma$, add $(-1)^{K-K'}$ (see (29)) to $m_{\Gamma'}$, where K, K' are the sizes of, respectively, Γ, Γ'. The sub-clusters $\Gamma' = \Gamma$ must be taken into account in the addition process.

C3. estimate the fields and the couplings through

$$\begin{aligned} h_i(\mathbf{p}) &= (h_0)_i(\mathbf{p}) + \sum_{\Gamma \in L_{sub}:(i) \subset \Gamma} m_\Gamma \left(h_{i,\Gamma}(\mathbf{p}) - (h_0)_{i,\Gamma}(\mathbf{p}) \right) \ , \\ J_{ij}(\mathbf{p}) &= (J_0)_{ij}(\mathbf{p}) + \sum_{\Gamma \in L_{sub}:(i,j) \subset \Gamma} m_\Gamma \left(J_{ij,\Gamma}(\mathbf{p}) - (J_0)_{ij,\Gamma}(\mathbf{p}) \right) \ . \end{aligned} \tag{74}$$

The fields $h_{i,\Gamma}$ and the couplings $J_{ij,\Gamma}$ in step **C3** above are the ones obtained through the minimization of $S_{Ising}(\mathbf{J}|\mathbf{p})$ over $\mathbf{J} = \{h_{i,\Gamma}, J_{ij,\Gamma}\}$ in step **B1**. The fields $(h_0)_i$ and the couplings $(J_0)_{ij}$ are (minus) the derivatives of the reference entropy $S_0(\mathbf{p})$ with respect to p_i and p_{ij}, see formulas (22). The fields $(h_0)_{i,\Gamma}$ and the couplings $(J_0)_{ij,\Gamma}$ are their counterparts for the subset Γ only, *i.e.* the derivatives of $(S_0)_\Gamma(\mathbf{p})$; their expressions are given by (22) again, upon substitution of the $N \times N$ matrix $M(\mathbf{p})$ with the $K \times K$ matrix $M_\Gamma(\mathbf{p})$ restricted to the K elements of Γ only.

D. Pseudo-code of the algorithm

We now give the pseudo-code useful for the implementation of the procedures above. To improve the readability the code is broken into several parts.

We start with Algorithm 1, which computes the cross-entropy and the reference entropy for a subset Γ. The energy function H_{Ising} is defined in (6). The minimization over $\mathbf{J}$ can be done using standard numerical algorithms for convex functions. A speed-up is generally obtained when we start with $\mathbf{J}_{MF}$, the value of the interaction parameters obtained from the MF approximation (22), as an initial guess for the value of $\mathbf{J}$ [45]. In the absence of regularization, the parameter γ is set to 0. It is straightforward to change the pseudo-code to introduce the L_1-regularization instead of the L_2-norm, see formulas (14).

Algorithm 1 Computation of entropy $S_\Gamma(\mathbf{p}) - (S_0)_\Gamma(\mathbf{p})$

Require: Γ (of size K), data $\mathbf{p}$, regularization parameter γ

Computation of S_Γ*:*

Define $\big(S_{Ising}\big)_\Gamma[\mathbf{J}|\mathbf{p}] \leftarrow \log\left(\sum_{\boldsymbol{\sigma}\in\{0,1\}^K} \exp\big(-H_{Ising}[\boldsymbol{\sigma}|\mathbf{J}]\big)\right) - \sum_{i\in\Gamma} h_i\, p_i - \sum_{i<j\in\Gamma} J_{ij}\, p_{ij} + \gamma \sum_{i<j} J_{ij}^2\; p_i\,(1-p_i)\,p_j\,(1-p_j)$,

where

$\mathbf{J} = \{h_i, J_{ij}\}$ is of dimension $\frac{1}{2}K(K+1)$.

$S_\Gamma(\mathbf{p}) \leftarrow \min_{\mathbf{J}} \big(S_{Ising}\big)_\Gamma[\mathbf{J}|\mathbf{p}]$

Computation of $(S_0)_\Gamma$*:*

$\mathbf{M}_\Gamma \leftarrow K \times K$ matrix with elements $(M_\Gamma)_{ij} = \frac{p_{ij}-p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}$ with $i,j \in \Gamma$

$(S_0)_\Gamma(\mathbf{p}) \leftarrow \frac{1}{2}\log\det\mathbf{M}_\Gamma$ if $\gamma = 0$, or use formula (36) if $\gamma > 0$.

Output: $S_\Gamma(\mathbf{p}) - (S_0)_\Gamma(\mathbf{p})$

Algorithm 2 calculates the entropy ΔS_Γ of the cluster Γ and maintains the list L_{all} of all cluster-entropies computed so far. It calls Algorithm 1 as a subroutine.

Algorithm 2 Computation of cluster-entropy $\Delta S_\Gamma(\mathbf{p})$

Require: Γ (of size K), data $\mathbf{p}$, list $L_{all} = \{\Gamma', \Delta S_{\Gamma'}(\mathbf{p})\}$ of known cluster-entropies.

$\hat{L} \leftarrow \{\Gamma' : \Gamma' \subset \Gamma \text{ and } \Gamma' \notin L_{all}\}$ (ordered in increasing sizes)
for $\Gamma' \in \hat{L}$ **do**
  $\Delta S_{\Gamma'}(\mathbf{p}) \leftarrow S_\Gamma(\mathbf{p}) - (S_0)_\Gamma(\mathbf{p}) - \sum_{\substack{\Gamma''\subset\Gamma' \\ (\Gamma''\neq\Gamma')}} \Delta S_{\Gamma''}(\mathbf{p})$ using list L_{all} of cluster-entropies calculated so far
  update $L_{all} \leftarrow L_{all}$ U $\{\Gamma', \Delta S_{\Gamma'}(\mathbf{p})\}$
end for

Output: $\Delta S_\Gamma(\mathbf{p})$ and L_{all}

We can now give the core part of the procedure, which produces the list of selected clusters:

Algorithm 4 calculates the estimates for the total cross-entropy, and for the interaction parameters once the list of selected clusters L_{sel} has been obtained. It requires Algorithms 1 and 2; function $(S_{Ising})_\Gamma$ and matrix M_Γ are defined in the pseudo-code of Algorithm 1.

Algorithm 3 Adaptive cluster algorithm for the inverse Ising problem

Require: data $\mathbf{p}$, threshold Θ

$L_1 \leftarrow \{(i) : i = 1, 2, \ldots, N\}$
$L_{sel} \leftarrow \emptyset$
$K \leftarrow 1$

while L_K is not empty
  $L_{sel} \leftarrow L_{sel} \cup L_K$
  $K \leftarrow K + 1$
  $L_K \leftarrow \emptyset$
  for $\Gamma_1, \Gamma_2 \in L_{K-1}$ **do**
    $\Gamma \leftarrow \Gamma_1 \cup \Gamma_2$
    if Γ is of size $K-2$ and if $|\Delta S_\Gamma(\mathbf{p})| < \Theta$, then $L_K \leftarrow L_K \cup \Gamma$
  end for
end while

Output: list L_{sel} of selected clusters

Algorithm 4 Estimates for the cross-entropy and for the interaction parameters

Require: data $\mathbf{p}$, list L_{sel} of selected clusters

Computation of cross-entropy S:
  compute $S_0(\mathbf{p})$ using formula (21) or (13)
  $S_\Gamma(\mathbf{p}) \leftarrow S_0(\mathbf{p}) + \sum_{\Gamma \in L_{sel}} \Delta S_\Gamma(\mathbf{p})$
Computation of fields and couplings $\mathbf{J} = \{h_i, J_{ij}\}$:
  compute $\{(h_0)_i(\mathbf{p}), (J_0)_{ij}(\mathbf{p})\}$ using formula (22)
  $L_{sub} \leftarrow \{\Gamma' \subset \Gamma : \Gamma \in L_{sel}\}$
  for $\Gamma' \in L_{sub}$ **do**
    $m_{\Gamma'} = \sum_{\Gamma \in L_{sel} : \Gamma' \subset \Gamma} (-1)^{|\Gamma| - |\Gamma'|}$, where $|\Gamma|, |\Gamma'|$ are the sizes of Γ, Γ'.
    $\{h_i(\mathbf{p}), J_{ij}(\mathbf{p})\} \leftarrow \arg\min_{\mathbf{J}} \left(S_{Ising}\right)_\Gamma[\mathbf{J}|\mathbf{p}]$
    compute $\{(h_0)_{i,\Gamma}(\mathbf{p}), (J_0)_{ij,\Gamma}(\mathbf{p})\}$ using formula (22), with M_Γ replacing M.
  end for
  compute $\{h_i(\mathbf{p}), J_{ij}(\mathbf{p})\}$ using formula (74)

Output: $S(\mathbf{p}), \{h_i(\mathbf{p}), J_{ij}(\mathbf{p})\}$

VI. APPLICATIONS

In this Section we report the results of our algorithm when applied to data generated from Ising models with diverse interaction structures and various numbers B of sampled configurations. We define:

- the number N_{clu} of clusters generated by the algorithm and the size K_{max} of the largest clusters.
- the average error on the inferred couplings and fields:

$$\epsilon_h = \left(\frac{1}{N} \sum_i (h_i^{inf} - h_i)^2 \right)^{\frac{1}{2}}, \quad \epsilon_J = \left(\frac{2}{N(N-1)} \sum_{i<j} (J_{ij}^{inf} - J_{ij})^2 \right)^{\frac{1}{2}}. \tag{75}$$

  Here, J_{ij}^{inf} and h_i^{inf} denote the values of, respectively, the inferred couplings and fields, while J_{ij} and h_i are the values of the couplings and fields in the model used to generate the data.
- The error bars δh_i and δJ_{kl} on the inferred couplings and fields, resulting from the finite sampling. Those statistical fluctuations are asymptotically given by the inverse of the susceptibility matrix of the cross-entropy S_{Ising}, see equation (17). The entries of $\boldsymbol{\chi}$ can be calculated from a Monte Carlo simulation, to estimate the multi-spins correlations. In practice, a good approximation of $\boldsymbol{\chi}$ can already be obtained from the empirical

average over the B configurations in the sampling set. This procedure avoids the use of the Monte Carlo. In the presence of a L_2-regularization (13), $\gamma\, p_k(1-p_k)p_l(1-p_l)$ is added to the diagonal element $\chi_{kl,kl}$ of the susceptibility matrix, before the inversion is performed. Hence, all the eigenvalues are strictly positive and the inverse is well defined. The inversion of χ can be done with standard linear algebra routines.

Inferred couplings are called 'reliable' when their absolute value is larger than three times their statistical error-bar: $|J_{kl}| > 3\ \delta J_{kl}$.

- The reconstructed observables, p_i^{rec} and c_{ij}^{rec}, which we compare to the data, p_i and c_{ij}. Those reconstructed averages are obtained using Monte Carlo simulations of the Ising model with the inferred fields, h_i^{inf}, and couplings, J_{ij}^{inf}. For those simulations the number of sampled configurations is chosen to be much larger than B, *e.g.* $100\,B$, to minimize the uncertainty on the reconstructed averages.

- the relative errors on the reconstructed averages p_i and connected correlations c_{ij}, with respect to their statistical fluctuations due to finite sampling:

$$\epsilon_p = \left(\frac{1}{N} \sum_i \frac{(p_i^{rec} - p_i)^2}{(\delta p_i)^2} \right)^{\frac{1}{2}} \quad , \quad \epsilon_c = \left(\frac{2}{N\,(N-1)} \sum_{k<l} \frac{(c_{kl}^{rec} - c_{kl})^2}{(\delta c_{kl})^2} \right)^{\frac{1}{2}} \quad . \tag{76}$$

where the denominators in (76) measure the typical fluctuations of the data expected at thermal equilibrium, see (18), and

$$\delta c_{kl} = \delta p_{kl} + p_k \delta p_l + p_l \delta p_k \ . \tag{77}$$

If not explicitly stated otherwise we start from the value $\Theta = 1$ for the threshold and run the algorithm several times, dividing the threshold by 1.01 after each execution. The algorithm is stopped when both errors ϵ_p and ϵ_c are close to 1. We call Θ^* the final value of the threshold corresponding to this criterion. Unless explicitly stated otherwise a L_2–regularization term (13) is present, with $\gamma = 1/(10\,B\,p^2\,(1-p)^2)$, where p is the average value of the p_i's (Appendix A). As explained in Section II B the regularization term is important in case of undersampling and guarantees the convergence of the numerical minimization of S_{Ising}.

A. Independent Spin Model

It is instructive to run first the algorithm on the Independent Spins model, where each spin has a probability p_i to be 1, $1-p_i$ to be 0, independently of the other variables. Due to the noise in the sampling (finite value of B), the connected correlations c_{ij} are not equal to zero. Figure 14 shows the outcome for a system of size $N = 40$, as a function of the threshold Θ. The errors of reconstruction, ϵ_p and ϵ_c, are already smaller than one for the initial threshold value $\Theta^* = 1$. For this value of the threshold, cluster of size one only are selected. In other words, the interaction network J_0, calculated from the reference entropy $S_0 = S_{MF}$ alone, is already overfitting the data as it attempts to reproduce the correlations due to statistical fluctuations. For smaller thresholds Θ contributions from clusters of size $K \geq 2$ allow for an even more precise reproduction of the data .

The histogram of the inferred couplings, J_{ij}^{inf}, is shown in Fig. 15. It is centered around zero, and is approximately Gaussian. The standard deviation of the distribution is compatible with the statistical error bar on couplings δJ_{ij} (17) averaged on all the couplings. For the particular case of Fig. 14, 815 of the 820 inferred couplings are away from zero by less than three error bars, and are, therefore, classified as unreliable. This result is compatible with the fact that the non-zero couplings are the consequence of overfitting and do not reflect any real interactions.

Another possibility to avoid overfitting in this case is to apply the cluster expansion to the entropy S in the absence of reference entropy ($S_0 = 0$). We find that, for $\Theta^* = 1$, the reconstruction errors are $\epsilon_p = 0.07$ and $\epsilon_c = 0.8$. Therefore only one–spin clusters are taken into account, and all couplings are equal to zero exactly (since $J_0 = 0$).

B. Unidimensional Ising model

We now test the algorithm on the unidimensional Ising model with first–neighbor interactions, $J_{i,i+1} = J$, and uniform fields, $h_i = h$. The model is placed on a ring with N sites (periodic boundary conditions). Data $\mathbf{p}$ can be computed exactly (Appendix F) or through an average over B configurations, sampled by Monte Carlo simulations. We compare the performance of the inference procedure for various values of B and two values of J, h, corresponding

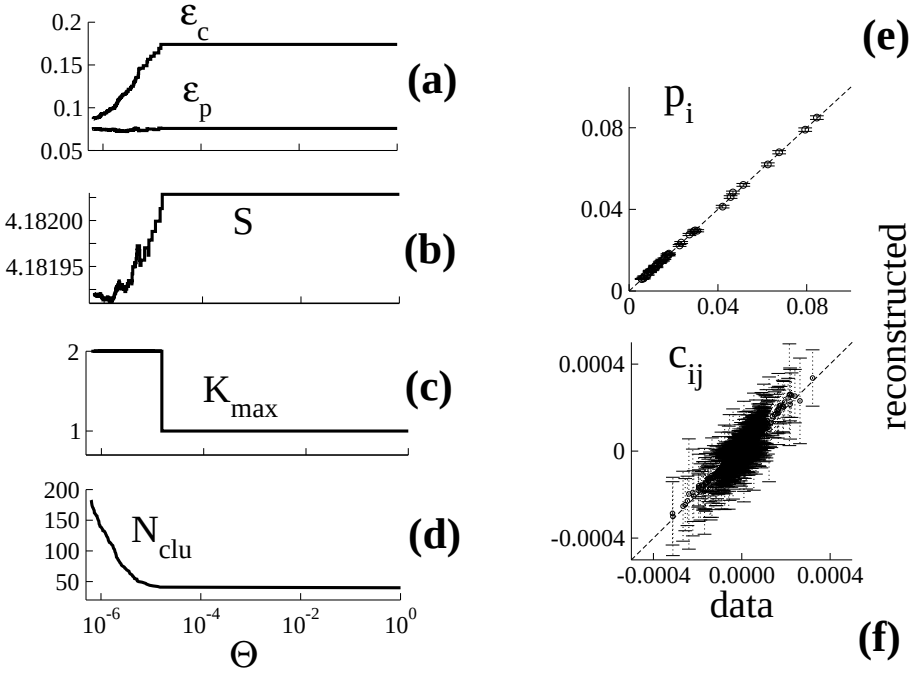


FIG. 14: Performance of the algorithm as a function of Θ for the Independent Spin model: **(a)** errors ϵ_p and ϵ_c; **(b)** cross-entropy S The entropy of the model in absence of sampling noise is $\simeq 4.182071$; **(c)** size K_{max} of largest clusters; **(d)** number N_{clu} of clusters. Panels **(e)** and **(f)** show the reconstructed p_i and c_{ij} vs. their values in the data. Error bars on p_i and c_{ij} are calculated through (77). Data were obtained by sampling $B = 10^5$ configurations of the $N = 40$ spins, with spin dependent means p_i equal to the average activity of the neurons in [2]. The optimal threshold Θ^* is already obtained with clusters of size $K = 1$.

