High-Dimensional Inference with the generalized Hopfield Model: Principal Component Analysis and Corrections

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

[1] *Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540, USA*

[2] *Laboratoire de Physique Statistique de l'Ecole Normale Supérieure, CNRS & Univ. Paris 6, Paris, France*

[3] *Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, CNRS & Univ. Paris 6, Paris, France*

We consider the problem of inferring the interactions between a set of N binary variables from the knowledge of their frequencies and pairwise correlations. The inference framework is based on the Hopfield model, a special case of the Ising model where the interaction matrix is defined through a set of patterns in the variable space, and is of rank much smaller than N. We show that Maximum Likelihood inference is deeply related to Principal Component Analysis when the amplitude of the pattern components, ξ, is negligible compared to $\sqrt{N}$. Using techniques from statistical mechanics, we calculate the corrections to the patterns to the first order in $\xi/\sqrt{N}$. We stress that it is important to generalize the Hopfield model and include both attractive and repulsive patterns, to correctly infer networks with sparse and strong interactions. We present a simple geometrical criterion to decide how many attractive and repulsive patterns should be considered as a function of the sampling noise. We moreover discuss how many sampled configurations are required for a good inference, as a function of the system size N and of the amplitude ξ. The inference approach is illustrated on synthetic and biological data.

I. INTRODUCTION

Understanding the patterns of correlations between the components of complex systems is a fundamental issue in various scientific fields, ranging from neurobiology to genomic, from finance to sociology, ... A recurrent problem is to distinguish between direct correlations, produced by physiological or functional interactions between the components, and network correlations, which are mediated by other, third-party components. Various approaches have been proposed to infer interactions from correlations, exploiting concepts related to statistical dimensional reduction [1], causality [2], the maximum entropy principle [3], Markov random fields [4] ... A major practical and theoretical difficulty in doing so is the paucity and the quality of data: reliable analysis should be able to unveil real patterns of interactions, even if measures are affected by under- or noisy sampling. The size of the interaction network can be comparable to or larger than the number of data, a situation referred to as high-dimensional inference.

The purpose of the present work is to establish a quantitative correspondence between two of those approaches, namely the inference of Boltzmann Machines (also called Ising model in statistical physics and undirected graphical models for discrete variables in statistical inference [4]) and Principal Component Analysis (PCA) [1]. Inverse Boltzmann Machines (BM) are a mathematically well-founded but computationally challenging approach to infer interactions from correlations. Our scope is to find the interactions among a set of N variables $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$. For simplicity, we consider variables σ_i taking binary values ± 1 only; the discussion below can be easily extended to the case of a larger number of values, *e.g.* to genomics where nucleotides are encoded by four-letter symbols, or to proteomics where amino-acids can take twenty values. Assume that the average values of the variables, $m_i = \langle \sigma_i \rangle$, and the pairwise correlations, $c_{ij} = \langle \sigma_i \sigma_j \rangle$ are measured, for instance, through the sampling of, say, B configurations $\boldsymbol{\sigma}^b, b = 1, \ldots, B$. Solving the inverse BM problem consists in finding the set of interactions, J_{ij}, and of local fields, h_i, defining an Ising model, such that the equilibrium magnetizations and pairwise correlations coincide with, respectively, m_i and c_{ij}. Many procedures have been designed to tackle this inverse problem, including learning algorithms [5], advanced mean-field techniques [6, 7], message-passing procedures [8, 9], cluster expansions [10, 11], graphical lasso [4] and its variants [12]. The performance (accuracy, running time) of those procedures depend on the structure of the underlying interaction network and on the quality of the sampling, *i.e.* how large B is.

Principal Component Analysis (PCA) is a widely popular tool in statistics to analyze the correlation structure of a set of variables $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$. The principle of PCA is simple. One starts with the correlation matrix,

$$\Gamma_{ij} = \frac{c_{ij} - m_i m_j}{\sqrt{(1 - m_i^2)\,(1 - m_j^2)}} \,, \tag{1}$$

which expresses the covariance between variables σ_i and σ_j, rescaled by the product of the expected fluctuations of the variables taken separately. Γ is then diagonalized. The projections of $\boldsymbol{\sigma}$ along the top eigenmodes (associated to the largest eigenvalues of Γ) identify the uncorrelated variables which contribute most to the total variance. If a few, say, p ($\ll N$), eigenvalues are notably larger than the remaining ones PCA achieves an important dimensional reduction. The determination of the number p of components to be retained is a delicate issue. It may be done by comparing the spectrum of Γ to the Marcenko-Pastur (MP) spectrum for the null hypothesis, that is, for the correlation matrix calculated from the sampling of B configurations of N independent

variables [13]. Generally those two spectra coincide when N is large, except for some large or small eigenvalues of Γ, retained as the relevant components.

The advantages of PCA are multiple, which explains its success. The method is very versatile and fast as it only requires to diagonalize the correlation matrix, which can be achieved in a time polynomial in the size N of the problem. In addition, PCA may be extended to incorporate *prior* information about the components, which is particularly helpful for processing noisy data. An illustration is sparse PCA, which looks for principal components with many vanishing entries [14].

In this paper we present a conceptual and practical framework which encompasses BM and PCA in a controlled way. We show that PCA, with appropriate modifications, can be used to infer BM and discuss in detail the amount of data necessary to do so. Our framework is based on an extension of a celebrated model of statistical mechanics, the Hopfield model [15]. The Hopfield model was originally introduced to model auto-associative memories, and relies on the notion of patterns [16]. Informally speaking, a pattern $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ defines an attractive direction in the N-dimensional space of the variable configurations, *i.e.* a direction along which $\boldsymbol{\sigma}$ has a tendency to align. The norm of $\boldsymbol{\xi}$ characterizes the strength of the attraction. While having only attractive patterns makes sense for auto-associative memories, it is an unnecessary assumption in the context of BM. We therefore generalize the Hopfield model by including repulsive patterns $\hat{\boldsymbol{\xi}}$, that is, directions in the N-dimensional space which $\boldsymbol{\sigma}$ tends to be orthogonal to [17]. From a technical point of view, the generalized Hopfield model with p attractive patterns and $\hat{p}$ repulsive patterns is simply a particular case of BM with an interaction matrix, $\mathbf{J}$, of rank equal to $p + \hat{p}$. If one knows *a priori* that the rank of the true $\mathbf{J}$ is indeed small, *i.e.* $p + \hat{p} \ll N$, using the generalized Hopfield model rather than a generic BM allows one to infer much less parameters and to avoid overfitting in the presence of noisy data.

We first consider the case where the components ξ_i and $\hat{\xi}_i$ are very small compared to $\sqrt{N}$. In this limit case we show that Maximum Likelihood (ML) inference with the generalized Hopfield model is closely related to PCA. The attractive patterns are in one-to-one correspondence with the largest components of the correlation matrix, while the repulsive patterns correspond to the smallest components, which are normally discarded by PCA. When all patterns are selected ($p + \hat{p} = N$) inference with the generalized Hopfield model is equivalent to the mean-field approximation [6]. Retaining only few significative components helps, in principle, to remove noise from the data. We present a simple geometrical criterion to decide in practice how many attractive and repulsive patterns should be considered. We also address the question of how many samples (B) are required for the inference to be meaningful. We calculate the error bars over the patterns due to the the finite sampling. We then analyze the case where the data are sampled from a generalized Hopfield model, and inference amounts to learn the patterns of that model. When the system size, N, and the number of samples, B, are both sent to infinity with a fixed ratio, $\alpha = \frac{B}{N}$, there is a critical value of the ratio, α_c, below which learning is not possible. The value of α_c depends on the amplitude of the pattern components. This transition corresponds to the retarded learning phenomenon discovered in the context of supervised learning with continuous variables and rigorously studied in random matrix and probability theories, see [13, 18, 19] for reviews. We validate our findings on synthetic data generated from various Ising models with known interactions, and present applications to neurobiological and proteomic data.

In the case of a small system size, N, or of very strong components, $\xi_i, \hat{\xi}_i$, the ML patterns do not coincide with the components identified by PCA. We make use of techniques from the statistical mechanics of disordered systems originally intended to calculate averages over ensembles of matrices to compute the likelihood to the second order in powers of $\frac{\xi_i}{\sqrt{N}}$ for a given correlation matrix. We give explicit expressions for the ML patterns in terms of non-linear combinations of the eigenvalues and eigenvectors of the correlation matrix. These corrections are validated on synthetic data. Furthermore, we discuss the issue of how many sampled configurations are necessary to improve over the leading–order ML patterns as a function of the amplitude of the pattern components and of the system size.

The plan of the paper is as follows. In Section II we define the generalized Hopfield model, the Bayesian inference framework and list our main results, that is, the expressions of the patterns without and with corrections, the criterion to decide the number of patterns, and the expressions for the error bars on the inferred patterns. Tests on synthetic data are presented in Section III. Section IV is devoted to the applications to real biological data, *i.e* recordings of the neocortical activity of a behaving rat and consensus multi-sequence alignment of the PDZ protein domain family. Readers interested in applying our results rather than in their derivation need not read the subsequent sections. Derivation of the log-likelihood with the generalized Hopfield model and of the main inference formulae can be found in Section V. In Section VI we study the minimal number B of samples necessary to achieve an accurate inference, and how this number depends on the number of patterns and on their amplitude. Perspectives and conclusions are given in Section VII.