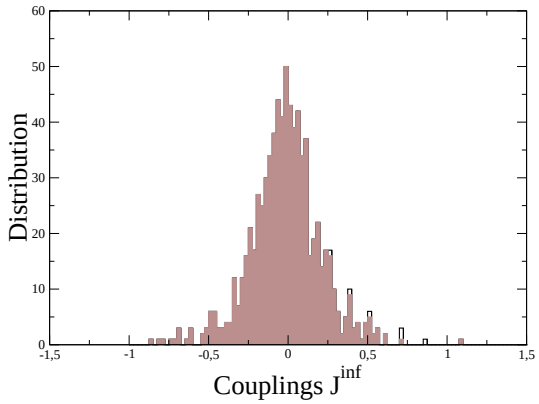


FIG. 15: Histogram of the inferred couplings for the Independent Spin model of Fig. 14, and $\Theta^* = 1.815$ of the 820 inferred couplings unreliable (in gray), because compatible with zero within three standard deviations

to the correlation lengths $\xi \simeq 1$ ($h = -5, J = 4$) and $\xi \simeq 9$ ($h = -5.95, J = 6$). These values are, respectively, smaller than $\xi_c \simeq 4.3$, the correlation length below which the cross-entropy expansion (65) is absolutely convergent, and larger than $2\,\xi_c \simeq 8.6$, above which condition (70) is violated, see Section IV D.

1. *Accuracy of the cluster expansion as a function of the threshold: errors on the entropy, couplings and fields*

We start with the small–ξ case. Figure 16(top) shows ϵ_S, the absolute value of the difference between the cross-entropy $S(\mathbf{p}, \Theta)$ (48) and the entropy of the model for perfect sampling, and the errors ϵ_J and ϵ_h, for various values of

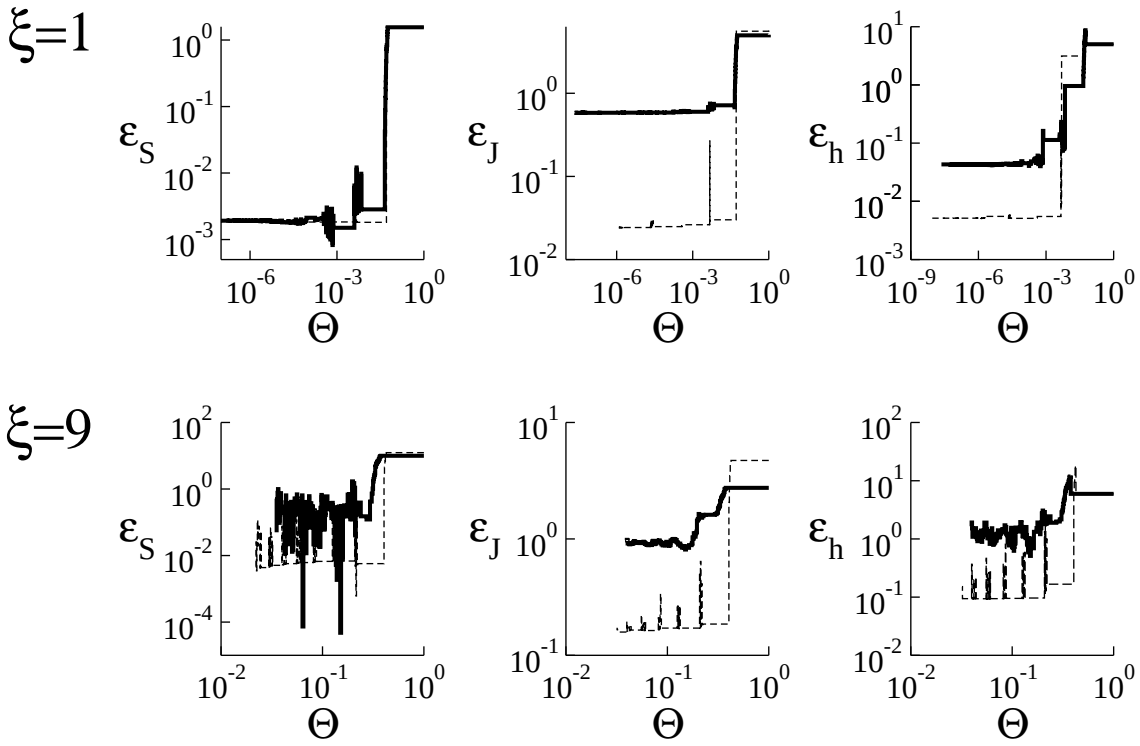


FIG. 16: Errors on the entropy (ϵ_S), the couplings (ϵ_J), and the fields (ϵ_h) vs. threshold Θ for the unidimensional Ising model with $\xi = 1$ (top) and $\xi = 9$ (bottom). Full lines correspond to large sampling noise ($B = 10^5$ for $\xi = 1$, $B = 10^3$ for $\xi = 9$), dashed line to weak sampling noise ($B = 10^7$ for $\xi = 1$, $B = 10^5$ for $\xi = 9$). The size is $N = 30$.

B. We observe that ϵ_S sharply decreases around $\Theta_1 = 0.05$, that is, the entropy of nearest-neighbor clusters $\Delta S_{(i,i+1)}$; discarding all entropies smaller than this value would be exact in the perfect sampling case (Section III C 2). As Θ is decreased, ϵ_S exhibits fluctuations centered around a discrete sequence of threshold values, Θ_d, with $d \geq 2$. As explained in Section III C 2, these values correspond to the cluster-entropies $\Delta S_{(i,i+d)}$ in the absence of noise, see identity (45). Fluctuations spread over a small window around Θ_d. They correspond to imperfect cancellations of 'packets' of entropies whose associated clusters share the same interaction path with length $L = 2d$ (Section IV D). Since the correlation length ξ is small, and, therefore, the cross entropy expansion is absolutely convergent, the magnitude of the fluctuations quickly decreases with d. For $B = 10^5$ two bursts of fluctuations are visible (corresponding to $d = 2, 3$). For $B = 10^7$, fluctuations are smaller and spread over narrower windows; only the $d = 2$ burst can be observed. In between two bursts of fluctuations, ϵ_S reaches a plateau. Note that the value of ϵ_S on the plateau is not zero due to the sampling noise (finite B).

The errors on the inferred couplings and fields have the same behaviour as ϵ_S (Fig. 16(top)). Their magnitude are comparable to the expected statistical fluctuations calculated from the 4-spin correlations (inverse susceptibility in (17)), which decrease as $B^{-1/2}$.

We now test our algorithm on the unidimensional Ising model with a large correlation length, $\xi = 9$. The errors $\epsilon_S, \epsilon_J, \epsilon_h$ are shown in Fig. 16(bottom) as functions of Θ. The correlation length of the model is larger than $2\xi_c$, at which the series of the squared cluster-entropies is divergent. Therefore, we expect large fluctuations of ϵ_S, corresponding to packets of clusters with the same interaction path (Section IV D). Figure 16 shows, indeed, that fluctuations are much larger for $\xi = 9$ (bottom) than for $\xi = 1$ (top). Furthermore, fluctuations do not decrease much in amplitude as Θ is decreased. In the case of severe undersampling ($B = 1000$) the notion of bursts of fluctuations separated by plateaus is blurred out. The distributions of cluster-entropies in a packet is so wide that it overlaps with the distributions of entropies associated to the neighbouring packets. As for the previous case the magnitude of ϵ_J, ϵ_h are comparable to the expected statistical fluctuations calculated from (17).

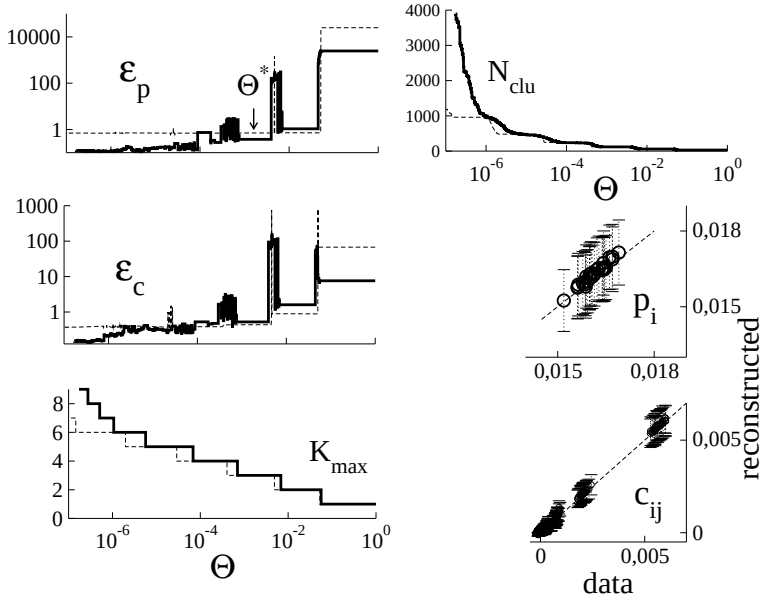


FIG. 17: Performance of the algorithm on the unidimensional Ising model with $\xi = 1$, $B = 10^5$ (full-bold line) $B = 10^7$ (dashed line). For both values of B the optimal threshold $0.0007 < \Theta^* < 0.003$ is reached with clusters of size $K = 3$ and length $L = 6$. The reconstruction of data p_i and c_{ij} is shown only for the $B = 10^5$ case (large sampling noise). Error bars on p_i and c_{ij} are computed from (77).

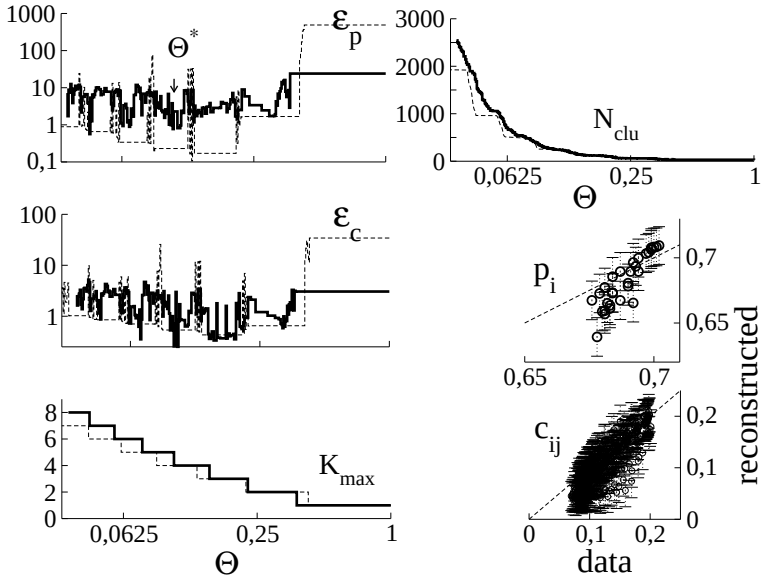


FIG. 18: Performance of the algorithm on the unidimensional Ising model with $\xi = 9$, $B = 10^3$ (full-bold line) and $B = 10^5$ (dashed line). For the largest sampling noise case ($B = 10^3$) at the threshold $\Theta^* \simeq .104$, 89 clusters of size $K = 2$, 92 clusters of size $K = 3$, 35 clusters of size $K = 4$, and 2 clusters of size $K = 5$ are selected. The reconstruction of data p_i and c_{ij} is shown for the $B = 10^3$ case. Error bars on p_i and c_{ij} are computed from (77).

2. Quality of reconstruction and choice of the threshold Θ^*

We now study the reconstruction errors ϵ_c and ϵ_p as functions of Θ, for the small and large correlation lengths and for the weak and strong sampling noises. We show in Fig. 17 the errors ϵ_c and ϵ_p, as well as the maximal size, K_{max}, and the number, N_{clu}, of clusters vs. the threshold Θ in the case of a small correlation length, $\xi = 1$. The threshold

at which both ϵ_c and ϵ_p are close to 1 can be chosen in the range $0.0007 < \Theta^* < 0.003$, for which all the clusters of lengths $L \leq 4$ are processed. It is possible to check in Fig. 16(top) that this threshold value gives the minimum values of ϵ_S, ϵ_J, ϵ_h . Contrary to the case of perfect sampling, it is not sufficient to take into account clusters with contour length $L = 2$ only. The selected clusters correspond to three groups of $N = 30$ clusters each; the first group gathers the clusters $(i, i+1)$ $(L = 2)$, the second one, the clusters $(i, i+2)$ $(L = 4)$ and the third one, the clusters $(i, i+1, i+2)$ $(L = 4)$. In Fig. 17 we show the reconstructed averages p_i and correlations c_{ij}, at Θ^* and for the largest sampling noise case $B = 10^5$, vs. their values in the data . The agreement is very good, and falls within the statistical fluctuations expected at equilibrium for the Ising model, given in (77). We stress that the optimal value of the threshold and the maximal size of selected clusters depend on the particular realization of the data $\mathbf{p}$, and can vary from sample to sample. By further decreasing the threshold Θ below Θ^*, the reconstruction errors ϵ_p and ϵ_c decrease to values smaller than one (Fig. 17). This regime corresponds to an overffiting of the data, as the errors on the couplings, ϵ_J, and on the fields, ϵ_h, cease to decrease (Fig. 16(top)).

Results for the case of a larger correlation length ($\xi = 9$) with $B = 10^5$ and $B = 10^3$ sampling configurations are reported in Fig. 18. For the poor sampling case ($B = 10^3$) plateaus are not present any longer, but a good inference is still obtained for a value Θ^* of the threshold, which, as in the $\xi = 1$ case, corresponds to the summation of all the clusters with contour length $L = 4$. This finding supports the discussion of Section IV C: the contour length required for a good inference is largely independent of the correlation length. However, in the poor sampling case, finding the right value for Θ^* is harder for larger ξ due to the mixing of packets.

3. Quality of the inference: histograms of couplings

To better understand the quality of the inference we plot in Fig. 19 (up and middle panels) the histogram of the inferred couplings J_{ij}^{inf} at the threshold Θ^*. The distribution is bimodal: a Gaussian-like peak centered in $J^{inf} = 0$ and a smaller distribution around $J^{inf} = 4$. The two sub-distributions are separated by a wide gap. The inference algorithm makes no classification error: the sub-distribution centered around $J^{inf} = 0$ contains all the pairs (i, j) such that $J_{ij} = 0$, and the one around $J^{inf} = 4$ includes all the pairs of nearest neighbours $(i, i+1)$. All the couplings centered around zero are unreliable. Moreover the standard deviation of the distribution of the couplings around the zero value (equal to the minimal value of ϵ_J reached on the plateau in Fig. 16) agrees with the statistical fluctuations (17); all the couplings around zero are therefore unreliable. The structure of the interaction network is perfectly recovered.

We show the histogram of couplings for $\Theta > \Theta^*$, *i.e.* $\epsilon_c > 1$, in Fig. 19 (bottom); the structure of the interaction network is still perfectly recovered but the values of the positive inferred couplings is less accurate. The histogram of inferred couplings for the large correlation length , $\xi = 9$, is shown in Fig. 20. Even when the sampling noise is large, a good separation of the two sub-distributions corresponding to interacting and non-interacting pairs of spins is achieved for large values of the threshold Θ^*, and the couplings are correctly inferred (up to the expected statistical fluctuations).

C. Regular bidimensional grid

We now analyze the performance of the algorithm on bidimensional grids of different sizes, $N = N' \times N'$. Nearest neighbors on the grid interact through the coupling J. The value of $J \simeq 1.778$ is chosen to make the grid critical in the thermodynamical limit, $N' \to \infty$ [47, 48]. Hence, the system is at the paramagnetic/ferromagnetic critical point, and the correlation length ξ' diverges with N'.

We have described in Section IV A and Fig. 6 the partial cancellation of the cluster-entropies for a small bidimensional grid ($N' = 3$), and no sampling noise. Due to this cancellation property, taking into account clusters of contour length $L \leq 4$ was sufficient to obtain a very accurate approximation to the cross-entropy. Hereafter, we show that this result is not affected by the presence of sampling noise. Furthermore, we will see that the size of the clusters necessary for a good inference of the interactions remains rather constant when the grid size is increased from $N' = 7$ to $N' = 20$, and is thus, as in Section VI B, largely independent of the correlation length ξ.

1. The small 3×3 grid revisited: influence of the sampling noise

We start with the 3×3 grid of Section IV A, for which the summations of all clusters up to size $K = 9$ gives the exact solution of the inverse problem, and all 1- and 2-point averages can be calculated exactly in the perfect

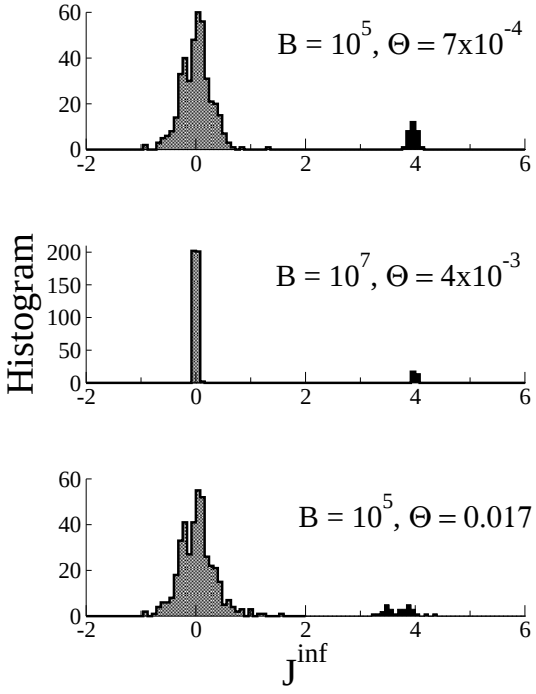


FIG. 19: Histograms of the inferred couplings J_{ij}^{inf} for the unidimensional Ising model with $\xi = 1$. Couplings equal to $J = 4$ and $J = 0$ in the model are shown in, respectively, black and gray. Gray couplings are unreliable, as they differ from zero by less than three standard deviations. Values of B and Θ are indicated on the figure. The bin width is $\Delta J = .08$.