II. DEFINITIONS AND MAIN RESULTS

A. Generalized Hopfield Model

We consider configurations $\boldsymbol{\sigma} = \{\sigma_1, , \sigma_2, \ldots, \sigma_N\}$ of N binary variables taking values $\sigma_i = \pm 1$, drawn according

to the probability

$$P_H[\boldsymbol{\sigma}|\mathbf{h},\{\boldsymbol{\xi}^\mu\},\{\hat{\boldsymbol{\xi}}^\mu\}] = \frac{\exp -E[\boldsymbol{\sigma},\mathbf{h},\{\boldsymbol{\xi}^\mu\},\{\hat{\boldsymbol{\xi}}^\mu\}]}{Z[\mathbf{h},\{\boldsymbol{\xi}^\mu\},\{\hat{\boldsymbol{\xi}}^\mu\}]} \ , \quad (2)$$

where the energy E is given by

$$E[\boldsymbol{\sigma},\mathbf{h},\{\boldsymbol{\xi}^\mu\},\{\hat{\boldsymbol{\xi}}^\mu\}] \ = \ -\sum_{i=1}^{N} h_i\sigma_i - \frac{1}{2N}\sum_{\mu=1}^{p}\left(\sum_{i=1}^{N}\xi_i^\mu\sigma_i\right)^2 + \ \frac{1}{2N}\sum_{\mu=1}^{\hat{p}}\left(\sum_{i=1}^{N}\hat{\xi}_i^\mu\sigma_i\right)^2 \ . \quad (3)$$

The partition function Z in (2) ensures the normalization of P_H. The components of $\mathbf{h} = (h_1, h_2, .., h_N)$ are the local fields acting on the variables. The patterns $\boldsymbol{\xi}^\mu = \{\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu\}$, with $\mu = 1, 2, \ldots, p$, are attractive patterns: they define preferred directions in the configuration space $\boldsymbol{\sigma}$, along which the energy E decreases (if the fields are weak enough). The patterns $\hat{\boldsymbol{\xi}}^\mu$, with $\mu = 1, 2, \ldots, \hat{p}$, are repulsive patterns: configurations $\boldsymbol{\sigma}$ aligned along those directions have a larger energy. The pattern components, $\xi_i^\mu, \hat{\xi}_i^\mu$, and the fields, h_i, are real-valued. Our model is a generalized version of the original Hopfield model [15], which has only attractive patterns and corresponds to $\hat{p} = 0$. In the following, we will assume that $p + \hat{p}$ is much smaller than N.

Energy function (3) implicitly defines the coupling J_{ij} between the variables σ_i and σ_j,

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\xi_i^\mu\,\xi_j^\mu - \frac{1}{N}\sum_{\mu=1}^{\hat{p}}\hat{\xi}_i^\mu\,\hat{\xi}_j^\mu \ . \quad (4)$$

Note that any interaction matrix J_{ij} can be written under the form (4), with p and $\hat{p}$ being, respectively, the number of positive and negative eigenvalues of J. Here, we assume that the total number of patterns, $p + \hat{p}$, *i.e.* the rank of the matrix J is (much) smaller than the system size, N.

The data to be analyzed consists of a set of B configurations of the N spins, $\boldsymbol{\sigma}^b$, $b = 1, \ldots, B$. We assume that those configurations are drawn, independently from each other, from the distribution P_H (2). The parameters defining P_H, that is, the fields $\mathbf{h}$ and the patterns $\{\boldsymbol{\xi}^\mu\}, \{\hat{\boldsymbol{\xi}}^\mu\}$ are unknown. Our scope is to determine the most likely values for those fields and patterns from the data. In Bayes inference framework the posterior distribution for the fields and the patterns given the data $\{\boldsymbol{\sigma}^b\}$ is

$$P[\mathbf{h},\{\boldsymbol{\xi}^\mu\},\{\hat{\boldsymbol{\xi}}^\mu\}|\{\boldsymbol{\sigma}^b\}] \ = \ \frac{P_0[\mathbf{h},\{\boldsymbol{\xi}^\mu\},\{\hat{\boldsymbol{\xi}}^\mu\}]}{P_1[\{\boldsymbol{\sigma}^b\}]} \times \prod_{b=1}^{B} P_H[\boldsymbol{\sigma}^b|\mathbf{h},\{\boldsymbol{\xi}^\mu\},\{\hat{\boldsymbol{\xi}}^\mu\}] \ , \quad (5)$$

where P_0 encodes some *a priori* information over the parameters to be inferred and P_1 is a normalization.

It is important to realize that many transformations affecting the patterns can actually leave the coupling matrix $\mathbf{J}$ (4) and the distribution P_H unchanged. A simple example is given by an orthogonal transformation $\mathcal{O}$ over the attractive patterns : $\xi_i^\mu \to \bar{\xi}_i^\mu = \sum_\nu \mathcal{O}^{\mu\nu}\xi_i^\nu$. This invariance entails that the the problem of inferring the patterns is not statistically consistent: even with an infinite number of sampled data no inference procedure can distinguish between a Hopfield model with patterns $\{\boldsymbol{\xi}^\mu\}$ and another one with patterns $\{\bar{\boldsymbol{\xi}}^\mu\}$. However, the inference of the couplings is statistically consistent: two distinct matrices $\mathbf{J}$ define two distinct distributions over the data.

In the presence of repulsive patterns the complete invariance group is the indefinite orthogonal group $O(p, \hat{p})$, which has $\frac{1}{2}(p + \hat{p})(p + \hat{p} - 1)$ generators. To select one particular set of most likely patterns, we explicitly break the invariance through P_0. A convenient choice we use throughout this paper is to impose that the weighted dot products of the pairs of attractive and/or repulsive patterns vanish:

$$\begin{aligned}
\sum_i \xi_i^\mu\xi_i^\nu(1-m_i^2) \ &= \ 0 \quad \left[\frac{1}{2}p(p-1) \text{ constraints}\right] , \\
\sum_i \xi_i^\mu\hat{\xi}_i^\nu(1-m_i^2) \ &= \ 0 \quad \left[p\hat{p} \text{ constraints}\right] , \\
\sum_i \hat{\xi}_i^\mu\hat{\xi}_i^\nu(1-m_i^2) \ &= \ 0 \quad \left[\frac{1}{2}\hat{p}(\hat{p}-1) \text{ constraints}\right] .
\end{aligned} \quad (6)$$

In the following we will use the vocable Maximum Likelihood inference to refer to the case where the prior P_0 is used to break the invariance only. P_0 may also be chosen to impose specific constraints on the pattern amplitude, see Section II E devoted to regularization.

B. Maximum Likelihood Inference: lowest order

Due to the absence of three- or higher order-body interactions in E (3), P depends on the data $\{\boldsymbol{\sigma}^b\}$ only through the N magnetizations, m_i, and the $\frac{1}{2}N(N-1)$ two-spin covariances, c_{ij}, of the sampled data:

$$m_i = \frac{1}{B}\sum_b \sigma_i^b \quad , \qquad c_{ij} = \frac{1}{B}\sum_b \sigma_i^b\,\sigma_j^b \ . \quad (7)$$

We consider the correlation matrix Γ (1), and call $\lambda^1 \geq \ldots \geq \lambda^k \geq \lambda^{k+1} \geq \ldots \geq \lambda^N$ its eigenvalues. $\mathbf{v}^k$ denotes the eigenvector attached to λ^k and normalized to unity. We also introduce another notation to label the same eigenvalues and eigenvectors in the reverse order: $\hat{\lambda}^k \equiv \lambda^{N+1-k}$ and $\hat{\mathbf{v}}^k = \mathbf{v}^{N+1-k}$, *e.g.* $\hat{\lambda}^1$ is the smallest eigenvalue of Γ; the motivation for doing so will be transparent below. Note that Γ is, by construction, a semi-definite positive matrix: all its eigenvalues are positive. In addition, the sum of the eigenvalues is equal to N since $\Gamma_{ii} = 1, \forall i$. Hence the largest and smallest

eigenvalues are guaranteed to be, respectively, larger and smaller than unity.

In the following Greek indices, *i.e.* μ, ν, ρ, correspond to integers comprised between 1 and p or $\hat{p}$, while roman letters, *i.e.* i, j, k denote integers ranging from 1 to N.

Finding the patterns and fields maximizing P (5) is a very hard computational task. We introduce an approximation scheme for those parameters

$$\begin{aligned} \xi_i^\mu &= (\xi^0)_i^\mu + (\xi^1)_i^\mu + \dots \ , \\ \hat{\xi}_i^\mu &= (\hat{\xi}^0)_i^\mu + (\hat{\xi}^1)_i^\mu + \dots \ , \\ h_i &= (h^0)_i + (h^1)_i + \dots \ . \end{aligned} \tag{8}$$

The derivation of this systematic approximation scheme and the discussion of how smaller the contributions get with the order of the approximation can be found in Section V A. To the lowest order the patterns are given by

$$\begin{aligned} (\xi^0)_i^\mu &= \sqrt{N\left(1-\frac{1}{\lambda^\mu}\right)}\,\frac{v_i^\mu}{\sqrt{1-m_i^2}} \quad (1 \le \mu \le p) \\ (\hat{\xi}^0)_i^\mu &= \sqrt{N\left(\frac{1}{\hat{\lambda}^\mu}-1\right)}\,\frac{\hat{v}_i^\mu}{\sqrt{1-m_i^2}} \quad (1 \le \mu \le \hat{p}) \ . \end{aligned} \tag{9}$$

The above expressions require that $\lambda^\mu > 1$ for an attractive pattern and $\hat{\lambda}^\mu < 1$ for a repulsive pattern. Once the patterns are computed the interactions, $(J^0)_{ij}$, can be calculated from (4),

$$\begin{aligned} (J^0)_{ij} &= \frac{1}{\sqrt{(1-m_i^2)(1-m_j^2)}}\left(\sum_{\mu=1}^{p}\left(1-\frac{1}{\lambda^\mu}\right)v_i^\mu\,v_j^\mu \right. \\ &- \left. \sum_{\mu=1}^{\hat{p}}\left(\frac{1}{\hat{\lambda}^\mu}-1\right)\hat{v}_i^\mu\,\hat{v}_j^\mu\right) \ . \end{aligned} \tag{10}$$

The values of the local fields are then obtained from

$$(h^0)_i = \tanh^{-1} m_i - \sum_j (J^0)_{ij}\, m_j \ , \tag{11}$$

which has a straightforward mean-field interpretation.