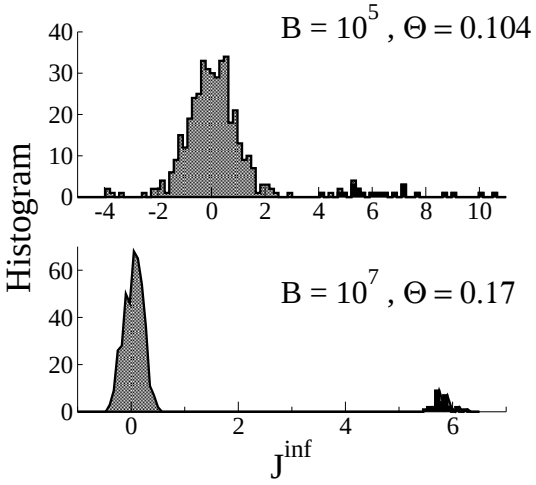


FIG. 20: Same as Fig. 19, but with $\xi = 9$ instead of $\xi = 1$. Couplings equal to $J = 6$ and $J = 0$ in the model are shown in, respectively, black and gray. Gray couplings are unreliable, as they differ from zero by less than three standard deviations.

sampling case. The reader is kindly referred to Fig. 6 and to the related discussion. Figure 21 shows the errors on the entropy, the couplings, the fields, the reconstructed 2- and 1-point averages, and the size and the number of clusters as functions of Θ. For a given amount of sampling noise (set by the value of B), the errors $\epsilon_S, \epsilon_J, \epsilon_h$ follow their perfect-sampling counterparts, until a threshold value Θ_{sat}, and they saturate for $\Theta < \Theta_{sat}$. The saturations are interrupted by fluctuations due to the imperfect cancellations of clusters within a packet (Section IV D). The saturation values of ϵ_J and ϵ_h decrease with increasing B, and are compatible with the expected statistical fluctuations δJ_{ij} given by (17). The value of Θ_{sat} approximately coincides with the threshold Θ^*, at which both ϵ_c and ϵ_p are $\simeq 1$ (Fig. 21). For $B = 4500$, only clusters of size $K = 2$ are taken into account at Θ^*. For $B = 10^7$, clusters made

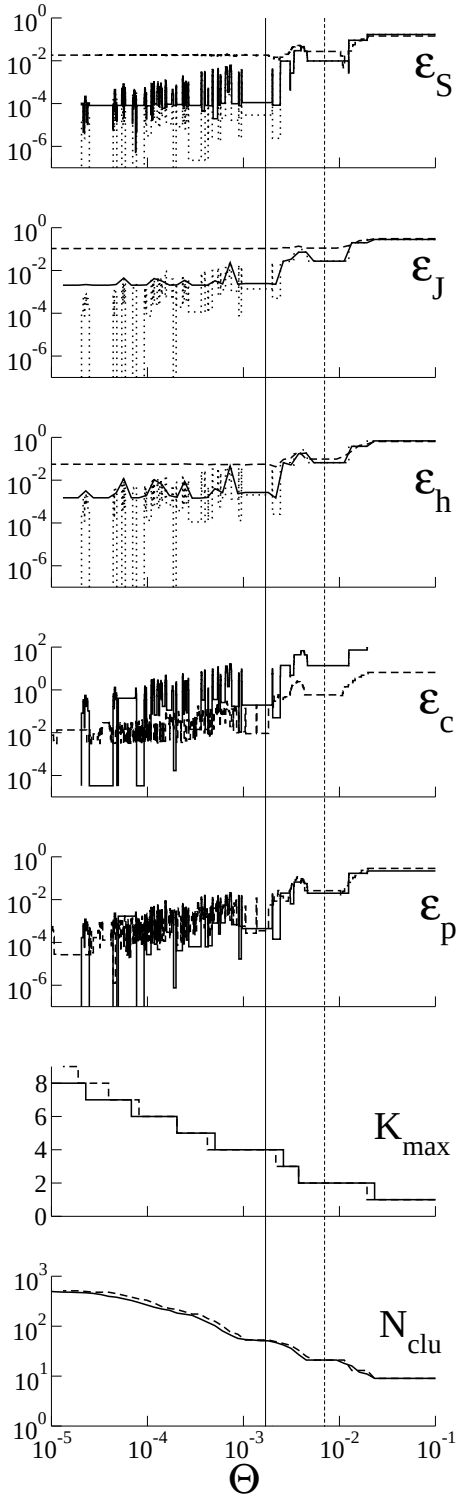


FIG. 21: Errors ϵ_s, ϵ_J, ϵ_h, ϵ_c, ϵ_p, size K_{max}, and number K_{clu} of clusters vs. Θ for a 3×3 grid with $J = 1.778$. The dashed line corresponds to $B = 4500$ sampled configurations, the full line to $B = 10^7$; the perfect sampling curves are shown with dotted lines in the top three panels (same data as Fig. 6). The values of Θ^* are shown with vertical lines: $\Theta^* = .0017$ (full, $B = 10^7$), and .007 (dashed, $B = 4500$). At Θ^* 21 clusters ($B = 4500$) and 49 clusters ($B = 10^7$) are selected

.

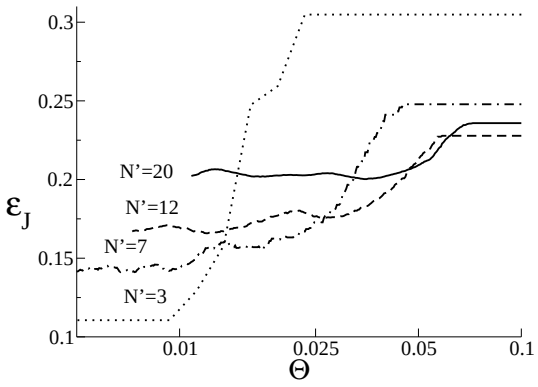


FIG. 22: Error ϵ_J for a bidimensional grid $N' \times N'$, with $N' = 3, 7, 12, 20$, and for $B = 4500$ sampled configurations.

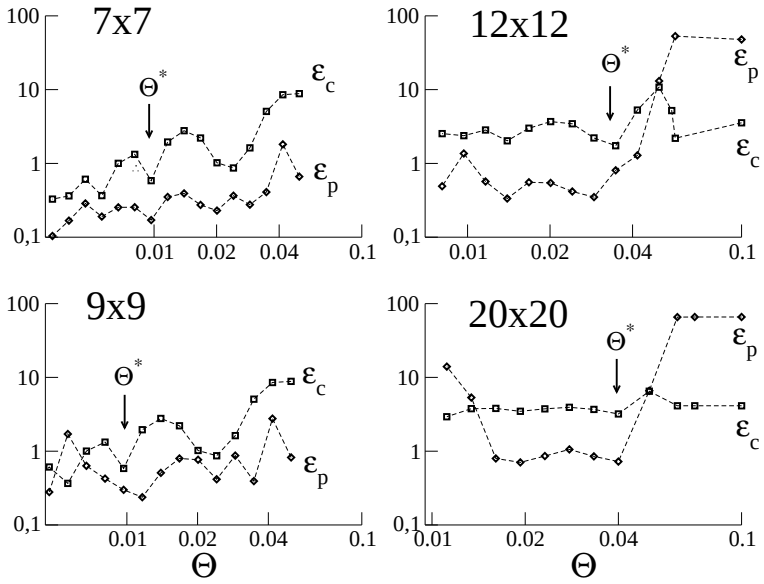


FIG. 23: Reconstruction errors ϵ_c (squares) and ϵ_p (diamonds) for grids of sizes 7×7, 9×9,12×12 and 20×20. Optimal thresholds Θ^* are located by arrows. Other possible choices for Θ^*, for instance $\Theta^* = .024$ for the 7×7 grid are equally possible.

of spins on the elementary squares of Fig. 5 and of size up to $K = 4$ are selected, *e.g.* $1, 5$ or $1, 2, 4, 5$ with $L = 4$ in Fig. 5, while clusters such as $(1, 2, 3)$ or $(1, 3)$, which have the same contour length but are not on elementary squares, are discarded. At Θ^* the histogram of the inferred couplings is made of two far apart sub-distributions (not shown): the first one corresponds to the 12 non zero couplings and is centered around $J^{inf} \simeq 1.8$, and the second one, peaked around $J^{inf} = 0$, contains the remaining 24 pairs of sites. Hence, the structure of the interaction graph is correctly found back.

2. Study of larger critical grids

We now turn to larger grids $N' \times N'$, where N' ranges between 7 and 20. Data are calculated from $B = 4500$ configurations sampled through a Monte Carlo simulation. The error ϵ_J on the couplings is shown in Fig. 22. As Θ decreases, ϵ_J saturates to a value close to the average of the expected statistical error, δJ_{ij}, which lies between .1 and .2. Saturation begins at large values of the threshold, even when the linear size N' of the grid is increased. The asymptotic values depends strongly on N due to the non periodic boundary conditions [57].

The reconstructions errors ϵ_p and ϵ_c are shown in Fig. 23. The maximal size of the clusters at the optimal threshold

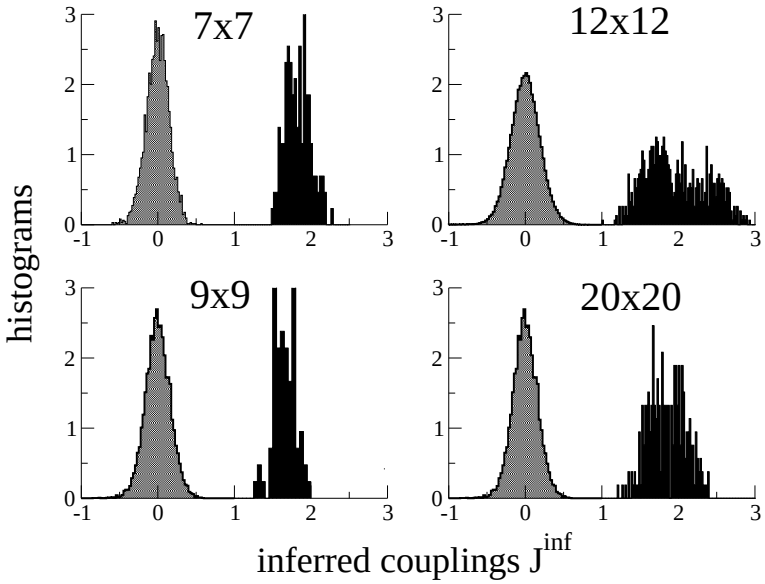


FIG. 24: Histograms of inferred couplings for bidimensional grids of different sizes. Couplings equal to $J = 1.778$ and 0 in the model are shown in, respectively, black and gray; the integral of each sub-distribution is normalized to one by hand. The values of Θ^* for each size are shown in Fig. 23.

Θ^* is bounded ($K_{max} \leq 4$), as the correlation length ξ diverges with N'. As a consequence, the running time of the algorithm increases linearly with N. For threshold values smaller than Θ^*, ϵ_p and ϵ_c decrease. The 7×7 grid in Fig. 23 provides a clear illustration of data overfitting. Keeping B fixed while N' and N increase make the data effectively more and more noisy. This is the reason why the value of Θ^* slightly increases with N.

The histograms of the inferred couplings at the threshold Θ^* are shown in Fig. 24. The structure of the grid is perfectly reconstructed for all sizes N'. We find that all the inferred couplings in the sub-distribution centered around zero are smaller (in absolute value) than three times their standard deviation (corresponding to the asymptotic value of Fig. 22) and are, therefore, unreliable.

D. Randomly diluted bidimensional grid

We now remove a fraction $1-\rho$ of the couplings on the grid, independently and at random. Our goal is to study how the algorithm performs on such disordered systems, at the phase transition and in the low temperature phase. We will compare the performance with another low complexity algorithm, the regularized logistic regression algorithm, guaranteed to perform well at high temperature and to fail at low temperature [15]. To compare with the numerical experiments of [16], we have generated 7×7 bidimensional grids, and keep each bond with probability $\rho = .7$. The remaining bonds are all equal to J [58]. We generate, for each value of J ranging from 0.4 to 4.4, eight randomly diluted grids. For each grid we calculate the data $\mathbf{p}$ by sampling over $B = 4500$ configurations generated by a Monte Carlo dynamics.

1. Inference of the network structure from the mean field entropy S_{MF}

Our first task consists, as in [16], in reconstructing the structure of the interaction graph only. We do not want to accurately determine the value of the coupling constants J_{ij}, but only if it is positive or null. This task is easier than the precise inference of the couplings, and we will first handle by approximating the cross-entropy S with the reference entropy $S_0 = S_{MF}$ only. Equivalently, we choose Θ^* to be large enough that no cluster is selected by our algorithm. We compute the mean-field couplings, $(J_0)_{ij}$, and, for each pair (i, j), and decide that a bond is present if $(J_0)_{ij} > J/2$, absent otherwise. The performances of this simple, Mean-Field algorithm are shown in Fig. 25. We say that the neighborhood of a vertex i is reconstructed if the sets of its neighbors j ($J_{ij} \neq 0$) and of its non-neighbors ($J_{ij} = 0$) are correctly inferred. In Fig. 25 (top panel) we report the fraction Q_{succ} of the neighborhoods which are reconstructed (straight line), as a function of the coupling strength J. We compare the performance of the mean-field

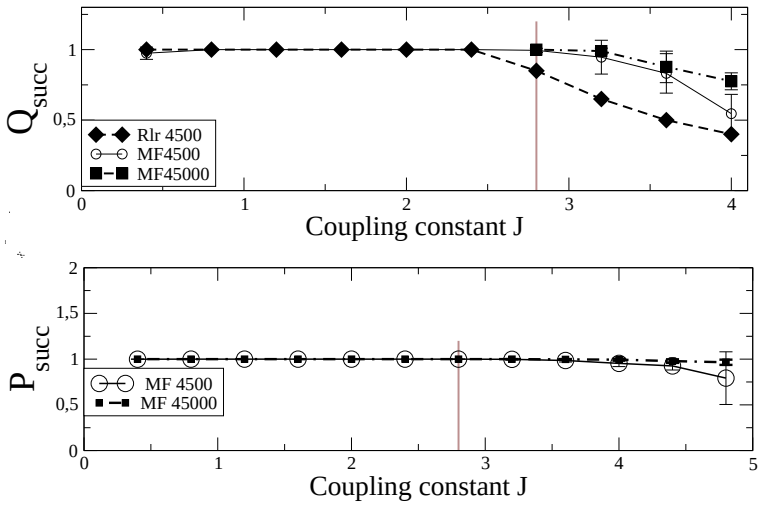


FIG. 25: Probabilities that a neighborhoods is reconstructed (top) and that a bond is inferred (bottom) as functions of the coupling J for a bidimensional random 7×7 grid density $\rho = .7$ and for various values of the numbers B of sampled configurations. Error bars are calculated from the standard deviations over eight random grids. Probabilities were obtained from the simple Mean-Field algorithm ($S_0 = S_{MF}$, $\Theta = 1$). The value of Q_{succ} (top) is compared to the performances of the pseudo-likelihood algorithm (Rlr) of [15, 16] in the top panel.

algorithm with the pseudo-likelihood algorithm of [15] in Fig. 25(right). Contrary to the pseudo-likelihood algorithm case, the neighborhoods are perfectly reconstructed at the phase transition. Furthermore, Q_{succ} remains large in the ferromagnetic phase: for instance, for $J = 3.2$, more than 80% of neighborhoods are perfectly inferred. For very large J (low temperatures) the average p_i are too close to 0 in the down state and to 1 in the up state. Most of the sampled configurations coincide with one of the two ground states, and the inference is difficult. Fig. 25(top panel) shows the increase of Q_{succ} resulting from a ten-fold increase of B.

Another measure of the performances is shown in Fig. 25(bottom panel). We plot P_{succ}, the fraction of bonds in the grid correctly predicted to exist, averaged over the data realizations, as a function of the coupling strength J. For $J = 3.2$ more than 99.78% (respectively 99.96%) of the bonds are correctly predicted with $B = 4500$ (respectively $B = 45000$) configurations. For even larger values of the coupling constant, $J = 4$, more than 95 % (respectively 99%) of the bonds are correctly predicted with $B = 4500$ (respectively $B = 45000$) with the Mean-Field algorithm.

In Section VI D 3 we show that the probability of success increases in the ferromagnetic phase, when using our algorithm with a well-chosen threshold Θ rather than the simple Mean-Field procedure ($\Theta = 1$), *e.g.* all neighborhoods are perfectly reconstructed for $J = 3.2$ ($Q_{succ} = 1$).

2. Is thermalization relevant to the inference at low temperatures?

The diluted bidimensional grid, with a fraction $\rho = .7$ of non-zero bonds, undergoes a transition from a paramagnetic to a ferromagnetic phase at the value $J_{crit}(\rho = .7) \simeq 2.8$ (vertical line in Fig. 25) in the infinite size limit [47, 48]. In the ferromagnetic phase, $J > J_{crit}(\rho = 0.7)$, and for small bidimensional grids, two competing 'states' coexist: the 'down' state, where most spins are 0, and the up state, where most spins are equal to 1. The system 'jumps' from one state to the other, as shown by the time-dependence of the average activity,

$$\mu(t) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i(t) \ , \tag{78}$$

where t is the Monte Carlo time. Figure 26(a) shows that the two states are equally sampled on a 9×9 grid, with $N_A = 10,000$ single spin-flip attempts with the Metropolis rule in between two sampled configurations (the results of Fig. 25 on the 7×7 grid were obtained with the same value of N_A). To investigate the performance of the algorithm when the two states are not well sampled we have studied a 9×9 grid, with $N_A = 100$ and $N_A = 1,000$. For $N_A = 100$, fig. 26(b) shows that few transitions occur, and that the two states will likely not be weighted equally. On larger 12×12 grids no jump occurs, even with $N_A = 1,000$ spin-flip attempts (Fig. 26(c)). The values of the spin averages

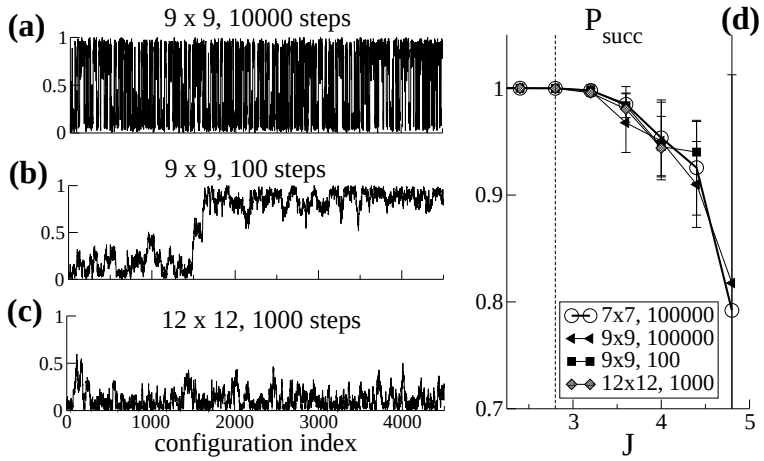


FIG. 26: **(a,b,c).** Average activity μ (78) of the $B = 4500$ sampled configurations for three grid sizes and numbers of spin-flip attempts in between two samplings. **(d)** Probability of success for $J > 2.8$, for the same data as in **(a,b,c)** and for the 7×7 grid of Fig. 25.