The above results are reminiscent of PCA, but differ in several significative aspects. First, the patterns do not coincide with the eigenvectors due to the presence of m_i-dependent terms. Secondly, the presence of the λ^μ-dependent factor in (9) discounts the patterns corresponding to eigenvalues close to unity. This effect is easy to understand in the limit case of independent spins and perfect sampling ($B \to \infty$): Γ is the identity matrix, which gives $\lambda^\mu = 1, \forall \mu$, and the patterns rightly vanish. Thirdly, and most importantly, not only the largest but also the smallest eigenmodes must be taken into account to calculate the interactions.

The couplings J^0 (10) calculated from the lowest-order approximation for the patterns are closely related to the mean-field (MF) interactions [6],

$$J_{ij}^{MF} = -\frac{(\Gamma^{-1})_{ij}}{\sqrt{(1-m_i^2)(1-m_j^2)}} \ , \tag{12}$$

where Γ^{-1} denotes the inverse matrix of Γ (1). However, while all the eigenmodes of Γ are taken into account in the MF interactions (12), our lowest-order interactions (10) include contributions from the p largest and the $\hat{p}$ smallest eigenmodes only. As the values of $p, \hat{p}$ can be chosen depending on the number of available data, the generalized Hopfield interactions (10) is *a priori* less sensitive to overfitting. In particular, it is possible to avoid considering the bulk part of the spectrum of Γ, which is essentially due to undersampling ([13] and Section VI B 2).

C. Sampling error bars on the patterns

The posterior distribution P can locally be approximated with a Gaussian distribution centered in the most likely values for the patterns, $\{(\boldsymbol{\xi}^0)^\mu\}, \{(\hat{\boldsymbol{\xi}}^0)^\mu\}$, and the fields, $\mathbf{h}^0$. We obtain the covariance matrix of the fluctuations of the patterns around their most likely values,

$$\langle \Delta\xi_i^\mu\, \Delta\xi_j^\nu \rangle = \frac{N\left[\mathbf{M}_{\xi\xi}\right]_{ij}^{\mu\nu}}{B\sqrt{(1-m_i^2)(1-m_j^2)}} \ . \tag{13}$$

and identical expressions for $\langle \Delta\xi_i^\mu\, \Delta\hat{\xi}_j^\nu \rangle$ and $\langle \Delta\hat{\xi}_i^\mu\, \Delta\hat{\xi}_j^\nu \rangle$ upon substitution of $\left[\mathbf{M}_{\xi\xi}\right]_{ij}^{\mu\nu}$ with, respectively, $\left[\mathbf{M}_{\xi\hat{\xi}}\right]_{ij}^{\mu\nu}$ and $\left[\mathbf{M}_{\hat{\xi}\hat{\xi}}\right]_{ij}^{\mu\nu}$. The entries of the $\mathbf{M}$ matrices are

$$\begin{aligned} \left[\mathbf{M}_{\xi\xi}\right]_{ij}^{\mu\nu} &= \delta^{\mu\nu}\left[\sum_{k=p+1}^{N-\hat{p}}\frac{v_i^k\,v_j^k}{|\lambda^k-\hat{\lambda}^\mu|} + \sum_{\rho=1}^{p}\frac{|\lambda^\mu-1|\lambda^\rho\,v_i^\rho\,v_j^\rho}{G_1(\lambda^\rho,\lambda^\mu)}\right. \\ &+ \left. \sum_{\rho=1}^{\hat{p}}\frac{|\lambda^\mu-1|\hat{\lambda}^\rho\,\hat{v}_i^\rho\,\hat{v}_j^\rho}{G_1(\hat{\lambda}^\rho,\lambda^\mu)}\right] + \frac{G_2(\lambda^\mu,\lambda^\nu)}{G_1(\lambda^\mu,\lambda^\nu)}\,v_j^\mu\,v_i^\nu \ , \\ \left[\mathbf{M}_{\xi\hat{\xi}}\right]_{ij}^{\mu\nu} &= \frac{G_2(\lambda^\mu,\hat{\lambda}^\nu)}{G_1(\lambda^\mu,\hat{\lambda}^\nu)}\,v_j^\mu\,\hat{v}_i^\nu \ , \end{aligned} \tag{14}$$

and $\left[\mathbf{M}_{\hat{\xi}\hat{\xi}}\right]_{ij}^{\mu\nu}$ is obtained from $\left[\mathbf{M}_{\xi\xi}\right]_{ij}^{\mu\nu}$ upon substitution of $\lambda^\mu, \lambda^\nu, v_i^\mu, v_i^\nu$ with, respectively, $\hat{\lambda}^\mu, \hat{\lambda}^\nu, \hat{v}_i^\mu, \hat{v}_i^\nu$. Functions G_1 and G_2 are defined through

$$\begin{aligned} G_1(x,y) &= \left(x\,|y-1| + y\,|x-1|\right)^2 \ , \\ G_2(x,y) &= \sqrt{x\,y\,|x-1|\,|y-1|} \ . \end{aligned} \tag{15}$$

The covariance matrix of the fluctuations of the fields is given in Section V D. Error bars on the couplings (4) can be calculated from the ones on the patterns.

Formula (13) tells us how significative are the inferred values of the patterns in the presence of finite sampling. For instance, if the error bar $\langle(\Delta\xi_i^\mu)^2\rangle^{1/2}$ is larger than, or comparable with the pattern component $(\xi^0)_i^\mu$ calculated from (9) then this component is statistically compatible with zero. According to formula (13) we expect error bars of the order of $\frac{1}{\sqrt{\alpha}}$ over the pattern components, where $\alpha = \frac{B}{N}$.

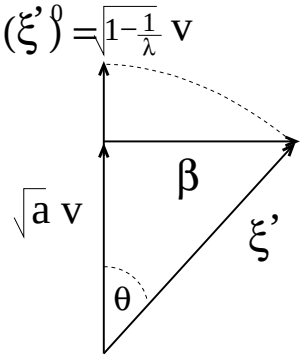

FIG. 1: Geometrical representation of identity (16), expressing the rescaled pattern $\boldsymbol{\xi}'$ as a linear combination of the eigenvector $\mathbf{v}$ and of the orthogonal fluctuations $\boldsymbol{\beta}$. The most likely rescaled pattern, $(\boldsymbol{\xi}')^0$, corresponds to $a = 1 - \frac{1}{\lambda}, \boldsymbol{\beta} = 0$.The dashed arc has radius $\sqrt{1 - \frac{1}{\lambda}}$. The subscript μ has been dropped to lighten notations.

D. Optimal numbers of attractive and repulsive patterns

We now determine the numbers of patterns, p and $\hat{p}$, based on a simple geometric criterion; the reader is referred to Section V E for detailed calculations. To each attractive pattern $\boldsymbol{\xi}^\mu$ we associate the rescaled pattern $(\boldsymbol{\xi}^\mu)'$, whose components are $(\xi_i^\mu)' = \xi_i^\mu \sqrt{1 - m_i^2}/\sqrt{N}$. We write

$$(\boldsymbol{\xi}^\mu)' = \sqrt{a^\mu}\, \mathbf{v}^\mu + \boldsymbol{\beta}^\mu \ , \tag{16}$$

where a^μ is a positive coefficient, and $\boldsymbol{\beta}^\mu$ is a vector orthogonal to all rescaled patterns by virtue of (6) (Fig. 1). Our lowest order formula (9) for the Maximum Likelihood estimators gives $a^\mu = 1 - \frac{1}{\lambda^\mu}$ and $\boldsymbol{\beta}^\mu = 0$, see Fig. 1. This result is, to some extent, misleading. While the most likely value for the vector $\boldsymbol{\beta}^\mu$ is indeed zero, its norm is almost surely not vanishing! The statement may appear paradoxical but is well-known to hold for stochastic variables: while the average or typical value of the location of an isotropic random walk vanishes, its average squared displacement does not. Here, $\boldsymbol{\beta}^\mu$ represents the stochastic difference between the pattern to be inferred and the direction of one of the largest eigenvectors of Γ. We expect the squared norm $(\boldsymbol{\beta}^\mu)^2$ to have a non-zero value in the $N, B \to \infty$ limit at fixed ratio $\alpha = \frac{B}{N} > 0$. Its average value can be straightforwardly computed from formula (14),

$$\langle (\boldsymbol{\beta}^\mu)^2 \rangle = \frac{1}{B} \sum_i \left[M_{\xi\xi} \right]_{ii}^{\mu\mu} = \frac{1}{B} \sum_{k=p+1}^{N-\hat{p}} \frac{1}{\lambda^\mu - \lambda^k} \ , \tag{17}$$

where μ is the index of the pattern. We define the angle θ^μ between the eigenvector $\mathbf{v}^\mu$ and the rescaled pattern $(\boldsymbol{\xi}^\mu)'$ through

$$\theta^\mu = \sin^{-1} \sqrt{\frac{\langle (\boldsymbol{\beta}^\mu)^2 \rangle}{1 - \frac{1}{\lambda^\mu}}} \ , \tag{18}$$

see Fig. 1. Small values of θ^μ correspond to reliable patterns, while large θ^μ indicate that the Maximum Likelihood estimator of the μ^{th} pattern is plagued by noise. The value of p such that θ^p is, say, about $\frac{\pi}{4}$ is our estimate for the number of attractive patterns.

The above approach can be easily repeated in the case of repulsive patterns. We obtain, with obvious notations,

$$\langle (\hat{\boldsymbol{\beta}}^\mu)^2 \rangle = \frac{1}{B} \sum_i \left[M_{\hat{\xi}\hat{\xi}} \right]_{ii}^{\mu\mu} = \frac{1}{B} \sum_{k=p+1}^{N-\hat{p}} \frac{1}{\lambda^k - \hat{\lambda}^\mu} \ , \tag{19}$$

and

$$\hat{\theta}^\mu = \sin^{-1} \sqrt{\frac{\langle (\hat{\boldsymbol{\beta}}^\mu)^2 \rangle}{\frac{1}{\hat{\lambda}^\mu} - 1}} \ . \tag{20}$$

The value of $\hat{p}$ such that $\hat{\theta}^{\hat{p}}$ is, say, about $\frac{\pi}{4}$ is our estimate for the number of repulsive patterns.

E. Regularization

So far we have considered that the prior probability P_0 over the patterns was uniform, and was used to break the invariance through the conditions (6). The prior probability can be used to constrain the amplitude of the patterns. For instance, we can introduce a Gaussian prior on the patterns,

$$P_0 \propto \exp \left[-\frac{\gamma}{2} \sum_{i=1}^{N} (1 - m_i^2) \left(\sum_{\mu=1}^{p} (\xi_i^\mu)^2 + \sum_{\mu=1}^{\hat{p}} (\hat{\xi}_i^\mu)^2 \right) \right] , \tag{21}$$

which penalizes large pattern components [11]. The presence of the $(1 - m_i^2)$ factor entails that the effective strength of the regularization term, $\gamma(1 - m_i^2)$, depends on the site magnetization. Regularization is particularly useful in the case of severe undersampling. With regularization (21) the lowest order expression for the pattern is still given by (9), after carrying out the following transformation on the eigenvalues,

$$\begin{aligned} \lambda^\mu &\to \lambda^\mu - \gamma \ , \quad (\mu = 1, \ldots, p) \ , \\ \lambda^k &\to \lambda^k \ , \qquad (k = p+1, \ldots, N - \hat{p}) \ , \\ \hat{\lambda}^\mu &\to \hat{\lambda}^\mu + \gamma \ , \quad (\mu = 1, \ldots, \hat{p}) \ . \end{aligned} \tag{22}$$

The values of p and $\hat{p}$ must be such that the transformed λ^p and $\hat{\lambda}^{\hat{p}}$ are, respectively, larger and smaller than unity. Regularization (21) ensures that the couplings do not blow up, even in the presence of zero eigenvalues in Γ. Applications will be presented in Sections III and IV. The value of the regularization strength γ can be chosen based on a Bayesian criterion [20].

F. Maximum likelihood inference: first corrections

We now give the expression for the first-order correction to the attractive patterns,

$$(\xi^1)_i^\mu = \sqrt{\frac{N}{1-m_i^2}} \sum_{k=1}^{N} A^{k\mu}\, B^{k\mu}\, v_i^k \ , \tag{23}$$

where

$$\begin{aligned} A^{k\mu} \ &= \ C^k C^\mu + \left(\sum_{\rho=1}^{p} + \sum_{\rho=N+1-\hat{p}}^{N} \right) (\lambda^\rho - 1) \\ &\times \ \sum_i v_i^k v_i^\mu \left[(v_i^\rho)^2 + \frac{2\, m_i\, C^\rho\, v_i^\rho}{\sqrt{1-m_i^2}} \right] \end{aligned} \tag{24}$$

and

$$B^{k\mu} = \begin{cases} \frac{1}{2}\sqrt{\frac{\lambda^\mu}{\lambda^\mu - 1}} & \text{if} \quad k \le p \ , \\ \frac{\sqrt{\lambda^\mu(\lambda^\mu - 1)}}{\lambda^\mu - \lambda^k} & \text{if} \ \ k \ge p+1 \ . \end{cases} \tag{25}$$

and

$$C^k = \sum_i \frac{m_i\, v_i^k}{\sqrt{1-m_i^2}} \left(\sum_{\rho=1}^{p} + \sum_{\rho=N+1-\hat{p}}^{N} \right) (\lambda^\rho - 1)\, (v_i^\rho)^2 \ . \tag{26}$$

Similarly, the first corrections to the repulsive patterns are

$$(\hat{\xi}^1)_i^\mu = \sqrt{\frac{N}{1-m_i^2}} \sum_{k=1}^{N} \hat{A}^{k\mu}\, \hat{B}^{k\mu}\, v_i^k \ . \tag{27}$$

The definition of $\hat{A}^{k\mu}$ is identical to (24), with C^μ and v_i^μ replaced with, respectively, $C^{N+1-\mu}$ and $\hat{v}_i^\mu$. Finally,

$$\hat{B}^{k\mu} = \begin{cases} \frac{1}{2}\sqrt{\frac{\hat{\lambda}^\mu}{1-\hat{\lambda}^\mu}} & \text{if} \ \ k \ge N - \hat{p} + 1 \ , \\ \frac{\sqrt{\hat{\lambda}^\mu(1-\hat{\lambda}^\mu)}}{\hat{\lambda}^\mu - \lambda^k} & \text{if} \quad k \le N - \hat{p} \ . \end{cases} \tag{28}$$

The first order corrections to the fields h_i can be found in Section V F.

It is interesting to note that the corrections to the pattern $\boldsymbol{\xi}^\mu$ involve non-linear interactions between the eigenmodes of Γ. Formula (24) for $A^{k\mu}$ shows that the modes μ and k interact through a multi-body overlap with mode ρ (provided $\lambda^\rho \neq 1$). In addition, $A^{k\mu}$ does not *a priori* vanish for $k \ge p+1$: corrections to the patterns have non–zero projections over the 'noisy' modes of Γ. In other words, valuable information over the true values of the patterns can be extracted from the eigenmodes of Γ associated to bulk eigenvalues.

G. Quality of the inference vs. size of the data set

The accuracy ϵ on the inferred pattern is limited both by the sampling error resulting from the finite number of data and the intrinsic error due to the expansion (8). According to Section II C, the sampling error on the pattern component is expected to decrease as $\sim \sqrt{\frac{N}{B}}$. The intrinsic error depends on the order of the expansion, on the size N and on the amplitude of the patterns.

No inference is possible unless the ratio $\alpha = \frac{B}{N}$ exceeds a critical value, referred to as α_c in the following (Section VI A 2). This phenomenon is similar to the retarded learning phase transition discovered in the context of unsupervised learning [18].

Assume that the pattern components ξ_i are of the order of one (compared to N), that is, that the couplings are almost all non zero and of the order of $\frac{1}{N}$. Then, the intrinsic error is of the order of $\frac{1}{N}$ with the lowest order formula (9), and of the order of $\frac{1}{N^2}$ when corrections (23) are taken into account; for a more precise statement see Section V A and formula (49). The corresponding values of B at which saturation takes place are, respectively, of the order of N^3 and N^5. The behaviour of the relative error between the true and inferred patterns, ϵ (32), is summarized in Fig. 2. In general we expect that $B \sim N^{1+2a}$ samples at least are required to have a more accurate inference with a^{th}-order patterns than with $(a-1)^{th}$-order patterns. Furthermore there is no need to sample more than N^{3+2a} configurations when using the a^{th}-order expression for the patterns.

If the system has $O(N)$ non vanishing couplings J_{ij} of the order of J, then patterns have few large components, of the order of $\sqrt{J}$. In this case the intrisic error over the patterns will be of the order of J with the lowest order inference formulae, and of the order of J^2 with the first corrections. The numbers of sampled configurations, B,

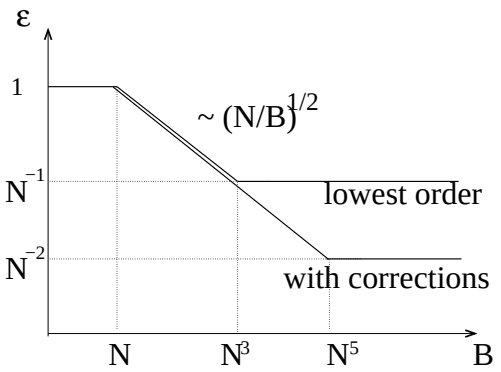

FIG. 2: Schematic behaviour of the error ϵ on the inferred patterns as a function of the number B of sampled configurations and for a problem size equal to N, when the pattern components are of the order of unity compared to N. See main text for the case of few large pattern components, of the order of $\sqrt{N}$, *i.e.* couplings J of the order of 1.

required to reach those minimal errors will be, respectively, of the order of $\frac{N}{J^2}$ and $\frac{N}{J^4}$.

III. TESTS ON SYNTHETIC DATA

In this Section we test the formulae of Section II for the patterns and fields against synthetic data generated from various Ising models with known interactions. We consider four models:

- *Model A* is a Hopfield model with $N = 100$ spins, p (= 1 or 3) attractive patterns and no repulsive pattern ($\hat{p} = 0$). The components of the patterns are Gaussian random variables with zero mean and standard deviation ξ, specified later. The local fields h_i are set equal to zero.
- *Model B:* Model B consists of N spins, grouped into four blocks of $\frac{N}{4}$ spins each. The $p = 3$ patterns have uniform components over the blocks: $\xi^1 = \frac{2\sqrt{3}}{5}(0, 1, 1, 1)$, $\xi^2 = \frac{2}{5}(\sqrt{3}, 1, -2, 1)$, $\xi^3 = \frac{2}{5}(\sqrt{3}, -2, 1, 1)$. The fields are set to zero. Those choices ensure that the pattern are orthogonal to each other, and have a weak intensity: on average, $|\xi|^2 = \frac{9}{25} < 1$.
- *Model C* is a very simple Ising model where all fields and couplings vanish, except coupling $J_{12} \equiv J$ between the first two spins.
- *Model D* is an Ising model with $N = 50$ spins, on an Erdos-Renyi random graph with average connectivity (number of neighbors for each spin) equal to $d = 5$ and coupling values J distributed uniformly between -1 and 1. Model D is an instance of the Viana-Bray model [21]. In the thermodynamic limit $N \to \infty$ this model is in the spin glass phase since $d\langle \tanh^2(J)\rangle_J > 1$ [21].