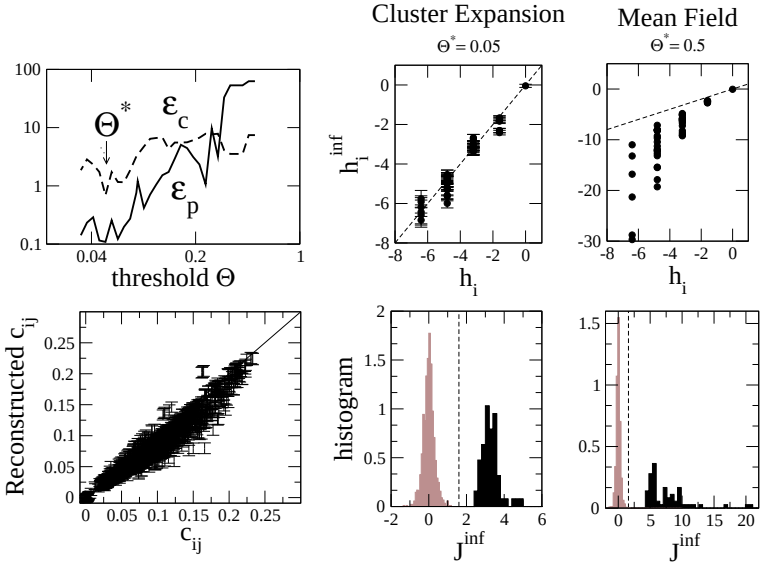


FIG. 27: Performances of the inference algorithm for the 7×7 randomly diluted grid, with $J = 3.2$, $B = 4500$, $N_A = 10^5$. Left: Relative errors ϵ_p, ϵ_c as functions of the threshold (top) and comparison of reconstructed and data correlations at Θ^* (bottom). Middle and Right: comparison of the inferred and true fields (top) and histograms of inferred couplings (bottom) for our cluster algorithm and the mean field procedure. Color code for the histograms: brown/gray: unreliable couplings (which also correspond to zero couplings in the true network), black : reliable couplings. The two sub-distributions are normalized separately.

p_i will strongly vary between the three cases above: $p_i \simeq .5$ in the mixed case (a), $p_i \simeq .7$ in the particular realization (b) of the partially mixed case, and p_i close to zero in case (c). Remarkably the probability of success of the algorithm is not sensitive to the nature of the mixing. Fig. 26(d) shows, indeed, that the reconstruction performances do not significantly decrease even in the partially and non mixed cases compared to the fully thermalized case.

3. Inference of the couplings and reconstruction of 1- and 2-point averages with the cluster expansion

It is harder to determine the values of the fields and of the couplings and to reconstruct the frequencies than to infer the structure of the interaction graph alone. To this aim the minimization of the MF entropy S^{MF} is generally not

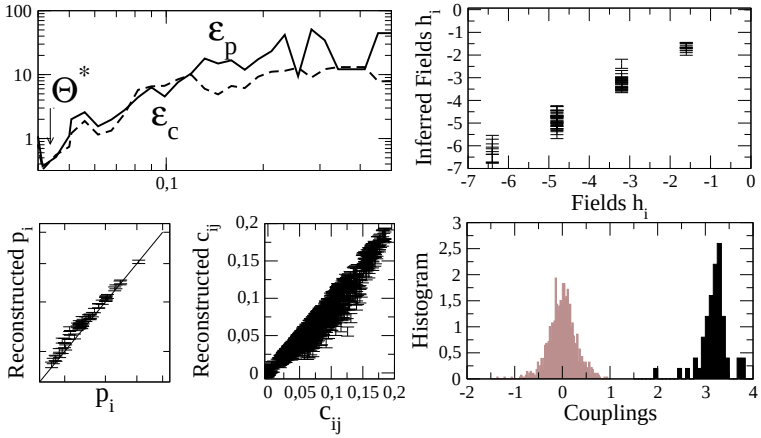


FIG. 28: Performance of the inference algorithm in the case of poor mixing, for a 7×7 randomly diluted grid, with $J = 3.2$, $B = 4500$, $N_A = 10^2$. Top & left: relative errors ϵ_p, ϵ_c as functions of the threshold Θ. Bottom & left: reconstructed p_i and c_{ij} versus their data values. Right: Inferred fields h_i^{inf} vs. their true values for $\Theta^* = 0.05$ (top) and histogram of the inferred couplings J^{inf} (bottom). As usual, unreliable couplings, which correspond to zero couplings in the true network, are depicted in gray, and reliable couplings in black.

sufficient, and the cluster expansion of $S - S^{MF}$ has to be carried out. In Fig. 27 we show the relative errors ϵ_p and ϵ_c in the reconstruction of one- and two-site frequencies as a function of the threshold Θ for the same 7×7 randomly diluted grid as in Section VI D 1. We also compare the fields and couplings inferred with our cluster algorithm (middle panels in Fig. 27, $\Theta^* = .05$) to the ones found with the simple MF procedure (right panels in Fig. 27, large value of Θ). Note that the small error done in the graph learning for $J = 3.2$ ($P_{succ} = 0.997$) can be avoided when the threshold value is optimized for each data realization.

We show in Fig. 28 the performances of the algorithm in the case of poor mixing, when the two states are not equally sampled. For the particular realization corresponding to Fig. 28, the frequencies are $p_i \simeq 0.3$ instead of $p_i \simeq 0.5$. In spite of the poor mixing the inference of the fields and couplings is as accurate as in the case of well-mixed sampling. The difference between the fields corresponding to the apparent frequencies $p_i \simeq 0.3$ and the true one ($p_i = .5$) are, indeed, smaller than the statistical uncertainty on the fields due to the limited sampling (finite value of B). The reason is that, near a critical point, a small variation in the field is sufficient to produce a large change in the average values of the spins.

E. Erdös-Renyi random graphs

In this Section we report the results of our inference algorithm when applied to disordered Ising models on random graphs. The random networks are generated from the Erdös-Renyi ensemble, where $M = \frac{d}{2}N$ edges are drawn, uniformly and at random, between N points. Parameter d is the average degree of a vertex on the network.

Figure 29 shows the outcome of the algorithm when data are generated from an Erdös-Renyi model of connectivity $d = 10$. On the selected bonds (i, j) the couplings J_{ij} were chosen uniformly at random in $[-3; 3]$. All other couplings J_{ij} were set to zero, and the fields were $h_i = -1$. Values of the parameters are such that the system is in the paramagnetic phase (in the thermodynamic limit). Panel A shows the inference with good sampling (the data are obtained by averaging over $B = 10^6$ Monte Carlo configurations), while Panel B shows the inference with poor sampling ($B = 10^3$). At Θ^* the data are reconstructed within the expected statistical fluctuations and, correspondingly, couplings are found back within the statistical error bars δJ_{ij}. In the case of a large sampling noise case $B = 10^3$, the statistical fluctuations δJ_{ij} are so large that most of the inferred couplings are unreliable. The inference of the complete network is thus not possible. The maximal size of clusters at θ^* increases with the average degree (Section VI F); we find $K_{max} = 9$ and $K_{max} = 7$ for, respectively, $B = 10^6$ and $B = 10^3$.

Fig. 30 shows the outcome of the algorithm on an Erdös-Renyi random graph with a smaller connectivity, $d = 5$, and for values of the couplings J_{ij} chosen uniformly at random in $[-4; 4]$. The fields are set to $h_i = -\frac{1}{2}\sum_{j(\neq i)} J_{ij}$, in such a way that the corresponding fields in spin variable ± 1 vanish. This is an example of a smaller connectivity system in the spin-glass phase. We have studied the performance of the algorithm as a function of the threshold Θ, by varying the system size N from 50 to 200 and the number of sampled configurations from $B = 1000$ to 10000. The

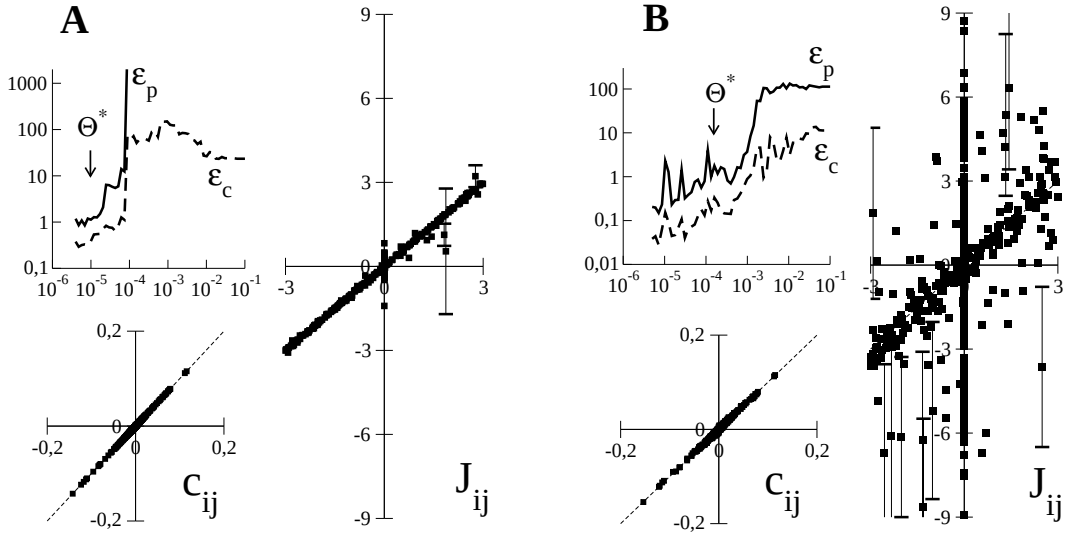


FIG. 29: Outcome of the inference algorithm for an Erdös-Renyi random graph with $N = 50$ spins, connectivity $d = 10$, and with $B = 10^6$ **(a)** and $B = 10^3$ **(b)** sampled configurations. For each value of B we show the errors ϵ_p, ϵ_c vs. θ, the inferred vs. data values of the correlations c_{ij}, and of the couplings J_{ij}. A few large error bars over J_{ij} (calculated from χ^{-1}) are shown. Values of J_{ij} were chosen uniformly at random in $[-3; 3]$, and fields were set to $h_i = -1$.

algorithm is able to reach $\epsilon_c = 1$ at large thresholds, with a small number of selected clusters, *e.g.* $N_{clu} < 1000$ and $K_{max} = 7$ for $N = 100$. The threshold Θ^* for which $\epsilon_c = 1$ corresponds to the beginning of the plateau for ϵ_J. For smaller thresholds ϵ_c decreases and data are overfitted. The height of the plateau for ϵ_J coincides with the calculated statistical error δJ; it scales as $1/\sqrt{B}$ and does not strongly depend on N.

F. Computational Time

For a given value Θ of the threshold the computational time can be estimated through

$$\text{time} \simeq \sum_{K=1}^{K_{max}} N_{clu}(K)\, 2^K \ , \tag{79}$$

where $N_{clu}(K)$ is the number of selected clusters of size K, and 2^K is the number of operations necessary to calculate exactly the partition function of a sub-system of size K. As the number of selected clusters depends on the interaction graph, the computational time is sensitive to the structure of the interaction graph, while it does not depend too much on the correlation length of the system. For instance, the number of processed clusters and the computational time for Erdös-Renyi graphs is larger for the connectivity $d = 10$ than for $d = 5$.

Moreover, as the sampling noise increases (the value of B is made smaller), so does the threshold value Θ^*. As less precision is needed in the reconstructed frequencies and correlations, the size of the selected clusters is reduced. As a consequence, the computational time is reduced. Figure 31 illustrates this statement for Erdös-Renyi random graphs: the running time increases with the quality of the sampling, *i.e.* increases with the value of B. In some very noisy cases, however, large size clusters which are due only to the noise and do not reflect the interaction network can be processed. As an example, the number of clusters for the unidimensional Ising model with $\xi = 9$, $B = 1000$ is larger than the one for $B = 4500$. Another illustration is given by the diluted 7×7 grid with $J = 3.2$, which requires the processing of many clusters of large size ($K_{max} = 8$).

VII. CONCLUSION AND PERSPECTIVES

In this paper, we have presented an adaptive cluster expansion to infer the interactions between a set of Ising variables from the measure of their equilibrium correlations. We have discussed the statistical mechanics of this

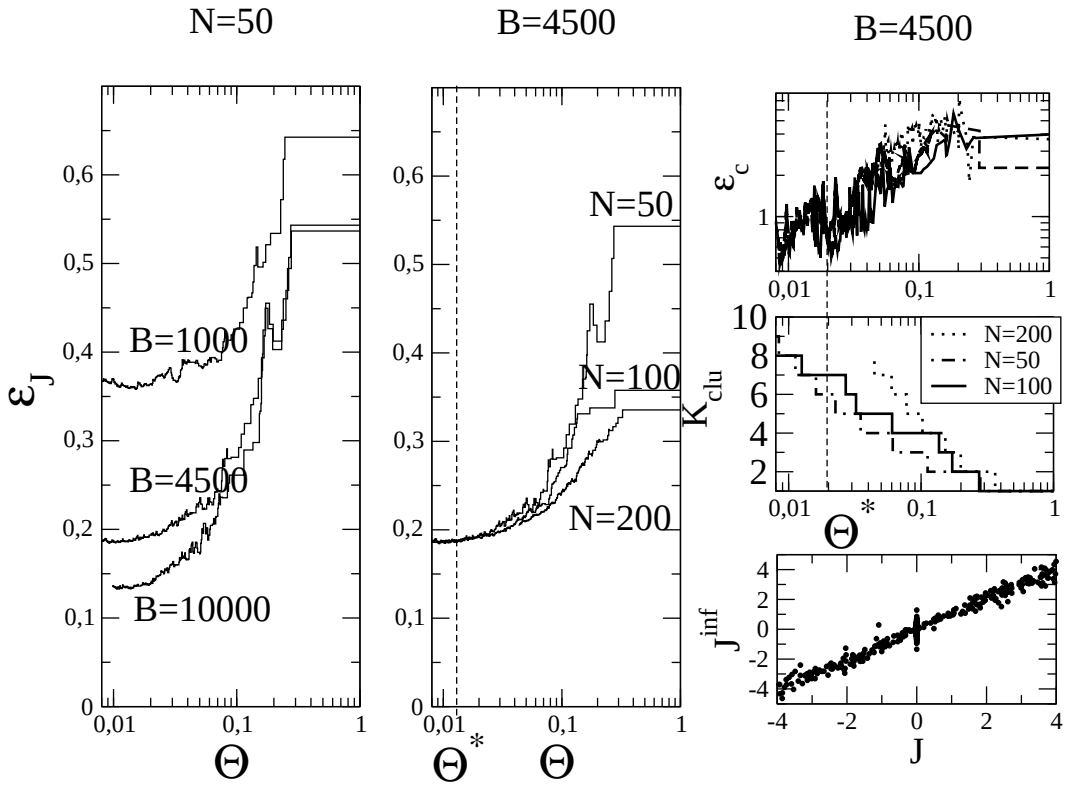


FIG. 30: Outcome of the inference algorithm for an Erdös-Renyi random graph with N spins, and connectivity $d = 5$. Values of J_{ij} were chosen uniformly at random in $[-4; 4]$, and fields were set to $h_i = -\frac{1}{2}\sum_{j(\neq i)} J_{ij}$. Left: error ϵ_J on the inferred couplings as a function of Θ for $N = 50$ and $B = 1000, 4500, 10000$ configurations. Middle: error ϵ_J vs. Θ for $N = 50, 100, 200$ and for $B = 4500$. Right: ϵ_c (top) and K_{clu} (middle) as functions of Θ; the inferred couplings $\mathbf{J^{inf}}$ are compared to their true values in the bottom panel for $N = 100$ and threshold Θ^*.

expansion, and shown applications of the algorithm to artificial data generated using Ising models on unidimensional and bidimensional lattices, as well as on Erdos-Renyi random graphs.

We have in particular underlined the important conditions on the inverse problem that should be fulfilled for our algorithm to be efficient. The essential condition is that the inverse susceptibility, which determines the change of a coupling (or a field) resulting from a change in the data (1- or 2-spin frequencies) should be well-conditioned. We stress that this property is not incompatible with the presence of a long-range susceptibility. Hence, the inverse problem can be easy to handle even in the presence of long-range correlations. As far as our algorithm is concerned, this condition entails that the maximal size K_{clu} of the clusters which need to be taken into account remains small even if the correlation length of the system is large.