For each one of the models above, the magnetizations and pairwise correlations can be estimated through the sampling of B configurations at equilibrium using Monte Carlo simulations. This allows us to estimate the consequence of sampling noise on the inference quality by varying the value of B. Furthermore, for models B and C, it is possible to obtained the exact Gibbs values for m_i and c_{ij} (corresponding to a perfect sampling, $B = \infty$)[40]. This allows us to study the systematic error resulting from formulae (9,23,27), irrespectively of the sampling noise.

Model A is used to test the lower order formula for the patterns, and how the quality of inference depends on the amplitude of the patterns. Models C and D are highly diluted networks with strong $J = O(1)$ interactions, while models A and B correspond to dense networks with weak $J = O(\frac{1}{N})$ couplings. Models C and D are, therefore, harder benchmarks for the generalized Hopfield model. In addition, the couplings implicitly define, through (4), both attractive and repulsive patterns.

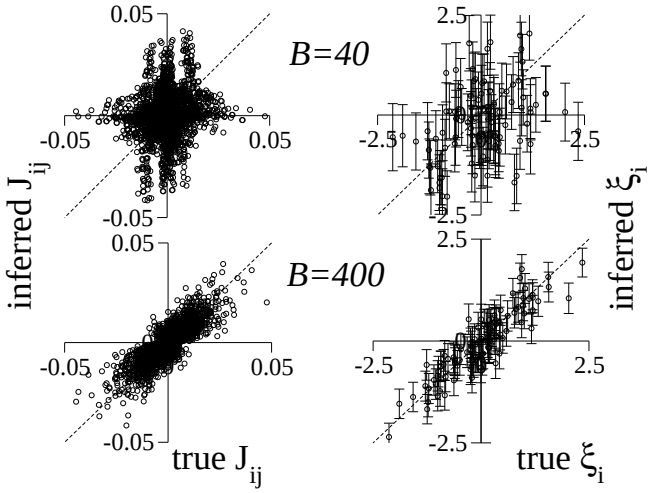

FIG. 3: Application of formula (9) to two sets of $B = 40$ (top) and 400 (bottom) configurations, randomly generated from the distribution P_H (2) for model A with $p = 1$ pattern. The standard deviation of the pattern components is $\xi = .7$. **Left:** comparison of the true and inferred couplings for each pair (i, j). **Right:** comparison of the true and inferred components ξ_i of the pattern, with the error bars calculated from (13). The dashed lines have slope unity. Inference is done with $p = 1$ attractive pattern and no repulsive pattern.

Those models can thus be used to determine how much repulsive patterns are required for an accurate inference of general Ising models.

A. Dominant order formula for the patterns

We start with Model A with $p = 1$ pattern. In this case, no ambiguity over the inferred pattern is possible since the energy E is not invariant under continuous transformations, see Section II A. We may therefore directly compare the true and the inferred patterns. Figures 3 and 4 show the accuracy of the lowest order formula for the patterns, eqn (9). If the pattern components are weak, each sampled configuration $\boldsymbol{\sigma}$ is weakly aligned along the pattern $\boldsymbol{\xi}$. If the number B of sampled configurations is small, the largest eigenvector of Γ is uncorrelated with the pattern direction (Fig. 3). When the size of the data set is sufficiently large, *i.e.* $B > \alpha_c N$ (Section VI A 2), formula (9) captures the right direction of the pattern, and the inferred couplings are representative of the true interactions. Conversely, if the amplitudes of the components of the pattern $\boldsymbol{\xi}$ are strong enough, each sampled configuration $\boldsymbol{\sigma}$ is likely to be aligned along the pattern. A small number B (compared to N) of those configurations suffice to determine the pattern (Fig. 4). In the latter case, we see that the largest components ξ_i are systematically underestimated. A systematic study of how large B should be for the inference to be reliable can be found in Section VI.

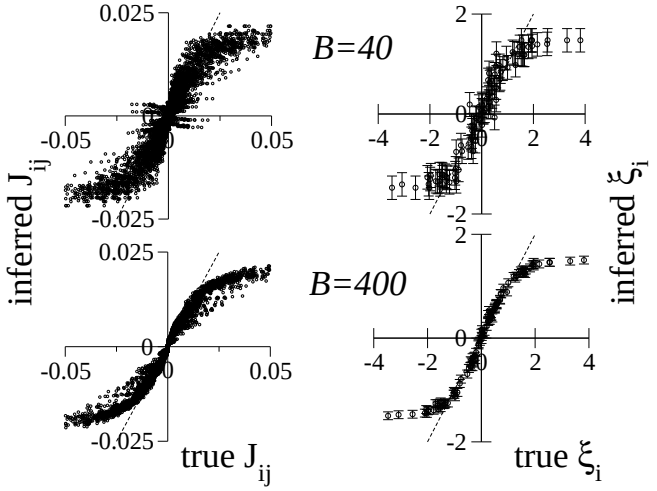

FIG. 4: Same as Fig. 3, but with a standard deviation $\xi = 1.3$ instead of $\xi = .7$. The amplitude is strong enough to magnetize the configurations along the pattern, see Sections VI A 1 and VI B 3.

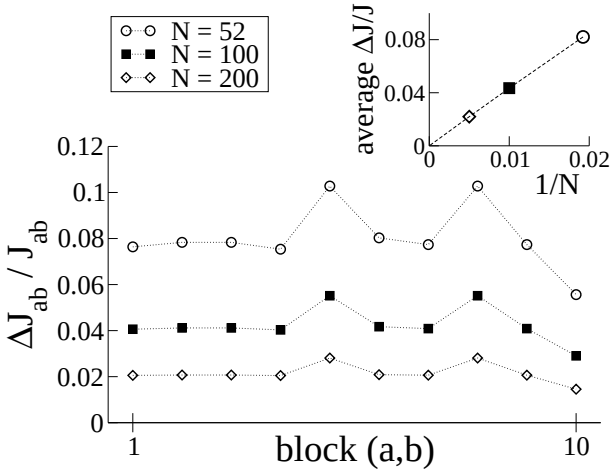

FIG. 5: Relative differences between the true and the inferred couplings, $\Delta J_{ab}/J_{ab}$ for three system sizes, N. The inference was done using the lowest order ML formulae (9) for the patterns. Data were generated from Model B (perfect sampling); there are *a priori* ten distinct values of the couplings, one for each pair of blocks a and b. Inset: average value of $\Delta J_{ab}/J_{ab}$ as a function of $\frac{1}{N}$. Circles, squares and diamonds correspond to, respectively, $N = 52$, 100 and 200 spins.

We now use model B to generate the data. As model B includes more than one pattern, the inferred patterns cannot be compared to the true one easily due to the invariance of Section II A. We therefore compare in Fig. 5 the true couplings and the interactions found using (9) for three sizes, $N = 52$, 100 and 200. The size N sets also the amplitude of the couplings, which decreases as $\frac{1}{N}$ from (4). As the patterns are uniform among each one of the four blocks there are ten possibles values for the couplings J_{ij}, depending on the labels a and b of the blocks to which i and j belong, with $1 \leq a \leq b \leq 4$. For $N = 100$ spins, the relative errors range between 3 and 5.5%. When the number of spins is doubled (respectively, halved) the relative errors are about twice smaller (respectively, larger). This result confirms that formula (9) is exact in the infinite N limit only, and that corrections of the order of $O(\frac{1}{N})$ are expected for finite system sizes (Inset of Fig. 5). This scaling was expected from Section II G.

We now consider model C. For perfect sampling ($B = \infty$) the correlation matrix (1) is

$$\Gamma = \begin{pmatrix} 1 & \tanh J & 0 & \dots & 0 \\ \tanh J & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix} . \qquad (29)$$

The top eigenvalue, $\lambda^1 = 1 + \tanh J > 1$, and the smallest eigenvalue, $\hat{\lambda}^1 = \lambda^N = 1 - \tanh J < 1$, are attached to the eigenvectors

$$\mathbf{v}^1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} , \quad \hat{\mathbf{v}}^1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} . \qquad (30)$$

The remaining $N - 2$ eigenvalues are equal to 1. Using formula (10) for the lowest order coupling, J^0, we find that those eigenmodes do not contribute and that the interaction can take three values, depending on the choices for p and $\hat{p}$:

$$\begin{aligned} (J^0)_{p=1,\hat{p}=0} &= \frac{\tanh J}{2\,(1+\tanh J)} \simeq \frac{J}{2} - \frac{J^2}{2} + \frac{J^3}{3} + \dots , \\ (J^0)_{p=0,\hat{p}=1} &= \frac{\tanh J}{2\,(1-\tanh J)} \simeq \frac{J}{2} + \frac{J^2}{2} + \frac{J^3}{3} + \dots , \\ (J^0)_{p=1,\hat{p}=1} &= \frac{\tanh J}{1-\tanh^2 J} \simeq J + \frac{2\,J^3}{3} + \dots . \end{aligned} \qquad (31)$$

Those expressions are plotted in Fig. 6. The coupling $(J^0)_{1,0}$ (dashed line), corresponding to the standard Hopfield model, saturates at the value $\frac{1}{4}$ and does not diverge with J. Even the small J behavior, $(J^0)_{1,0} \simeq \frac{J}{2}$, is erroneous. Adding the repulsive pattern leads to a visible improvement, as fluctuations of the spin configurations along the eigenvector $\hat{\mathbf{v}}^{\mathbf{1}}$ (one spin up and the other down) are penalized. The inferred coupling, $(J^0)_{1,1}$ (bold line), is now correct for small J, $(J^0)_{1,1} \simeq J$, and diverges for large values of J.

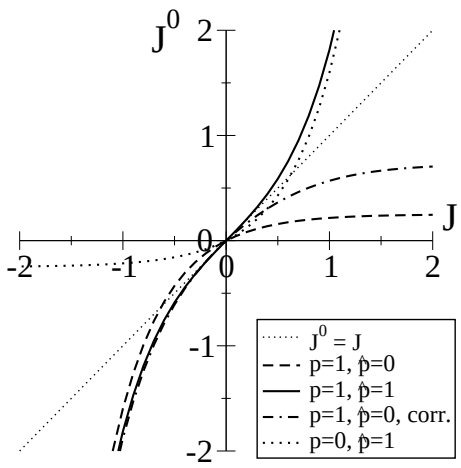

FIG. 6: Inferred coupling J^0 between the first two spins of Model C, within lowest order ML, and as a function of the true coupling J. Values of p and $\hat{p}$ are shown in the Figure.

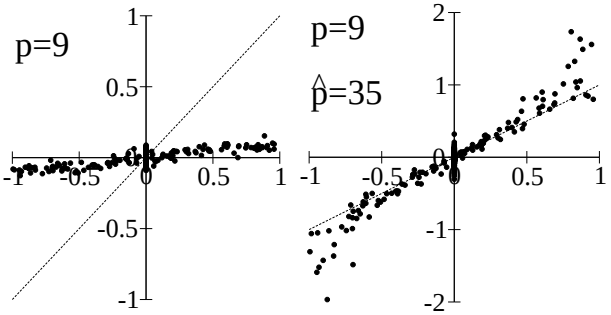

FIG. 7: Inferred vs. true couplings for Model D, with $B = 4500$ sampled configurations. **Left**: Hopfield model with $p = 9$ (corresponding to the optimal number of patterns selected by the geometrical criterion); no repulsive pattern is considered ($\hat{p} = 0$). **Right:** Generalized Hopfield model with $(p, \hat{p}) = (9, 35)$ (optimal numbers).

We now turn to Model D. Figure 7 compares the inferred and true couplings for $B = 4500$ sampled configurations. The generalized Hopfield model outperforms the standard Hopfield model ($\hat{p} = 0$), showing the importance of repulsive patterns in the inference of sparse networks with strong interactions. Large couplings, either positive or negative, are overestimated by the lowest order ML estimators for the patterns.

B. Error bars and criterion for $p, \hat{p}$

An illustration of formula (13) for the error bars is shown in Fig. 3, where we compare the components of the true pattern used to generate data in Model A with the inferred one, $(\xi^0)_i$, and the error bar, $\sqrt{\langle(\Delta\xi_i)^2\rangle}$. For small $\alpha = \frac{B}{N}$ the inferred pattern components are uncorrelated with the true pattern and compatible with zero within the error bars. For larger values of α, the discrepancy between the inferred and the true components are stochastic quantities of the order of the calculated error bars.

We report in Fig. 8 the tests of the criterion for determining p and $\hat{p}$ on artificially generated data from an extension of model A with $p = 3$ patterns. For very poor sampling (Fig. 8, top) the angle θ^1 is close to $\frac{\pi}{4}$: even the first pattern cannot be inferred correctly. This prediction is confirmed by the very poor comparison of the true interactions and the inferred couplings calculated from the first inferred pattern. For moderately accurate sampling (Fig. 8, middle) the strongest pattern can be inferred; the accuracy on the inferred couplings worsens when the second pattern is added. Excellent sampling allows for a good inference of the structure of the underlying model: the angle θ^μ is small for $\mu = 1, 2, 3$ (Fig. 8, bottom), and larger than $\frac{\pi}{4}$ for $\mu \geq 4$ (not shown). Not surprisingly large couplings are systematically affected by errors. Those errors can be corrected by taking into account $O(\frac{\xi}{\sqrt{N}})$ corrections to the patterns if the number of data, B , is large enough (Section VI).

Figure 9 compares the inferred and true couplings for $B = 4500$ sampled configurations of Model D. The optimal number of patterns given by the geometrical criterion is $(p = 9, \hat{p} = 35)$, see Fig. 7. Hence most of the components of Γ are retained and the interactions inferred with the generalized Hopfield model do not differ much from the MF couplings.

C. Corrections to the patterns

Formula (23) for the corrections to the patterns was tested on model B in the case of perfect sampling. Results are reported in Fig. 10 and show that the errors in the inferred couplings are much smaller than in Fig. 5. Inset of Fig. 10 shows that the relative errors are of the order of $\frac{1}{N^2}$ only. This scaling was expected from Section II G. Pushing our expansion of ξ to the next order in powers of $\frac{1}{N}$ could in principle give explicit expressions for those corrections. We have also tested our higher order formula when the fields h_i are non-zero. For instance we have considered the same Hopfield model with $p = 3$ patterns as above, and with block pseudo-magnetizations $\mathbf{t} = \frac{1}{15}(2\sqrt{3}, 2, 2, -4)$. Hence, $\mathbf{t}$ was orthogonal to the patterns, and the field components were simply given by $h_i = \tanh^{-1} t_i$, according to (38) [41]. For $N = 52$ spins the relative error over the pseudo-magnetizations (aver-

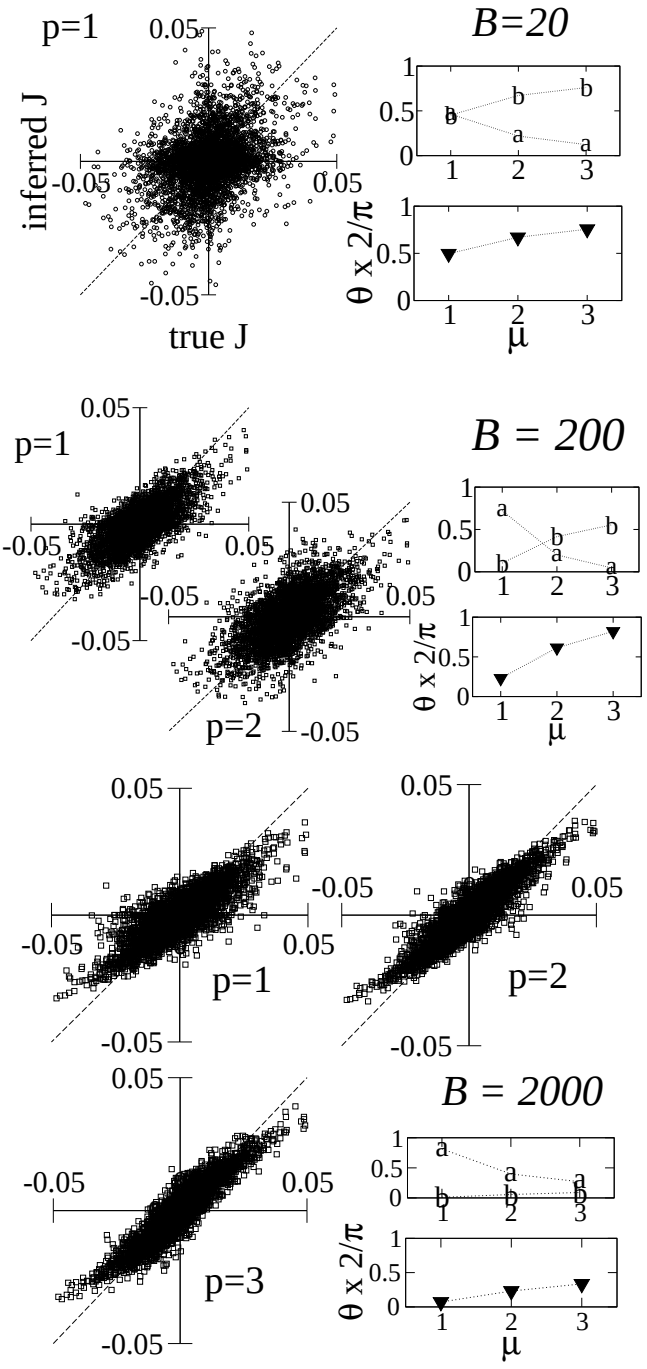

FIG. 8: Criterion to decide the number p of patterns and performance of the ML inference procedure for three different sizes of the data set, B. **Left:** inferred vs. true interactions with $p = 1, 2$ or 3 patterns; the dashed line has slope unity. **Right:** coefficients $a^\mu = (\rho^\mu)^2$ and $b^\mu = \langle(\boldsymbol{\beta}^\mu)^2\rangle$ vs. pattern index μ, and angles θ^μ, divided by $\frac{\pi}{2}$, see definitions (16) and (18). For each value of B one data set was generated from Model A with $p = 3$ patterns, and standard deviations $\xi^1 = .95$, $\xi^2 = .83$, and $\xi^3 = .77$.

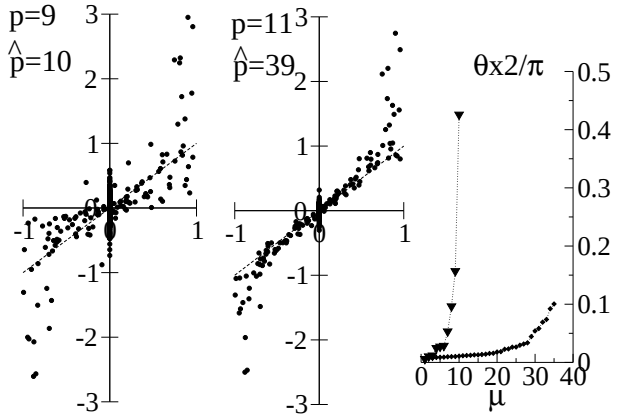

FIG. 9: Inferred vs. true couplings for Model D, with $B = 4500$ sampled configurations. **Left:** Generalized Hopfield model with $(p, \hat{p}) = (9, 10)$ and $(11, 39)$ (corresponding to the numbers of eigenvalues, respectively, larger and smaller than unity). **Right:** angles θ^μ and $\hat{\theta}^\mu$ for, respectively, attractive (triangle) and repulsive (diamond) patterns.