The origin of this condition is that our algorithm builds up, by definition, an interaction network defining a well-conditioned Ising model. Indeed, in the absence of reference entropy ($S_0 = 0$), the cross-entropy $S(\mathbf{p})$ is approximated through a sum of a restricted number of cluster-entropies, see (30). For sufficiently large thresholds Θ, most quadruplets of variables, say, i, j, k, l, do not appear in any selected cluster (of size $K \geq 4$); hence, most of the entries $(\chi^{-1})_{ij,kl}$ of the inverse susceptibility matrix entries vanish according to (55). In the presence of the reference entropy $S_0 = S_{MF}$, $\boldsymbol{\chi}^{-1}$ is not guaranteed to be sparse any more due to the contribution $\boldsymbol{\chi}_0^{-1} = -\frac{\partial S_0}{\partial \mathbf{p} \partial \mathbf{p}}$. However, when a regularization is introduced, *e.g.* based on the norm L_1 (14), the network of interactions $(J_0)_{ij}$ is highly diluted, and we expect $\boldsymbol{\chi}_0^{-1}$ to be well-conditioned, too. Further investigations of this point would be very useful.

According to the discussion of Section IV C inverse problems corresponding to Ising models on finite-dimensional lattices are well-conditioned in the perfect sampling limit. The introduction of a threshold over the minimal values of cluster-entropies allows us to force the inverse problem to be well-conditioned even in the presence of sampling noise. We have checked this statement on inverse problems corresponding to 'critical' Ising models. While the correlation length increases with the size of the system, the maximal size of the clusters, K_{clu}, remains roughly constant. Therefore, the computational complexity of the algorithm increases only linearly with the system size.

An essential feature of inverse problems is that data are generally obtained from a finite sampling and, therefore,

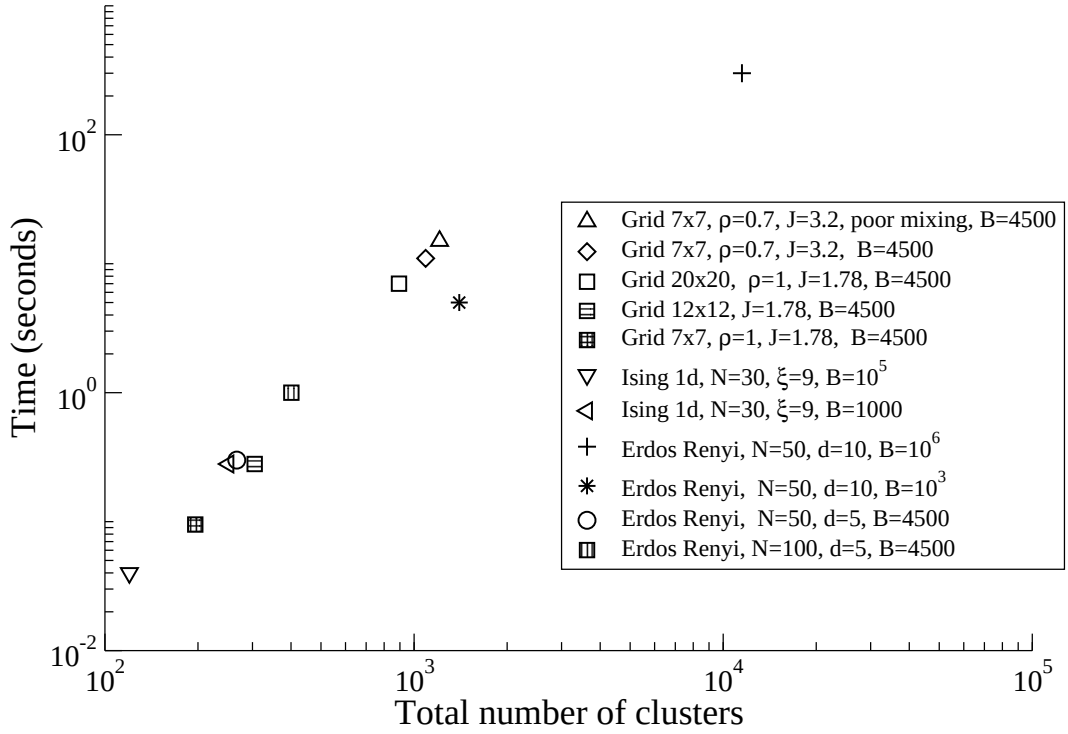


FIG. 31: Computational time of the cluster algorithm at threshold Θ^* for the different examples shown Section VI. The computational time grows with the number of processed clusters, which depends on the structure of the interaction graph and on the number of sampled configurations more than the sole number of variables, N. For a fixed maximal size of clusters, K_{max}, the running time is roughly proportional to the number of clusters. Unless explicitly stated otherwise, the sampling is realized in good mixing conditions. Times were measured on one core of a 2.8 GHz Intel Core 2 Quad desktop computer.

frequencies and pairwise correlations are plagued by sampling noise. Avoiding overfitting is a primary goal for an inference algorithm. This goal is achieved, in our algorithm, by the introduction of the threshold Θ. As a result most of the clusters are discarded, and in particular, those whose contributions would convey very little information about the true nature of the underlying interaction network. Fixing the threshold value such that the relative reconstruction errors ϵ_p and ϵ_c are of the order of one corresponds to the maximal accuracy allowed by the quality of the data.

The cluster expansion introduced here differs from other classical cluster expansions, developed in the contexts of the theory of liquids and of computational physics. In particular we do not impose consistency equations for the marginal probabilities over the clusters. Our expansion scheme is simpler, and requires only the knowledge of the individual and pairwise frequencies of the variables in the cluster. Moreover, the cluster construction and selection rules prevents any combinatorial explosion of the computational time.

Several points would deserve further investigations. Among them the discussion of the convergence properties of the expansion, started in Section IV D, should be expanded and improved. A natural and interesting question is to ask how the series behaves when the packets of Fig. 12 start mixing, *i.e.* in the presence of a strong sampling noise. Another aspect which should be better understood is the influence of the construction rule. Our heuristic consists in merging two almost completely overlapping clusters of size K to build a new cluster of size $K+1$ (provided its entropy is larger than Θ). This rule has a simple intuitive interpretation, compatible with the notion of interaction path, and attempts with other rules have been less fruitful. However, a deeper theoretical understanding and justification is clearly needed. Last of all, the *a posteriori* validation of the method relies on the use of a Monte Carlo simulation to calculate ϵ_c and ϵ_p. We have tested another procedure to avoid the use of a Monte Carlo calculation, based on a partial resummation of the cluster contributions corresponding to the free-energy (at fixed couplings and fields). This procedure, whose applicability goes beyond the inverse problem, will be detailed in a further publication.

Acknowledgements: We are grateful to J. Barton, J. Lebowitz, E. Speer for very useful and stimulating discussions, in particular regarding the correspondence between the inverse susceptibility and the direct correlation functions and the practical implementation of the inference algorithm. We thank E. Aurell for pointing to us the difference between P and Q, see Section VI D 1.

Appendix A: Optimal choice for the regularization parameter γ

In this Appendix we discuss how the optimal value for the parameter γ in the L_2-regularization (13) can be determined. As explained in Section II B, the regularization term can be interpreted as a Gaussian prior P_0 over the couplings. Let us call σ^2 the variance of this prior. Parameters γ and $1/\sigma^2$ are related through

$$\gamma\, p^2\,(1-p)^2 = \frac{1}{2\,\sigma^2\,B}\ , \tag{A1}$$

where we have assumed that the 1-site frequencies p_i are uniformly equal to p. To calculate the optimal value for γ, or, equivalently, for σ^2, we start with the case of a single spin for the sake of simplicity, and then turn to the general case of more than one spin.

a. Case of $N = 1$ spin

For a unique spin subjected to a field h the likelihood of the set of sampled spin values, $\{\sigma^\tau\}$, is $P_h[\sigma] = \exp(B\,p\,h)/(1+e^h)^B$. Here p denotes the average value of the spin over the sampled configurations (2). We obtain the *a posteriori* probability (15) for the field h given the frequency p,

$$P_{post}[h|p] = \frac{\exp(-h^2/(2\sigma^2) + B\ p\,h - B\log(1+e^h))/\sqrt{2\pi\sigma^2}}{\mathcal{P}(p,B,\sigma^2)} \tag{A2}$$

where the denominator $\mathcal{P}(p,B,\sigma^2)$ (marginal likelihood) is simply the integral of the numerator over all real-valued fields h. Given p and B we plot $I = -\log\mathcal{P}(p,B,\sigma^2)/B$ as a function of σ^2. The general shape of I is shown in Fig. 32. The value of on Bayesian grounds.

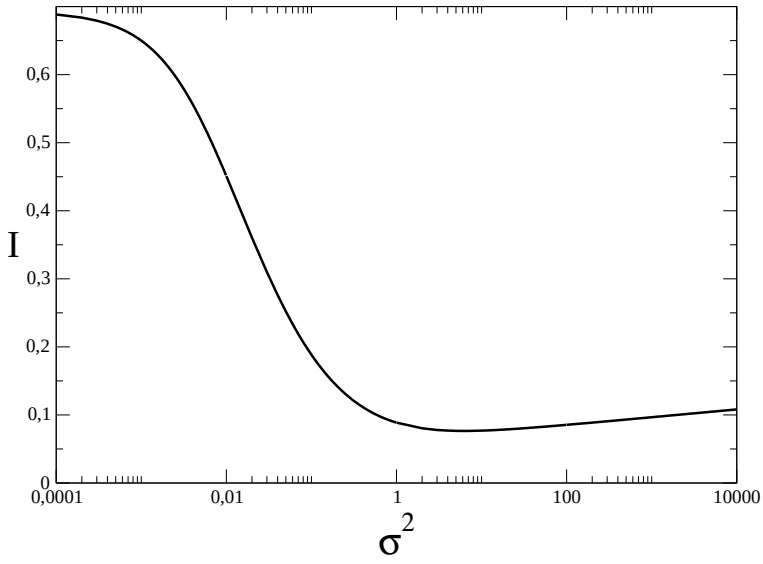


FIG. 32: Logarithm of the marginal likelihood (with a minus sign, and divided by the size B of the data set) versus variance σ^2 of the prior distribution of the field. Parameters are $B = 100$, $p = .02$.

For more than one spin calculating the marginal likelihood would be difficult. We thus need an alternative way of obtaining the best value for σ^2. The idea is to calculate I through a saddle-point method, and include the Gaussian corrections which turn out to be crucial. This approach is correct when the size of the data set is large. A straightforward calculation leads to

$$I \simeq \log\left(1+\exp(h^*)\right) - p\ h^* + \frac{\Gamma}{2}\ (h^*)^2 + \frac{1}{2B}\log\left[1+\frac{1}{\Gamma}\,\frac{\exp(-h^*)}{(1+\exp(-h^*))^2}\right] \tag{A3}$$

where $\Gamma = 1/(B\sigma^2)$ and h^* denotes the root of $(1+\exp(-h^*))^{-1} - p + \Gamma\ h^* = 0$. I decreases from $I(\sigma^2=0) = \log 2$ with a strong negative slope, $dI/d\sigma^2(0) \simeq -B\ p^2$, and increases as $\log\sigma^2/(2B)$ for large values of the variance. Expression (A3) cannot be distinguished from the logarithm of the true marginal likelihood I shown in Fig. 32.

b. Case of $N \geq 2$ spins

The above saddle-point approach can be generalized to any number N of spins, with the result

$$I \simeq S_{Ising}[\{h_i, J_{ij}\}|\mathbf{p}] + \frac{1}{2B} \log\ \det\left(1 + \frac{\mathbf{H}}{\Gamma}\right) \tag{A4}$$

where S_{Ising} was defined in (7) and $\boldsymbol{\chi}$ is the $N + \frac{1}{2}N(N-1) = \frac{1}{2}N(N+1)$-dimensional Hessian matrix composed of the second derivatives of S_{Ising} with respect to the couplings and fields (10). In principle $\boldsymbol{\chi}$ could be diagonalized and the expression (A4) calculated. However this task would be time-consuming. As we have seen in the previous subsection we expect I not to increase too quickly with σ^2 (for not too small variances) and approximate calculations of I can be done under some data-dependent hypothesis. We now give an example of such an approximation, valid in the case of multi-electrode recordings of neural cell populations.

A simplification arises when the number B of configurations and the frequency p are such that: (a) each spin i is active ($= 1$) in a number of configurations much larger than 1 and much smaller than B *i.e.* $1 \ll B \times p \ll B$; (b) the number n_2 of pairs of spins that are never active together is much larger than one and much smaller than $\frac{N(N-1)}{2}$. These assumptions are generically true for the applications to neurobiological data. For instance, the recording of the activity of $N = 40$ salamander retinal ganglion cells in [2] fulfills conditions (a) and (b) for a binning time $\Delta t = 5$ msec: a cell i firing at least once in a time-bin corresponds to $\sigma_i = 1$, while a silent cell is indicated by $\sigma_i = 0$. More precisely: (a) the least and most active neurons respectively fire 891 and 17,163 times (among $B = 636,000$ configurations); (b) $n_2 = 34$ pairs of cells (among 780 pairs) are never active together.

Condition (a) allows us to omit the presence of Γ in the calculation of the fields, $h_i \simeq \log p_i$, to the first order of a large (negative) field expansion. Condition (b) forces us to introduce a non–zero Γ to calculate the couplings, with the result that interactions between pairs i, j of cells not active together are equal to $J_{ij} \simeq \log\Gamma + O(\log\log(1/\Gamma))$. Finally we obtain the asymptotic scaling of the entropy when $\Gamma \to 0$,

$$S_{Ising} \simeq n_2\ \frac{\Gamma}{2}\ (\log\Gamma)^2 + O\Big(\Gamma \log\Gamma \log\log\frac{1}{\Gamma}\Big)\ . \tag{A5}$$

We are now left with the calculation of the determinant in (A4). From assumption (b) the number of pairs of neurons not spiking together is small with respect to N^2, meaning that most of the eigenvalues λ^a of the Hessian matrix of S_{Ising} are non zero. Hence,

$$\log\ \det\left(1 + \frac{\boldsymbol{\chi}}{\Gamma}\right) = \sum_{a=1}^{N(N-1)/2} \log\left(1 + \frac{\lambda^a}{\Gamma}\right) \simeq -\frac{N^2}{2}\ \log\Gamma\ . \tag{A6}$$

Putting both contributions to I together we get

$$I(\Gamma) \simeq n_2\ \frac{\Gamma}{2}\ (\log\Gamma)^2 - \frac{N^2}{4B}\ \log\Gamma\ . \tag{A7}$$

The optimal value for the variance σ^2 is the root of

$$\frac{dI}{d\Gamma}(\Gamma) = 0 \simeq \frac{n_2}{2}\ (\log\Gamma)^2 - \frac{N^2}{4\ B\ \Gamma} \simeq \frac{n_2}{2}\ (\log B)^2 - \frac{N^2}{4}\ \sigma^2\ . \tag{A8}$$

We finally deduce the optimal variance

$$\sigma^2 \simeq 2\ n_2\ \left(\frac{\log B}{N}\right)^2\ . \tag{A9}$$

For the data described above we find $\sigma^2 \simeq 8$.

Appendix B: Expression of the entropy of clusters with size $K = 3$

In this Appendix, we give the analytical expression for the entropy of a cluster with $K = 3$ spins. Using this expression instead of minimizing the cross-entropy (7) offers a valuable computational speed-up as there are $O(N^3)$ clusters of size $K = 3$. We start with the definition of the entropy $P(\boldsymbol{\sigma})$:

$$S_3 = -\sum_{\substack{\sigma_1=0,1\\ \sigma_2=0,1\\ \sigma_3=0,1}} P(\sigma_1,\sigma_2,\sigma_3)\ \log P(\sigma_1,\sigma_2,\sigma_3)\ . \tag{B1}$$

We then replace the probabilities $P(\sigma_1, \sigma_2, \sigma_3)$ of the eight configurations of the three spins above with their expressions in terms of the probabilities $\{p_i, p_{kl}\}$ in the data, and of the probability p_{123} that the three spins are equal to 1:

$$\begin{aligned}
P(1,1,1) &= p_{123} \\
P(1,1,0) &= p_{12} - p_{123} \\
P(1,0,1) &= p_{13} - p_{123} \\
P(0,1,1) &= p_{23} - p_{123} \\
P(0,0,1) &= p_3 - p_{23} - p_{13} + p_{123} \\
P(0,1,0) &= p_2 - p_{12} - p_{23} + p_{123} \\
P(1,0,0) &= p_1 - p_{13} - p_{12} + p_{123} \\
P(0,0,0) &= 1 - p_1 - p_2 - p_3 + p_{12} + p_{13} + p_{23} - p_{123}
\end{aligned} \tag{B2}$$

The only unknown quantity (not available in $\mathbf{p}$) is the probability p_{123}. To determine p_{123} we impose

$$\frac{dS_3}{dp_{123}} = 0 \ , \tag{B3}$$

which means that the three-body coupling J_{123} vanishes. Condition (B3) gives a third degree equation on p_{123},

$$p_{123}^3 + \alpha\, p_{123}^2 + \beta\, p_{123} + \gamma = 0 \tag{B4}$$

with

$$\begin{aligned}
\alpha &= p_1\, p_2 + p_1\, p_3 + p_2\, p_3 - p_1\, p_{23} - p_2\, p_{13} - p_3\, p_{12} - p_{12} - p_{23} - p_{13} \ , \\
\beta &= p_1\, p_{23}^2 + p_2\, p_{13}^2 + p_3\, p_{12}^2 - p_1\, p_2\, p_{23} - p_1\, p_2\, p_{13} - p_1\, p_3\, p_{12} - p_1\, p_3\, p_{23} - \\
&\quad p_2\, p_3\, p_{12} - p_2\, p_3\, p_{13} + 2\, p_{12}\, p_{13}\, p_{23} + p_{12}\, p_{13} + p_{12}\, p_{23} + p_{13}\, p_{23} + p_1\, p_2\, p_3 \ , \\
\gamma &= p_1\, p_2 + p_3\, (1 - p_1 - p_2 - p_3 + p_{12} + p_{13} + p_{23}) \ .
\end{aligned} \tag{B5}$$

Upon substitution of p_{123} in (B1) we obtain the desired cross-entropy S_3, as a function of the three average values p_i and the three two-point averages p_{kl}. The expression of the cluster-entropy is given by,

$$\Delta S_{(i,j,k)} = S_3(p_i, p_j, p_k, p_{ij}, p_{ik}, p_{jk}) - \Delta S_{(i,j)} - \Delta S_{(i,k)} - \Delta S_{(j,k)} - \Delta S_{(i)} - \Delta S_{(j)} - \Delta S_{(k)} \ , \tag{B6}$$

according to (28). The expressions of the cluster-entropies for one and two spins are given by, respectively, (23) and (25). Similarly, one obtains the expressions for the contributions of the 3-spin cluster to the values of the interactions parameters by differentiating ΔS with respect to the p_i's and the p_{kl}'s.