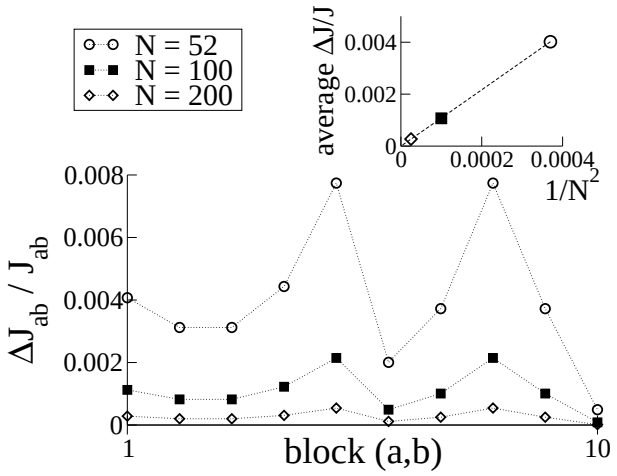

FIG. 10: Relative differences between the true and the inferred couplings, $\Delta J_{ab}/J_{ab}$ as a function of the system size, N. The inference was done using the finite–N ML formulae (9) and (23) for the patterns. Data were generated from a perfect sampling of the equilibrium distribution of a Hopfield model with $p = 3$ patterns and four blocks of $\frac{N}{4}$ spins, see main text; a and b are the block indices. Inset: average value of $\Delta J_{ab}/J_{ab}$ as a function of $\frac{1}{N^2}$. Circles, squares and diamonds correspond to, respectively, $N = 52$, 100 and 200 spins.

aged over the four blocks a) was $\frac{\Delta t_a}{t_a} \simeq .0301$ with the large-N formula (9) and $\frac{\Delta t_a}{t_a} \simeq 0.0029$ with the finite-N formulae (23) and (78).

Corrections to the PCA were also tested when data are corrupted by sampling noise. We compare in Fig. 11 the components of the pattern of Model A found with the lowest order approximation (9) and with our first order

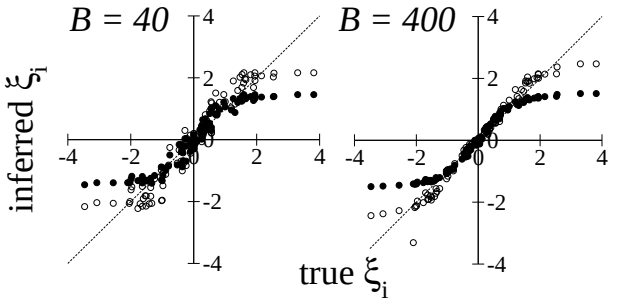

FIG. 11: True vs. inferred components of the patterns, ξ_i, for the model with $N = 100$ spins described in Fig. 4. Full circles are the result of the lowest order inference formula (9), while empty circles show the outcome of the first order formulae (23).

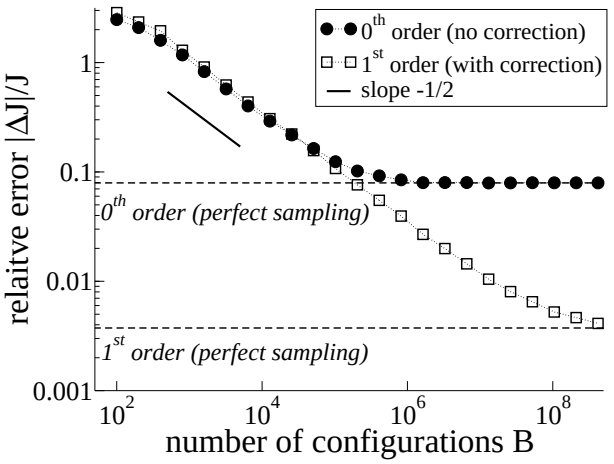

FIG. 12: Relative error between the inferred and true couplings for Model B (with $N = 52$ spins) vs. number of sampled configurations, B. The two curves correspond to the inference done with the 0^{th} order formula (9) (black circles) and the 1^{st} order formula (23) (squares). Each data point is the average over 10 samples; relative error bars are about 1%, and are much smaller than the symbol size. The asymptotic value of the errors, corresponding to perfect sampling ($B = \infty$), are extracted from Figs. 5 and 10.

formulae (23) (case of strong pattern). A clear improvement in the quality of the inference is observed, even when the sampling noise is strong. Our second example is Model B. We show in Fig. 12 the relative errors

$$\epsilon_J = \frac{2}{N(N-1)} \sum_{i<j} \left| \frac{\Delta J_{ij}}{J_{ij}} \right| \qquad (32)$$

between the true and the inferred couplings, with formulas (9) and (23), as a function of the number of sampled configurations, B, and for $N = 52$ spins. As B increases the relative error with the lowest order patterns (PCA) first decreases as $B^{-1/2}$, then saturates to the value $\simeq .0794$, as expected from Fig. 5. The relative error with the correction to the patterns also decreases as $B^{-1/2}$, and is expected to saturate to the lower value $\simeq .00374$ (Fig. 10). We remark that the gain in accuracy over the inferred couplings resulting from the corrections (23) to the patterns is obtained only when B is very large. $B \sim N^3$ configurations at least should be sampled to obtain an improvement over the lowest order formula (9). This scaling holds when the couplings are weak, and decrease as $\frac{1}{N}$. If the interaction network is diluted and carries couplings $J = O(1)$, we expect that $B \sim N/J^2$ configurations have to be sampled to make the first-corrections to the patterns effective.

We have applied our formula (23) to calculate the first correction to the couplings (31) for Models C and D. As for Model C, we find that the correction to the coupling $(J^0)_{1,1}$ vanishes; this result is due to the fact that $(J^0)_{1,1}$ is already correct to the second order in J, and that higher order corrections would be needed. The corrections to the coupling $(J^0)_{1,0}$ are equal to

$$\begin{aligned} (J^1)_{1,0} &= \frac{\tanh J}{2\sqrt{2}} + \frac{\tanh J(1+\tanh J)}{16} \\ &= \left(\frac{1}{16} + \frac{1}{2\sqrt{2}}\right) J + \frac{J^2}{16} + \ldots\ldots . \end{aligned} \qquad (33)$$

The resulting coupling, $(J^0 + J^1)_{10}$, is plotted as a function of J in Fig. 6, and qualitatively improves over the lowest order result (31). In particular, for small J, the inferred coupling is now $(J^0 + J^1)_{10} \simeq .916\, J - .438\, J^2$, which is definitely closer to J than (31). In the case of Model D, the first-order corrections improve only slightly the estimates for the large couplings.

IV. APPLICATION TO BIOLOGICAL DATA

In this Section we show how the inference approach can be applied to real biological data, and compared to other Boltzmann Machine learning procedures.

A. Cortical activity of the rat

We have first analyzed data coming from the recording of 37 neurons in the prefrontal cortex of rats. The experiment, done by A. Peyrache, F. Battaglia and their collaborators, consists in recording the neural activity during a task and during the Slow Wave sleep preceding and following the learning of the task [22]. PCA allowed Peyrache et al. to identify patterns in the activity, which are generated when the rat learns a task and are replayed during the sleep [22].

We have analyzed with the generalized Hopfield model the data corresponding to a 20 minute-long recording of the activity of a rat during the task (data shown in Fig. 1 of [22]). The raster plot was binned with a 10 msec window to obtain binary configurations of the neurons (active or silent in the time-bin). We have then calcu-

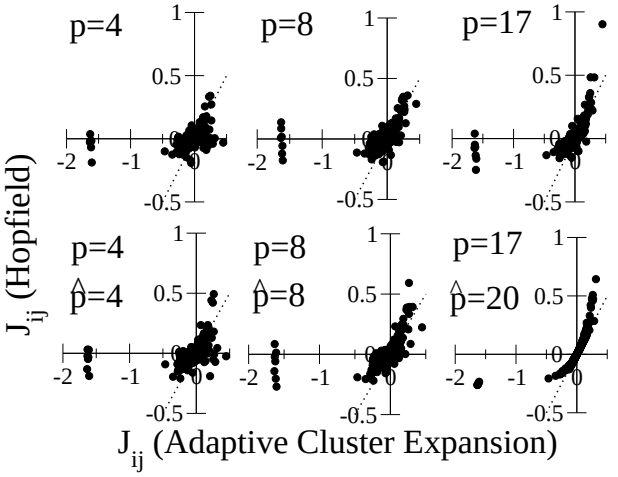

FIG. 13: Couplings calculated with the generalized Hopfield model vs. couplings calculated with the adaptive cluster expansion of [11] for 37 cells recorded in the prefrontal cortex of a behaving rat. **Top**: Hopfield model with $p = 4$, 8 (corresponding to the optimal number of patterns selected by the geometrical criterion) and 17; no repulsive pattern is considered ($\hat{p} = 0$). **Bottom.** Generalized Hopfield model with $(p, \hat{p}) = (4, 4)$, $(8, 8)$ (optimal numbers) and $(17, 20)$ (corresponding to the numbers of eigenvalues, respectively, larger and smaller than unity).