Appendix C: Leading diagrammatic contributions to small cluster-entropies

We analyze the dominant diagrams contributing to the cluster-entropies for the various values of the cluster sizes, K, in the limit of small connected correlations c_{kl}.

1. Case $K = 2$

The entropy $\Delta S_{(i,j)}$ of a 2-spin cluster is the sum of all diagrammatic contributions containing two spins and an arbitrary number of links between them, corresponding to the power of the expansion parameter $M_{ij} = c_{ij}/(p_i\,(1 - p_i)\,p_j\,(1 - p_j))$ (Fig. 3) and Section III C 1. For small values of M_{ij} the largest contribution to $\Delta S_{(i,j)}$ is the one with three links (cubic power of M_{ij}), if the reference entropy $S_0 = S_{MF}$ removes the two-link loop diagram. The entropy contribution of this diagram was computed in [30], with the result

$$\Delta S_{i,j}^{(3)} = \alpha_{i,j}\, (c_{ij})^3 \ , \tag{C1}$$

where

$$\alpha_{i,j}^{(3)} = \frac{(2\,p_i - 1)\,(2\,p_j - 1)}{6\,(p_i)^2\,(1 - p_i)^2\,(p_j)^2\,(1 - p_j)^2} \ . \tag{C2}$$

The superscript 3 refers to the power of the connected correlation.

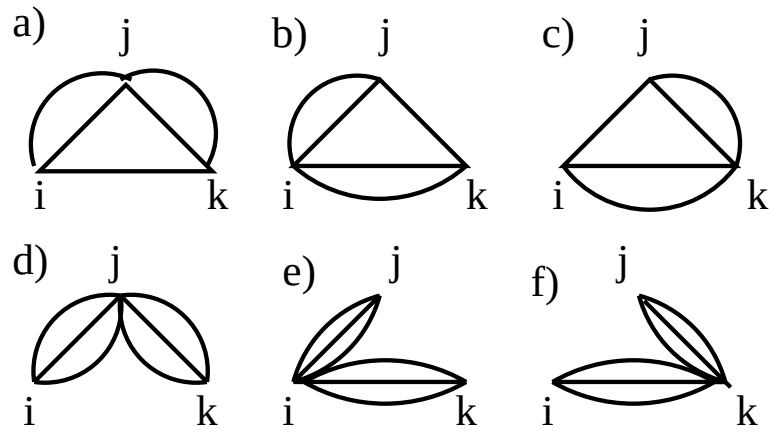


FIG. 33: Leading diagrams to the order 5 (top) and 6 (bottom) in the connected correlation for the entropy of 3-clusters.

2. Case $K = 3$

For $K = 3$ the leading term to $\Delta S_{(i,j,k)}$ in powers of M_{ij} was not derived analytically in [30]. Based on the studies of the unidimensional Ising model and the independent spin models (Appendix F), we find that the leading diagrams are diagrams (a), (b), (c) in Fig. 33 (bold diagrams in Fig. 3), whose sum is given by

$$\Delta S^{(5)}_{i,j,k} = \alpha^{(5)}_{ijk}\,(c_{ij})^2\,(c_{jk})^2\,c_{ki} + \alpha^{(5)}_{jik}\,(c_{ij})^2\,c_{jk}\,(c_{ki})^2 + \alpha^{(5)}_{jki}\,c_{ij}\,(c_{jk})^2\,(c_{ki})^2\ , \tag{C3}$$

with

$$\alpha^{(5)}_{ijk} = -\frac{(2p_i-1)\,(2p_k-1)}{2\,(p_i)^2\,(1-p_i)^2\,(p_j)^2\,(1-p_j)^2\,(p_k)^2\,(1-p_k)^2}. \tag{C4}$$

Note that $\alpha^{(5)}_{ijk}$ differs from $\alpha^{(5)}_{jik}$. We have also found the coefficients of the subsequent diagrams, of the order of M^6. These diagrams are labelled by (d), (e), (f) in Fig. 33. Their total contribution to the cluster-entropy is

$$\Delta S^{(6)}_{i,j,k} = \alpha^{(6)}_{ijk}\,(c_{ij})^3\,(c_{jk})^3 + \alpha^{(6)}_{jik}\,(c_{ij})^3\,(c_{ki})^3 + \alpha^{(6)}_{jki}\,(c_{jk})^3\,(c_{ki})^3 \tag{C5}$$

with

$$\alpha^{(6)}_{ijk} = \frac{(2p_i-1)\,(2p_k-1)}{3\,(p_i)^2\,(1-p_i)^2\,(p_j)^3\,(1-p_j)^3\,(p_k)^2\,(1-p_k)^2}. \tag{C6}$$

3. Generic case $K \geq 4$

The above results for $K = 3$ are easily generalized to any value of the cluster size $K \geq 4$. The diagrammatic expansion of a K–spin cluster includes all circuits where pairs of spins are linked together. Each diagram with (one or two) links between i_l and i_{l+1} $(l = 1, \ldots, K-1)$ and (one or two) links between i_1 and i_K gives

$$\begin{aligned}\Delta S^{(2K-1)}_{i_1,\ldots i_k} &= \frac{(-1)^K}{2\prod_{l=1}^K (p_{i_l})^2\,(1-p_{i_l})^2}\Big[(2p_{i_{k-1}}-1)\,(2p_{i_k}-1)\,(c_{i_1,i_2})^2(c_{i_2,i_3})^2\ldots(c_{i_{k-2},i_{k-1}})^2 c_{i_{k-1},i_k} \\ &+ (2p_{i_{k-2}}-1)\,(2p_{i_{k-1}}-1)\,(c_{i_1,i_2})^2\,(c_{i_2,i_3})^2\ldots c_{i_{k-2},i_{k-1}}(c_{i_{k-1},i_k})^2+\ldots \\ &+ (2p_{i_1}-1)\,(2p_{i_2}-1)\,c_{i_1,i_2}\,(c_{i_2,i_3})^2\ldots(c_{i_{k-2},i_{k-1}})^2\,(c_{i_{k-1},i_k})^2\Big]\ .\end{aligned} \tag{C7}$$

At the next order in power of M_{ij}, each diagram with three links between i_l and i_{l+1} $(l = 1, \ldots, K-1)$ gives a contribution

$$\Delta S^{(3K-3)}_{i_1,\ldots i_k} = \frac{(-1)^{K-1}\,(2p_{i_1}-1)\,(2p_{i_k}-1)}{3(p_{i_1})^2\,(1-p_{i_1})^2\,(p_{i_k})^2\,(1-p_{i_k})^2\prod_{l=2}^{K-1}(p_{i_l})^3\,(1-p_{i_l})^3}\prod_{l=1}^{K-1}(c_{i_l,i_{l+1}})^3\ . \tag{C8}$$

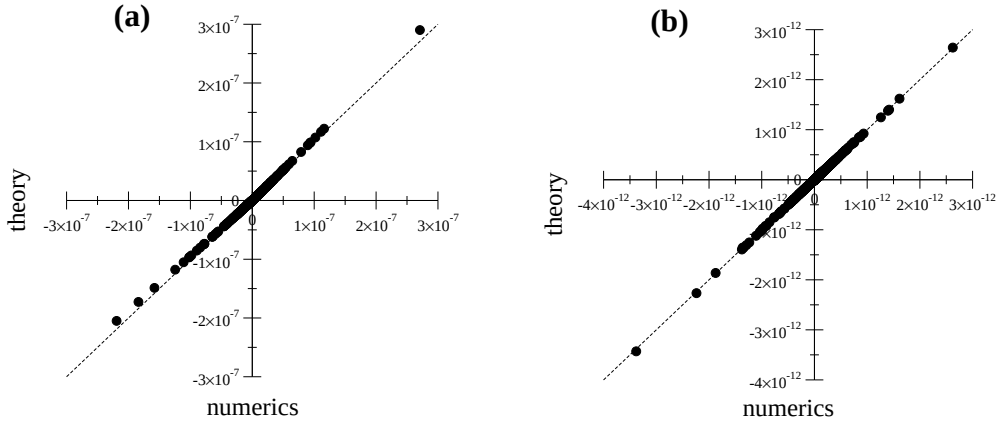


FIG. 34: Comparison of the numerical (x-axis) and theoretical (y-axis) values for the entropies of clusters (i,j) **(a)** and (i,j,k) **(b)**. The system is made of $N = 40$ independent spins, with the same p_i as in the neural data of Ref. [2]; the average value of the p_i's is $p = 0.0238$. Theoretical predictions correspond to (C3) and (C5). The number of sampled configurations is $B = 10^6$.

Appendix D: Critical correlation length ξ_c for the absolute convergence

In this Appendix, we briefly explain why the cluster-entropy series is absolutely convergent if and only if the correlation length ξ is smaller than

$$\xi_c = \frac{\Omega}{\log v} \ . \tag{D1}$$

Here, $\Omega = 2$ when the reference entropy is $S_0 = 0$, and $\Omega = 3$ when $S_0 = S_{MF}$. Parameter v denotes the number of neighbours of a site on the lattice, supposed to be uniform. For instance, $v = 2D$ on a hypercubic lattice in dimension $D \geq 1$.

Consider a set of K distinct points on the lattice. Let $\mathcal{N}(L)$ be the number of closed paths of length L visiting all K points. We obviously have $\mathcal{N}(L) \leq v^L$. Hence, the series

$$\sum_L \mathcal{N}(L) \ \exp\Big(-\,\Omega\, L/\xi\Big) \tag{D2}$$

is convergent if $\xi < \xi_c$. Reciprocally, let L_0 be the length of the shortest closed path $\mathcal{C}_0$ encircling the K points. A closed path of length $L_1 + L_0$ can be built from $\mathcal{C}_0$ by attaching a closed loop of length L_1 to any one of the sites in $\mathcal{C}_0$. Hence, for $L \geq L_0 + 2$, $\mathcal{N}(L) \geq L_0\ v^{L-L_0}$. We deduce that the series (D2) is divergent if $\xi > \xi_c$.

Appendix E: Distribution of cluster-entropies for the Independent Spin model

We generate B configurations of N independent spins σ_i. Spin i is equal to 1 with probability p and to zero with probability $1 - p$ (for simplicity we assume here that all the frequencies p_i are equal to the same value p). The empirical connected correlations c_{ij} computed from the B sampled configurations of spins are generally non zero. The marginal distribution of c_{ij} is a normal law, with zero mean and standard deviation (49). The largest values of the correlations are, for a system with N spins, of the order of

$$c_{ij}^{MAX} = c_B\ \sqrt{4 \log N}\ , \tag{E1}$$

according to extreme value theory.

We compare in Fig. 34 formulas (C3) and (C5) for, respectively, the cluster-entropies ΔS_{ij} and ΔS_{ijk} with numerics carried out from randomly sampled configurations. Each pair (i,j) (Fig. 34(a)) and triplet (i,j,k) (Fig. 34(b)) define a point, whose coordinates are the numerical and theoretical values of the entropy corresponding to the pair- or triplet-cluster. The agreement, for $B = 10^6$ sampled configurations, is excellent due to the small value of $c_B \simeq 2\,10^{-5}$.

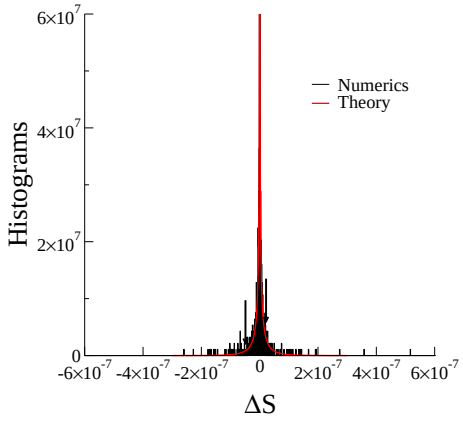


FIG. 35: Theoretical (red) and numerical (black) histograms H_{IS} for the entropies ΔS of 2-spin clusters in a system of independent spins and $B = 10^6$ configurations. Simulations were done with $N = 40$ spins, with heterogeneous $p_i \simeq .0238$, see caption of Fig. 34.

1. Distribution of cluster-entropies for $K = 2$

The distribution of the entropy of $K = 2$-clusters for a set of $B = 10^6$ configurations is shown in Fig. 35. To derive the analytical expression of the distribution in the $N \to \infty$ limit, we use the small-correlation formula (C1) for $\Delta S_{(1,2)}$, and the fact that the distribution of the connected correlation is Gaussian. As a result, approximating $\alpha_{1,2}$ with its average value α obtained by substituting p_1 and p_2 with p in (C2), we obtain

$$H_{IS}(\Delta S_{(1,2)}) = \frac{\exp\Big(-\frac{(\Delta S_{(1,2)})^{2/3}}{2\,(c_B)^2\alpha^{2/3}}\Big)}{3\,\alpha^{1/3}\,\sqrt{2\pi(c_B)^2}\,(\Delta S_{(1,2)})^{2/3}} \tag{E2}$$

This distribution is a stretched exponential at infinity, and diverges in zero. Its standard deviation is

$$\sigma_{\Delta S_{(1,2)}} = \sqrt{15}\,\alpha\,(c_B)^3 = \frac{\sqrt{15}\,(2p-1)^2}{6\,p(1-p)\,B^{3/2}} \tag{E3}$$

For $B = 10^6$ and $p = 0.0238$ we obtain that the standard deviation is $\simeq 2.7\,10^{-8}$. Distribution (E2) is compared to the histogram obtains from numerics in Fig. 35. The standard deviation and the distribution at small entropies are in good agreement. Large values of the correlations (E1) give rise to isolated values of $\Delta S_{(1,2)}$, of the order of

$$\Delta S^{MAX}_{(1,2)} \simeq (4\log N)^{3/2}\,\Big(\langle(\Delta S_{(i,j)})^2\rangle\Big)^{1/2}\,, \tag{E4}$$

approximately equal to $1.2\,10^{-6}$ for $N = 40$. This value is about twice the largest cluster-entropy observed in Fig. 35 for one particular realization of the sampled configurations.

2. Distribution of the cluster-entropies for $K = 3$

The leading order contribution to the entropy of a 3-cluster is given by (C3). We want to calculate the distribution of $\Delta S_{(1,2,3)}$ when the connected correlations c_{ij} are random Gaussian variables, of zero mean and variance $(c_B)^2$. We neglect the correlations between c_{12}, c_{13}, c_{23}, which is legitimate for large B. Let us call $x = \Delta S_{(1,2,3)}/(\alpha(c_B)^5)$, with α given by (C4), and let $P(x)$ be the probability density of x. Though we have not been able to find a closed expression for $P(x)$, the asymptotics behavior of P for large or small arguments can characterized analytically.

a. Large x behaviour

The Mellin transform of P [36] is

$$\int_0^\infty dx P(x) x^\lambda = \left(\frac{2}{\pi}\right)^{3/2} \int_0^\infty dc_{12}\, dc_{13}\, dc_{23}\; e^{-F(c_{12},c_{13},c_{23})} \tag{E5}$$

where

$$F(c_{12},c_{13},c_{23}) = -\frac{1}{2}(c_{12}^2 + c_{13}^2 + c_{13}^3) + \lambda\, \log(c_{12}c_{13}^2c_{23}^2 + c_{12}^2c_{13}c_{23}^2 + c_{12}^2c_{13}^2c_{23})\ . \tag{E6}$$

The tail of $P(x)$ at large x can be studied by considering large values of λ. We expect the dominant contribution to the multiple integral on the right hand side of (E5) to come from large correlations. The location of the main contribution to the integral is the value of (c_{12}, c_{13}, c_{23}) which maximizes F. As F is invariant under any permutation of its arguments, we look for a maximum where $c_{12} = c_{13} = c_{23} \equiv c^*$. A straigthforward calculation shows that

$$c^*(\lambda) = \sqrt{\frac{5}{3}\lambda}, \quad F^*(\lambda) = \frac{5}{2}\lambda\log\lambda + \lambda\left(\log 3 + \frac{5}{2}\log\frac{5}{3} - \frac{5}{2}\right)\ . \tag{E7}$$

We now use the saddle-point method again, this time to estimate the integral on the left hand side of (E5). We obtain

$$\max_x \left[\log P(x) + \lambda\, \log x\right] = F^*(\lambda)\ , \tag{E8}$$

which is true when λ is very large. Hence, $F^*(\lambda)$ is the Legendre transform of $\log P(x)$. Solving (E8) gives

$$\log P(x) \simeq -\frac{3}{2}\,\left(\frac{x}{3}\right)^{2/5} \tag{E9}$$

at large x. The distribution of the cluster entropies $\Delta S_{(1,2,3)}$ thus follows a stretched exponential with exponent $\frac{2}{5}$. This decay is much slower than an exponential, and leads to large tails as can be seen from Fig. 9.

b. Small x behavior

In order for the rescaled entropy x to be small, at least one among the three correlations should be small according to (C3). Without restriction, we may assume that c_{12} is the smallest of the three correlations. As c_{12} appears once with power one, and twice with power two in (C3), we approximate $x \simeq c_{12}c_{13}^2c_{23}^2$. The Mellin transform of P is, for negative λ,

$$\int_0^\infty dx P(x) x^\lambda \simeq 3\left(\int_0^\infty dc \frac{2}{\sqrt{2\pi}}\, c^\lambda\, e^{-c^2/2}\right)\,\left(\int_c^\infty dc \frac{2}{\sqrt{2\pi}}\, c^{2\lambda}\, e^{-c^2/2}\right)^2\ . \tag{E10}$$

The largest pole is located in $\lambda = -\frac{1}{2}$, and is of order 2. According to standard results on the inversion of Mellin transforms [36], we obtain a precise characterization of the divergence of the probability density at small x,

$$P(x) \simeq C\,\frac{(-\log x)}{\sqrt{x}}\ , \tag{E11}$$

where C is a constant.

c. Typical value of x

The typical value of the x is defined through

$$x_{typ} = \exp\left(\int_0^\infty dx\, P(x)\, \log x\right)\ . \tag{E12}$$

This quantity is less sensitive than the average value of x to the presence of the long tails in $P(x)$ at large x. We write $x = (c_{12}c_{13}c_{23})^2\, z$ where

$$z = \frac{1}{c_{12}} + \frac{1}{c_{13}} + \frac{1}{c_{23}} \ . \tag{E13}$$

Taking the logarithm, and averaging over the correlation, we obtain the following expression for the average value of the logarithm of x,

$$\langle \log x \rangle = 6 \left(\int_0^\infty dc \, \frac{2 \log c}{\sqrt{2\pi}} \, e^{-c^2/2} \right) + \langle \log z \rangle_z \ . \tag{E14}$$

The integral over c in the above equation can be calculated numerically, with a value $\simeq -.63518$. To calculate the average value of $\log z$, we first use the identity

$$\log z = \int_0^\infty \frac{du}{u} \left(e^{-u} - e^{-u\, z} \right) \ . \tag{E15}$$

Taking the average on both sides, we have

$$\langle \log z \rangle_z = \int_0^\infty \frac{du}{u} \left(e^{-u} - \langle e^{-u\, z} \rangle_z \right) \ . \tag{E16}$$

As z is a sum of independent random variables its Laplace transform is the product of their Laplace transforms,

$$\langle e^{-u\, z} \rangle_z = \left(\int_0^\infty dc \, \frac{2}{\sqrt{2\pi}} \, e^{-c^2/2 - u/c} \right)^3 = \left(\frac{\lambda}{2\pi\sqrt{2}} \, G^{30}_{03} \left(\frac{\lambda^2}{8} \middle| \begin{matrix} - \\ -\frac{1}{2}, 0, 0 \end{matrix} \right) \right)^3 \ , \tag{E17}$$

where G is the Meijer-G function. we have calculated the integral (E16) using the Mathematica software. Some care must be taken for the numerical accuracy when $z \to 0$. The outcome is $\langle \log z \rangle_z \simeq 2.09643$. Putting all contributions together we obtain $x_{typ} \simeq 0.18$. The corresponding values of $\Delta S_{(1,2,3)}$ are 3.5×10^{-15} for $B = 10^6$, and 1.1×10^{-12} for $B = 10^5$, in good agreement with the numerical value, respectively, 3×10^{-15} and 9×10^{-13}.

d. Standard deviation of x

We can easily evaluate the variance of each of the three terms of the sum in (C3) as the product of the variances of the three terms in the product, based on the approximation that the connected correlations c_{ij} are independent stochastic variables. We obtain

$$\sigma_{\Delta S_{ijk}} = \frac{3\sqrt{3}(2p-1)^2\,(c_B)^5}{2\,p^6\,(1-p)^6} = \frac{3\sqrt{3}\,(2\,p-1)^2}{2\,p\,(1-p)\,B^{5/2}} \ . \tag{E18}$$

With the values of N and p chosen in Fig. 35, we find that the standard deviation is of the order of 10^{-13} for $B = 10^6$, and $2\,10^{-11}$ for $B = 10^5$, see Fig. 9.

3. Distribution of cluster-entropies for generic $K \geq 4$

In general, for $K \geq 3$, the leading contribution to $\Delta S_{(i_1,i_2,\ldots i_K)}$ (C7) contains the sum of $K \times (K-1)!/2$ terms, each one being the product of K random variables, among which $(K-1)$ are elevated to power two, and 1 is elevated to power 1. The factor K comes from the fact that there are K way of choosing the single link in the circuits with K spins. The factor $(K-1)!/2$ is the number of non equivalent circuits going through K spins. We define the rescaled entropy x through

$$x = |\Delta S_{(i_1,i_2,\ldots i_K)}| \times \frac{2\,(p(1-p))^{2K}}{\sqrt{\frac{K!}{2}}\,(2p-1)^2\,(c_B)^{2K-1}} \tag{E19}$$

The approach followed in Section E 2 to calculate the asymptotic behaviour of the probability density P of x for $K = 3$ can be extended without difficulty to any value of $K > 3$. We find that $P(x)$ diverges when $x \to 0$, with

$$P(x) = C \, \frac{(-\log x)^{K-2}}{\sqrt{x}} \ . \tag{E20}$$

where C is a constant. Hence the shape of the distribution of x is, up to logarithmic terms, independent of K. On the contrary, the tail of the distribution for large x is very sensitive to K,

$$\log P(x) \simeq -\frac{K}{2(K-\frac{1}{2})^2} \, \left(\frac{x}{K}\right)^{2/(2K-1)} \tag{E21}$$

As in the $K = 3$ case, the distribution of the cluster entropies ΔS follows a stretched exponential. The exponent of the stretched exponential decreases with K. The variance of the distribution can be easily evaluated, with the result

$$\sigma_{\Delta S_{(i_1,i_2,\dots i_K)}} = \frac{\sqrt{K!/2}(\sqrt{3})^{K-1}(2p-1)^2 \, (c_B)^{2K-1}}{2(p\,(1-p))^{2K}} \ . \tag{E22}$$

Appendix F: Properties of the cluster-entropies of the one-dimensional Ising model

Consider the one-dimensional Ising model with nearest-neighbour couplings and periodic boundary conditions. The Hamiltonian of the model is

$$H = -h \sum_i \sigma_i - J \sum_i \sigma_i \, \sigma_{i+1} \ , \tag{F1}$$

where the spins σ_i take 0,1 values. The parameters of the model are the N identical fields $h_i = h$, the N couplings $J_{i,i+1} = J$ between neighbours and the remaining $N \times (N-3)/2$ zero couplings $J_{i,j} = 0$ between non neighbours.

We recall a few elementary facts about the model. The transfer matrix is

$$T = \begin{pmatrix} e^{J+h} & e^{h/2} \\ e^{h/2} & 1 \end{pmatrix} \ . \tag{F2}$$

The eigenvalues are $\lambda_\pm = \frac{1}{2}\left(e^{J+h} + 1 \pm \sqrt{(e^{J+h}-1)^2 + 4\,e^h}\right)$, and the two components of the eigenvectors are, respectively, $v_\pm(1) = -(1-\lambda_\pm)/\sqrt{e^h + (1-\lambda_\pm)^2}$ and $v_\pm(2) = e^{h/2}/\sqrt{e^h + (1-\lambda_\pm)^2}$. The probability that a spin is up is given by, in the $N \to \infty$ limit,

$$p = \langle \sigma_i \rangle_{\mathbf{J}} = \big(v_+(1)\big)^2 \ , \tag{F3}$$

and the connected correlation at distance d is

$$c_{i,i+d} = \langle \sigma_i \, \sigma_{i+d} \rangle_{\mathbf{J}} - \langle \sigma_i \rangle_{\mathbf{J}} \langle \sigma_{i+d} \rangle_{\mathbf{J}} = p(1-p) \, \left(\frac{\lambda_-}{\lambda_+}\right)^d = p(1-p) \exp(-d/\xi) \ , \tag{F4}$$

where the correlation length is given by $\xi = -1/\log(\lambda_-/\lambda_+)$.

1. Calculation of the cluster-entropies and cancellation property

In this Section, we show the exact cancellation property between the entropies of clusters with different sizes discussed in Section III C 2. We will see that this property is a direct consequence of the existence of a unique interaction path along the unidimensional chain.

a. Case $S_0 = 0$

We first consider the case where the reference entropy is zero. Let $\Gamma = (i_1, i_2, \dots, i_K)$ be a cluster of size K, with $i_1 < i_2 < \dots < i_K$. Due to the unidimensional nature of the interactions, the Gibbs distribution over the K-spin configurations $\boldsymbol{\sigma}$ obeys the chain rule,

$$P_{\mathbf{J}}[\boldsymbol{\sigma}] = P_{\mathbf{J}}(\sigma_{i_K}|\sigma_{i_{K-1}}) \dots P_{\mathbf{J}}(\sigma_{i_4}|\sigma_{i_3}) \, P_{\mathbf{J}}(\sigma_{i_3}|\sigma_{i_2}) \, P_{\mathbf{J}}(\sigma_{i_2}, \sigma_{i_1}) \ , \tag{F5}$$

where $P(\cdot,\cdot)$ and $P(\cdot|\cdot)$ denote, respectively, joint and conditional probabilites. Inserting the above formula into expression (9) for the cross-entropy, we obtain

$$\begin{aligned} S_{Ising}(\mathbf{J}|\mathbf{p}) &= -\sum_{\boldsymbol{\sigma}} P_{obs}[\boldsymbol{\sigma}]\Big(\sum_{l=2}^{K-1}\log P_{\mathbf{J}}(\sigma_{i_{l+1}}|\sigma_{i_l}) + \log P_{\mathbf{J}}(\sigma_{i_2},\sigma_{i_1})\Big) \\ &= -\sum_{\boldsymbol{\sigma}} P_{obs}[\boldsymbol{\sigma}]\Big(\sum_{l=1}^{K-1}\log P_{\mathbf{J}}(\sigma_{i_{l+1}},\sigma_{i_l}) - \sum_{l=2}^{K-1}\log P_{\mathbf{J}}(\sigma_{i_l})\Big) \\ &= \sum_{l=1}^{K-1} S_{Ising}(h^{\rightarrow}_{i_l}, h^{\leftarrow}_{i_{l+1}}, J_{i_l,i_{l+1}}|p_{i_l},p_{i_{l+1}},p_{i_l,i_{l+1}}) - \sum_{l=2}^{K-1} S_{Ising}(h^0_{i_l}|p_{i_l})\ . \end{aligned} \tag{F6}$$

Each variable σ_{i_l}, with $l = 2,\dots,K-1$, appears three times in (F6), which explains the presence of three fields h with the same index i_l. After optimization over $\mathbf{J} = (\{J_{i_l,i_{l+1}}\},\{h^{\rightarrow}_{i_l}\},\{h^{\leftarrow}_{i_l}\},\{h^0_{i_l}\})$ all these fields are equal, and we obtain

$$S(\mathbf{p}) = \sum_{l=1}^{K-1} S(p_{i_l},p_{i_{l+1}},p_{i_l,i_{l+1}}) - \sum_{l=2}^{K-1} S(p_{i_l}) = \sum_{l=1}^{K-1}\Delta S_{(i_l,i_{l+1})}(\mathbf{p}) + \sum_{l=2}^{K-1}\Delta S_{(i_l)}(\mathbf{p})\ . \tag{F7}$$

Hence the cross-entroy $S(\mathbf{p})$ is the sum of the 1-cluster entropies and of the entropies of the 2-clusters made of adjacent sites. None of the other cluster-entropies appear, which proves that they cancel each other. To illustrate the cancellation mechanism, consider the case $K = 3$. According to (F7),

$$S(\mathbf{p}) = \Delta S_{(i_1,i_2)}(\mathbf{p}) + \Delta S_{(i_2,i_3)}(\mathbf{p}) + \Delta S_{(i_1)}(\mathbf{p}) + \Delta S_{(i_2)}(\mathbf{p}) + \Delta S_{(i_3)}(\mathbf{p})\ . \tag{F8}$$

Comparing with (28) we obtain

$$\Delta S_{(i_1,i_2,i_3)}(\mathbf{p}) = -\Delta S_{(i_1,i_3)}(\mathbf{p})\ , \tag{F9}$$

which shows that the entropy of a 3-cluster and the one of a 2-cluster with the same extremities i_1, i_3 are opposite to one another. By a recursive applications of (F7) this result can be immediately generalized to higher values of K. The entropy of a K-cluster is simply the entropy of the 2-cluster with the same extremities, multiplied by $(-1)^{K-2}$. Hence, identity (44) is established.

According to formula (F4) for the connected correlation, the entropy of a two-site cluster is a function of the distance d between the two sites:

$$\Delta S_{(i,i+d)} = F\Big(\exp(-d/\xi)\Big)\ , \tag{F10}$$

where

$$\begin{aligned} F(u) &= -2p(1-p)(1-u)\log(1-u) - p\big(p+(1-p)u\big)\log\Big(1+\frac{(1-p)u}{p}\Big) \\ &- (1-p)\big(1-p+p\,u\big)\log\Big(1+\frac{p\,u}{1-p}\Big)\ . \end{aligned} \tag{F11}$$

To obtain the expression (F11) for F, we have used formula (26) for the 2-spin cluster-entropy, with $p_1 = p_2 = p$ and $p_{12} = p^2 + c_{12}$, where the correlation c_{12} is given by (F4). Note that $F(u) = O(u^2)$ for small u, in agreement with scaling (41).

b. Case $S_0 = S_{MF}$

We now introduce the reference entropy $S_0 = S_{MF}$. The matrix M defined in (21) has elements

$$M_{ij} = \frac{c_{ij}}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} = \exp(-|i-j|/\xi)\ . \tag{F12}$$

The inverse of M, $G = M^{-1}$, is a tridiagonal matrix, whose non zero elements are

$$G_{ii} = \frac{1+\exp(-2/\xi)}{1-\exp(-2/\xi)}\ ,\quad G_{i,i\pm1} = -\frac{\exp(-1/\xi)}{1-\exp(-2/\xi)}\ . \tag{F13}$$

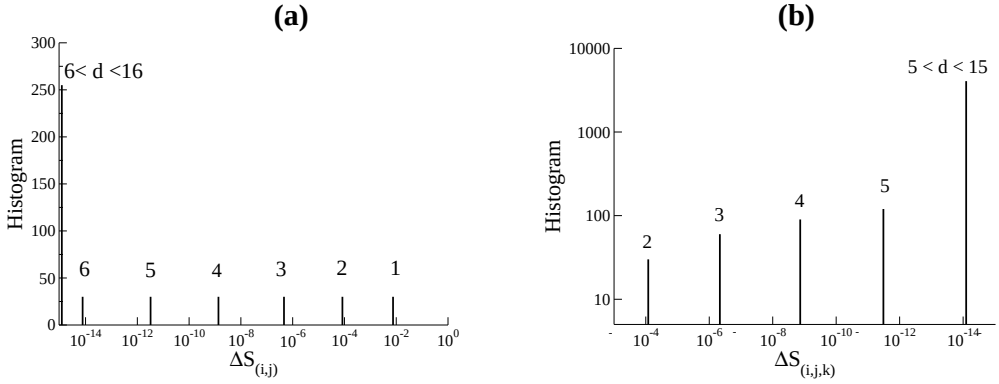


FIG. 36: Histograms of the entropies for clusters of size 2 **(a)** and 3 **(b)**; in the latter case, entropies are negative. Data are generated from the unidimensional Ising model (F1) with $N = 30$ spins, and parameters $J = 4$ and $h = -6$. Each peak is labelled by the distance d between the extremities of the clusters. The reference entropy is $S_0 = S_{MF}$.

Consider now the Gaussian model over N real-values variables φ_i, whose energy function is given by

$$E[\boldsymbol{\varphi}] = \frac{1}{2} \sum_{i,j} G_{ij}\, \varphi_i\, \varphi_j \ . \tag{F14}$$

For this Gaussian model, the logarithm of the partition function is (up an irrelevant additional constant), $\log Z[\mathbf{G}] = -\frac{1}{2} \log \det G$. By construction, model (F14) is the solution of the inverse Gaussian problem, with data: $\langle \varphi_i \rangle = 0, \langle \varphi_i\, \varphi_j \rangle = M_{ij}$. Hence, S_0 can be interpreted as the cross-entropy of Gaussian model (F14) under those data. A key feature of the Gaussian model above is that its interaction matrix G_{ij} is tridiagonal. Only nearest neighbour variables are coupled to each other according to (F13). We conclude that the Gaussian model is a one-dimensional model. Consequently, it obeys a chain rule similar to (F5). This is the only requirement for the main conclusion of Section F 1 a to hold: in the cluster expansion of S_0, the entropy of a K–cluster is simply equal to the entropy of the 2-cluster with the same extremities, multiplied by $(-1)^{K-2}$. As both the expansions of S and the one of S_0 enjoy this property, so does the expansion of $S - S_0$.

We conclude this Section by the expression of the 2-cluster entropy $\Delta S_{(i,i+d)}$. In the presence of the reference entropy $S_0 = S_{MF}$, we substract the following contribution to expression (F10), see (21),

$$\left(\Delta S_0\right)_{(i,i+d)} = \frac{1}{2} \log \det \begin{pmatrix} 1 & M_{i,i+d} \\ M_{i,i+d} & 1 \end{pmatrix} \ . \tag{F15}$$

Hence, function $F(u)$ defined in (F11) should be substracted $\frac{1}{2}\log(1-u^2)$. It is a simple check that $F(u) - \frac{1}{2}\log(1-u^2) = O(u^3)$, in agreement with scaling (46).

2. Examples and calculation of diagrammatic coefficients

We now show the histograms of cluster-entropies for $K = 2$ and $K = 3$ for specific choices of J, h. The averages p_i and p_{ij} were calculated exactly through formulas (F3) and (F4) (perfect sampling). Figure 36(a) shows the histogram of entropies for clusters of the type $(i, i+d)$. Entropy values are discrete and labelled by the distance d. They range from 10^{-2} (for nearest neighbours, distance $d = 1$) to values smaller than 10^{-15} for $6 < d < 15$. All entropies smaller than the numerical accuracy $\simeq 10^{-15}$ are put in the peak at the origin. Expanding $F(u)$ to the lowest order in u (for $S_0 = S_{MF}$) we find the asymptotic formula for the 2-cluster entropy:

$$\Delta S_{i,i+d} \simeq \frac{(2p-1)^2}{6\,p\,(1-p)}\, e^{-3\,d/\xi} \ , \tag{F16}$$

in agreement with (C1). We have verified that this formula is in very good agreement with the numerics as soon as $d \geq 4$ for the parameters of Fig. 36.

Figure 36(b) show the histogram of the entropies of 3-clusters (i, j, k). Let $d = k - i$ be the distance between the extremities. We observe that the entropies are gathered into peaks, and are exactly the opposite of the ones found in Fig. 36(a) as expected. Two differences are:

- The peak in $d = 1$ is not present because the minimal distance between three spins is $d = 2$. The largest 3-spins entropy thus corresponds to triplets of the type $(i, i+1, i+2)$.
- The height of the peak (number of clusters) corresponding to distance d is $(d-1)\,N$. The degeneracy $(d-1)$ is the number of ways of choosing the location of site i_2 in between i_1 and i_3.

We now show how the value of the cluster entropy can be found back from the leading terms in the diagrammatic expansion calculated in Section C 2. Let us call $d' = j - i < d$ the distance between the first two sites in the cluster. For each diagram in Fig. 33 we give in Table F 2 the sum of the distances of its links, *i.e.* the power of $\exp(-1/\xi)$.

diagram	sum of distances
a)	$d + 2d' + 2(d - d') = 3d$
b)	$2d + 2d' + (d - d') = 3d + d'$
c)	$2d + d' + 2(d - d') = 4d - d'$
d)	$3d' + 3(d - d') = 3d$
e)	$3d + 3d'$
f)	$3d + 3(d - d') = 6d - 3d'$

Interestingly, the lowest total distances are found in diagrams a) and d), while the latter diagram is of a higher power (6) in terms of the correlated function than the former (5). Hence, contrary to the case of independent spins (Section E), diagrams a) and d) give the dominant contributions to the entropy. Summing the contributions of a) and d) we find

$$\Delta S_{(i,j,k)} = \alpha^{(5)}_{ijk}\,(c_{ij})^2\,(c_{jk})^2\,c_{ki} + \alpha^{(6)}_{ijk}\,(c_{ij})^3\,(c_{jk})^3 = \left(\alpha^{(5)}_{ijk} + \alpha^{(6)}_{ijk}\right)\,\left(p(1-p)\,\exp(-1/\xi)\right)^{3d}\,. \tag{F17}$$

To derive the coefficients $\alpha^{(5)}$ and $\alpha^{(6)}$, we impose that $\Delta S_{(i,j,k)}$ is the opposite of (F16). We deduce that $\alpha^{(5)}$ and $\alpha^{(6)}$ are given by, respectively, (C4) and (C6).

The exact cancellation property discussed above has important consequences for the inferred fields and couplings. Consider for instance the coupling $J_{i,i+2}$, which vanishes in the 1D-Ising model with nearest-neighbour interactions (F1). As the connected correlation $c_{i,i+2}$ is not equal to zero, a contribution to the coupling will be collected from the cluster $(i, i+2)$ itself, equal to

$$\Delta J_{i,i+2;(i,i+2)} = -\frac{\partial \Delta S_{(i,i+2)}}{\partial p_{i,i+2}}\,. \tag{F18}$$

Other contributions will come from larger clusters. For instance the cluster $(i, i+1, i+2)$ will give an additional

$$\Delta J_{i,i+2;(i,i+1,i+2)} = -\frac{\partial \Delta S_{(i,i+1,i+2)}}{\partial p_{i,i+2}}\,. \tag{F19}$$

The sum of the two contributions above vanishes due to the cancellation property. It can be checked that the contributions coming from all the other clusters vanish, too, which makes the coupling $J_{i,i+2} = 0$ as it should.

Appendix G: Inverse susceptibility matrix for the unidimensional Ising model

Hereafter, we want to invert the matrix $\boldsymbol{\chi}$, whose elements are given in (61). The matrix is of dimension $\frac{1}{2}N(N-1)$, and each element is labelled by two indices (i, j) and (k, l), with $i < j$ and $k < l$. Each index (i, j) can be represented by a site of coordinates i and j on the half-grid of Fig. 37(a). We now show that the non-zero entries of the inverse susceptibility matrix, $\left(\chi^{-1}\right)_{ij,kl}$, are in one-to-one correspondence with the sites (i, j) and (k, l) that are either identical, or nearest neighbours, or diagonally opposed on the elementary mesh of the half-grid (Fig. 37(b,c,d)). Depending on the value of the difference $j - i$, the number of those sites is equal to 9, 8, or 6.

We start with the case $j - i \geq 3$ (Fig. 37(b)). By symmetry, the nine unknown matrix elements $\left(\chi^{-1}\right)_{ij,kl}$ take only three independent values, denoted by γ for $(k, l) = (i, j)$, β for (k, l) and (i, j) nearest neighbours, and α for $(k, l) = (i \pm 1, j \pm 1)$. We now write the matrix inversion identity,

$$\sum_{k<l} \left(\chi^{-1}\right)_{ij,kl}\,\chi_{kl,mn} = \delta_{i,m}\,\delta_{j,n}\,, \tag{G1}$$

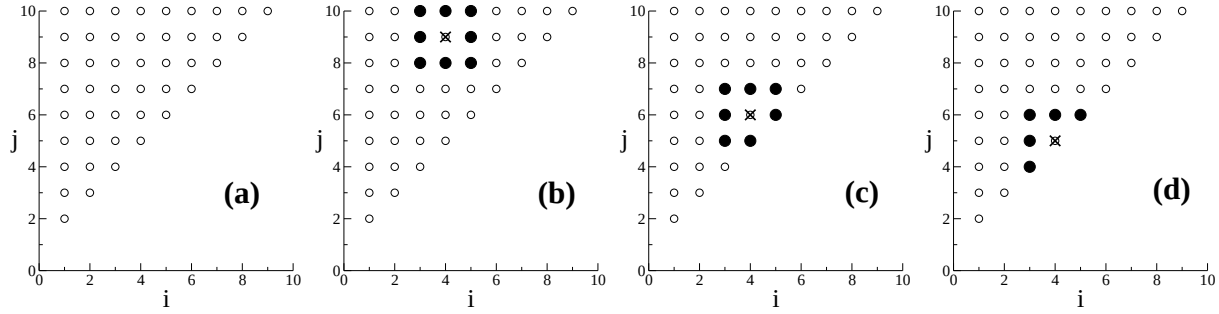


FIG. 37: Half grid representing the index (i, j) of the entries of the inverse susceptibility matrix, with $i < j$ **(a)**. Black circles locate the nearest-neighbours and the diagonally opposed sites (k, l) of (i, j) (cross), with $i = 4$ and $j = 9$ **(b)**, 6 **(c)**, 5 **(d)**.

for various values of (m, n). Let $d = j - i$. For $m = i, n = j$, constraint (G1) gives

$$\gamma(1 - x^{2d}) + 2\beta\big(2x - x^{2d-1} - x^{2d+1}\big) + \alpha\big(4x^2 - x^{2d-2} - x^{2d} - x^{2d+2}\big) = 1 \ , \tag{G2}$$

which should hold for all $d \geq 3$. We deduce two coupled equations for the three unknown variables:

$$\gamma + 2\left(x + \frac{1}{x}\right)\beta + 4\left(x + \frac{1}{x}\right)^2 \alpha \ = \ 0 \ , \tag{G3}$$

$$\gamma + 4\,x\,\beta + 4\,x^2\,\alpha \ = \ 1 \ . \tag{G4}$$

For $m = i + 1, n = j$, constraint (G1) is equivalent to

$$\gamma(x - x^{2d-1}) + \beta\big(1 + 3x^2 - x^{2d-2} - 3x^{2d}\big) + \alpha\big(2x + 2x^3 - x^{2d-3} - 2x^{2d-1} - x^{2d+1}\big) = 0 \ . \tag{G5}$$

The d-dependent term in the equation above cancels by virtue of (G3). We are left with an additional equation over α, β, γ:

$$\gamma\,x + \beta\,(1 + 3x^2) + 2\,x\,(1 + x^2)\,\alpha = 0 \ . \tag{G6}$$

By symmetry of the matrices $\boldsymbol{\chi}, \boldsymbol{\chi}^{-1}$, no new constraint is obtained when the values of m, n are further varied. Solving (G3), (G4), (G6) we obtain

$$\alpha = \frac{x^2}{(1 - x^2)^2} \ , \quad \beta = -\frac{x(1 + x^2)}{(1 - x^2)^2} \ , \quad \gamma = \frac{(1 + x^2)^2}{(1 - x^2)^2} \ . \tag{G7}$$

The analysis of the other cases $j = i + 2$ (Fig. 37(c)) and $j = i + 1$ (Fig. 37(d)) can be done along the same lines. We do not write the calculations in details, and simply report the results. The case $j = i + 2$ is very similar to the previous case. There are 8 coefficients to be calculated, with three independent values, α', β', γ'. It turns out that

$$\alpha' = \alpha \ , \quad \beta' = \beta \ , \quad \gamma' = \gamma \ . \tag{G8}$$

As for the last case, $j = i+1$, we call α'' the values of the entries of $\boldsymbol{\chi}^{-1}$ with $(k, l) = (i-1, j-1), (i-1, j+1), (i+1, j+1)$, β'' the values of the entries with $(k, l) = (i-1, j), (i, j+1)$, and γ'' the diagonal element corresponding to $(k, l) = (i, j)$. After some elementary algebra, we find

$$\alpha'' = \alpha \ , \quad \beta'' = \beta \ , \quad \gamma'' = \frac{1 + x^2 + x^4}{(1 - x^2)^2} \ . \tag{G9}$$

All those results are reported in (62).

---

[1] S.G. Brush, *Rev. Mod. Phys.* **39**, 883 (1967).

[2] E. Schneidman, M.J. Berry II, R. Segev, W. Bialek, *Nature* **440**, 1007 (2006); G. Tkacik, E. Schneidman, M.J. Berry II, W. Bialek, *arXiv:q-Bio.NC*, 0611072 (2006).
[3] O. Marre, S. El Boustani, Y. Frégnac, A. Destexhe, *Phys. Rev. Lett.* **102**, 138101 (2009)
[4] A. Peyrache *et al.*, *Nature Neurosci.* **12**, 919 (2009).
[5] M. Weigt *et al.*, *Proc. Nat. Acad. Sci.* **106**, 67 (2009).
[6] S. Balakrishnan *et al.*, *Proteins* **79**, 1061 (2011).
[7] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, A.M. Walczak, *arxiv 1107.0604* (2011).
[8] S. Cocco, R. Monasson, *Phys. Rev. Lett.* **106**, 090601 (2011).
[9] E.T. Jaynes, *Proc. IEEE* **70**, 939 (1982).
[10] S. Cocco, S. Leibler, R. Monasson, *Proc. Nat. Acad. Sci.* **106**, 14058 (2009).
[11] M. Opper, D. Saad (eds), *Advanced Mean Field Methods: Theory and Practice*, MIT Press (2001).
[12] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, *Cognitive Science* **9**, 147 (1985).
[13] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer (2009).
[14] C.S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, Springer-Verlag (Information Science and Statistics) (2005).
[15] P. Ravikumar, M.J. Wainwright, J. Lafferty, *Annals of Statistics* **38**, 1287 (2010).
[16] J. Bento, A. Montanari ,Which graphical models are difficult to learn?, NIPS (2009).
[17] M. Jerrum, A. Sinclair, The Markov chain Monte Carlo method: an approach to approximate counting and integration in *Approximation Algorithms for NP-hard Problems*, D.S.Hochbaum ed., PWS Publishing, Boston (1996).
[18] E. Cagliotti, T. Kuna, J.L. Lebowitz, E.R. Speer, *Markov Processes Relat. Fields* **12**, 257 (2006); T. Kuna, J.L. Lebowitz, E.R. Speer, *J. Stat. Phys.* **129**, 417 (2007).
[19] R.H. Swendsen, Phys. Rev. Lett.**52**, 1165 (1984).
[20] N. Meinshausen, P. Bühlmann, *Ann. Statist.* **34**, 1436 (2006).
[21] E. Aurell, M. Ekeberg, *arXiv:1107.3536* (2011).
[22] Y. Roudi, J.Tyrcha, J. Hertz, *Phys. Rev. E* **79**, 051915 (2009).
[23] T. Plefka, *J. Phys. A: Math. Gen.* **15**, 1971 (1982).
[24] A. Georges, J. Yedidia, *J. Phys. A: Math. Gen.* **24**, 2173 (1991).
[25] A. Georges, Lectures on the Physics of Highly Correlated Electron Systems VIII: 8th Training Course in the Physics Correlated Electron Systems and High-Tc Superconductors **715**, 3 (2004)
[26] D.J. Thouless, P.W. Anderson, R.G. Palmer, *Phil. Mag.* **35** 593 (1977).
[27] D. Sherrington, S. Kirkpatrick, *Phys. Rev. Lett.* **35**, 1792 (1975).
[28] T. Tanaka, *Phys Rev E* **58**, 2302 (1998).
[29] Y. Roudi, E. Aurell, J. Hertz, *Front. Comput. Neurosci.* **3**, 22 (2009)
[30] V. Sessak, R. Monasson *J. Phys. A* **42**, 055001 (2009).
[31] A. Pelizzola, *J. Phys. A* **38**, R 309 (2005).
[32] M. Mézard, T. Mora, *J. Physiol. Paris* **103**, 107 (2009); E. Marinari, V. Van Kerrebroeck, *J. Stat. Mech.* P02008 (2010).
[33] C. de Dominicis, J. Math. Phys., 3, 983 (1962).
[34] J.P. Hansen, I. R. McDonald Theory of Simple Liquids (New York: Academic) (1976).
[35] The Equilibrium Theory of Classical Fluids (A lecture note and reprint volume) (With H. Frisch), H. L. Frisch and J. L. Lebowitz. Benjamin, New York, (1964).
[36] P. Flajolet, X. Gourdon, P. Dumas, *Theoretical Computer Science* **144**, 3 (1995).
[37] M.J. Schnitzer, M. Meister, *Neuron* **37**, 499-511 (2003).
[38] E.J. Gumbel, *Statistics of Extremes*, Dover (2004)
[39] D.J.C. MacKay, *Neural Computation* **4**, 415 (1991).
[40] A.J. Bray, M.A. Moore, *J. Phys A* **10**, 1927 (1977).
[41] I.M. Johnstone, *Proc. ICM 2006* **1**, 307 (2006).
[42] J.K. Percus, G.J. Yevick, *Phys. Rev.***110**, 1 (1958).
[43] C. Borzi, G. Ord, J.K. Percus, *J. Stat. Phys.* **46**, 51 (1986).
[44] J. K. Percus, L. Šamaj, *J. Stat. Phys.* **77**, 421 (1993).
[45] J. Barton, *private communication.*
[46] M. Fisher, *J. Math. Phys. (N.Y.)* **5**, 944 (1964).
[47] M. Fisher, *Phys Rev* **162**, 480 (1967).
[48] D. Zobin, *Phys Rev* **5**, 2387 (1978).
[49] In the present work where spins are equal to 0,1, the couplings (J_{ij}) and fields (h_i) are the one's in spin 0,1 ($\hat{J}_{ij}, \hat{h}_i$) by the transformation: $J_{ij} = 4\hat{J}_{i,j}$ and $h_i = -\frac{1}{2}\sum_{j\neq i} J_{ij} + \hat{h}_i$. The numerical experiments of [16] where done with ± 1 spins.
[50] We have to minimize here rather than maximize since the true Lagrange multipliers take imaginary values, the couplings and fields being their imaginary part.
[51] The vocable 'field', should strictly speaking, be used when the variables σ_i are spins taking ± 1 values. For $0, 1$ variable, the use of the denomination 'chemical potential' would be more appropriate. We keep to the simpler denomination 'field' hereafter.
[52] The minimal strength of the couplings which can be 'detected' depend on the quality of the sampling, and scales as $\sqrt{\log N/B}$, see Section IV B 2.

[53] In mean-field spin-glasses, as the couplings scale as the inverse square root of the number N of spins, only loop diagrams have non-zero weights in the thermodynamical limit.
[54] A further theoretical argument supporting the existence of the cancellation property is, in the case of perfect sampling, the fact that the entropy S must be extensive in N. As S is the sum of $\sim 2^N$ cluster entropies, those contributions must compensate each other.
[55] If $p = \frac{1}{2}$, ΔS_{IS} is of the order of B^{-K}
[56] This statement is widely believed to be true in the theory of liquids literature. The fast decay of the inverse susceptibility, $\chi^{-1}(r)$, or, equivalently, of the direct pair correlation, $g(r)$, is used to approximately close the hierarchy of correlation functions. The Percus-Yevik closure scheme, which gives an accurate equation of state for liquids of hard spheres, assumes that the inverse susceptibility vanishes above the interaction range of the potential (diameter of a particle).
[57] We have verified that the dependence on N is weaker in the case of periodic boundary conditions.
[58] The numerical experiments of [16] were done with ± 1 spins and with coupling parameter θ and field $\nu = 0$; in the present work where spins are equal to 0,1, the corresponding couplings and fields are: $J = 4\theta$ and $h_i = -\frac{1}{2}\sum_{j \neq i} J_{ij}$.