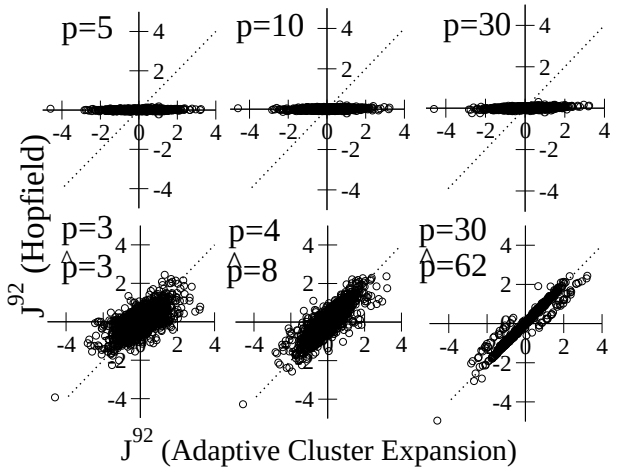

FIG. 14: Couplings calculated with the Generalized Hopfield model versus coupling calculated with the adaptive cluster expansion for 92 amino-acids in the PDZ domain.The values of $p, \hat{p}$ are given in the Figure. Note that $\hat{p} = 0$ for the top panels. The middle panels correspond to the optimal values for the number of patterns.

lated the average frequencies, m_i, and the pairwise correlations, c_{ij}. We calculate the couplings with p attractive and $\hat{p}$ repulsive patterns according to (9) and (10). The numbers p and $\hat{p}$ are calculated according to the geometrical criteria (18) and (20). Hereafter, we compare the couplings obtained this way to the ones found with the adaptive cluster expansion (ACE) of [11], which is not based on the expansion of the loglikelihood used in the present work.

In Fig. 13 (top) we compare the Hopfield ($\hat{p} = 0$) couplings with $p = 4, 8, 17$ selected patterns to the ACE couplings. The agreement is quite good for $p^* = 8$. In [22] $p = 6$ patterns were kept in the PCA; this value is close to the optimal value, $p = 8$, we find using the geometrical criterion. Addition of repulsive patterns (bottom of Fig. 13) slightly improves the similarity with the ACE couplings. We find, indeed, that the couplings J_{ij} are rather weak, and that repulsive patterns do not play an important role. Calculating the couplings with all eigenmodes ($p = 17, \hat{p} = 20$) is equivalent ot the mean-field (MF) approximation. A clear discrepancy between the Hopfield and the ACE couplings is found for the largest (in absolute value) interactions. We have checked that this discrepancy is not reduced when the first order corrections to the patterns are included, presumably because the number of data is not sufficient. Couplings are not significatively changed in the presence of the regularization (21) for sensible values of γ.

B. Protein-domain families

We have next analyzed the alignement of a family of 240 sequences of PDZ, a commonly encountered domain binding the C-terminus of proteins, with 92 amino-acids [24]. R. Ranganathan and collaborators have elaborated an approach, called Statistical Coupling Analysis(SCA), to extract interactions between residues by using evolutionary data for the protein, *i.e.* by sampling the single-site and pairwise frequencies from multi-sequence alignments of the family [23]. Briefly speaking, SCA consists in doing a PCA analysis of a weighted correlation matrix, $D_i \Gamma_{ij} D_j$, where the weight D_i on site i is small for poorly conserved residues [24].

We have taken the binary data representation of the 240 PDZ sequences in the alignement given in [25] (Supplementary Material). This consensus approximation amounts to replace the amino-acid on each site (20 possible types) with a binary variable σ_i^b, equal to $+1$ if the amino-acid i in the b^{th} sequence is the most common amino-acid at that position in the alignment, to -1 otherwise. The consensus representation does not allow to keep track of all the information contained in the alignment but is indicative of the conservation pattern in the family.

The inferred couplings, denoted by J^{92}, are shown in Fig. 14. As in the case of Model D in Section III we find that proteomic data are better accounted with the generalized Hopfield model than with the standard Hopfield model: repulsive patterns seems necessary to recover the couplings found with the ACE method. The couplings found with attractive patterns only are not correlated with the ACE couplings (top of Fig. 14), while the agree-

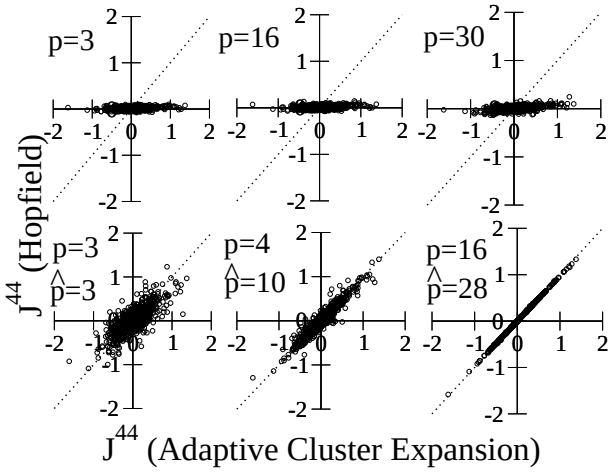

FIG. 15: Same as Fig. 14 when retaining the 44 residues with the largest weights D_i only [24]. The values of $p, \hat{p}$ are given in the Figure. Note that $\hat{p} = 0$ for the top panels. The middle panels correspond to the optimal values for the number of patterns.

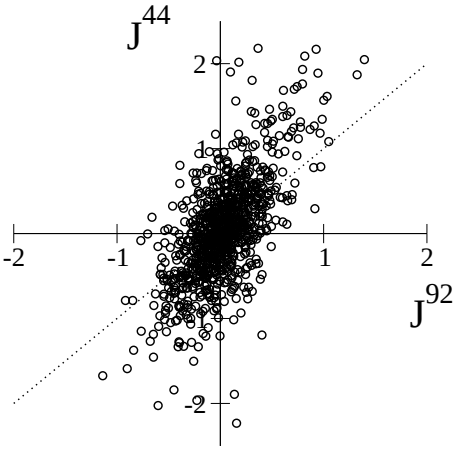

FIG. 16: Comparison between the couplings J_{ij} calculated with all 92 residues and with the 44 most weighted residues only, for each one of the $44 \times 43/2$ pairs (i, j) of residues.

ment is quite good when taking into account attractive and repulsive patterns; the optimal numbers of patterns are $p = 4$ and $\hat{p} = 10$.

We have also calculated the couplings when discarding all but the most weighted sites. More precisely, we have recalculated the distribution of the weights D_i as in [24, 25], and found a bimodal distribution, which suggests a natural cut-off between large and small weights. We have redone the previous inference when keeping only the 44 residues (out of 92) with the largest weights, corresponding to the red sites in Fig. C of [24]. The resulting interactions, denoted by J^{44}, are shown in Fig. 15. Again we compare the couplings found with the Hopfield model and with the ACE. The agreement is not good with attractive patterns only (as done in usual PCA), and is very good when repulsive patterns are included.

An interesting question is whether the couplings obtained between the 44 most conserved residues are strongly affected by the presence or the absence of the remaining 48 residues in the inference. The interactions in the 44-site model are effective and *a priori* differ from their values in the 92-site model, in that they account for chains of interactions going through the remaining 48 sites. Nevertheless, we find that the couplings calculated with all 92 residues and the couplings obtained from the subset of 44 sites with large weights are similar, see Fig.16. This result suggests that the 48 residues removed from our second analysis are not strongly interacting with the 44 retained sites.

V. EXPANSION OF THE CROSS ENTROPY AND MAXIMUM LIKELIHOOD INFERENCE

This Section is intended to provide the derivations of the results announced in Section II. Maximizing the posterior probability (5) with respect to the patterns and the fields is equivalent to minimizing the cross entropy of the Hopfield model given the data,

$$\Phi[\mathbf{h}, \{\boldsymbol{\xi}^\mu\}, \{\boldsymbol{\sigma}^b\}] = \log Z[\mathbf{h}, \{\boldsymbol{\xi}^\mu\}] + U[\mathbf{h}, \{\boldsymbol{\xi}^\mu\}, \{\boldsymbol{\sigma}^b\}] \ , \tag{34}$$

where Z is the partition function appearing in (2),

$$Z[\mathbf{h}, \{\boldsymbol{\xi}^\mu\}] = \sum_{\boldsymbol{\sigma}} \exp\big(- E[\boldsymbol{\sigma}, \mathbf{h}, \{\boldsymbol{\xi}^\mu\}]\big) \ , \tag{35}$$

and U is the average value of the energy E (3) over the sampled configurations:

$$U[\mathbf{h}, \{\boldsymbol{\xi}^\mu\}, \{\boldsymbol{\sigma}^b\}] = -\sum_{i=1}^{N} h_i m_i - \frac{1}{2}\sum_{i,j} J_{ij}\, c_{ij} \ , \tag{36}$$

where the couplings J_{ij} are calculated from the patterns according to (4). The calculation of the partition function, which is defined as a sum over 2^N configurations, cannot generally be done in a reasonable time for large sizes N. In the next section we show how the use of statistical mechanics techniques allows one to obtain a systematic expansion of Z, and, thus, of the cross entropy

$$\Phi = \Phi^0 + \Phi^1 + \dots \ , \tag{37}$$

in powers on $\frac{\xi_i}{\sqrt{N}}$ and $\frac{\hat{\xi}_i}{\sqrt{N}}$.

A. Expansion of the free energy of the Hopfield model in powers of $\frac{\xi_i}{\sqrt{N}}, \frac{\hat{\xi}_i}{\sqrt{N}}$

To lighten notations calculations are presented for the case of attractive patterns only. We explain at the end of the Section how formulae are modified in the presence of repulsive patterns.

For technical reasons to be made clear below it results convenient to make the change of variables $\mathbf{h} \to \mathbf{t}$ described by

$$h_i \;=\; \tanh^{-1} t_i - \frac{1}{N}\sum_{\mu}\sum_{j}\xi_i^\mu\xi_j^\mu\, t_j \;, \tag{38}$$

where the t_i, hereafter called pseudo-magnetizations, are real-valued numbers comprised between -1 and 1. Hereafter, we will infer the most likely values for $\mathbf{t}$, and will recover the fields $\mathbf{h}$ through (38). The change $\mathbf{h} \to \mathbf{t}$ amounts to consider the energy function

$$E = -\sum_{i=1}^{N}\sigma_i \tanh^{-1} t_i - \frac{1}{2N}\sum_{\mu=1}^{p}\left(\sum_{i=1}^{N}\xi_i^\mu\big(\sigma_i - t_i\big)\right)^2 \;, \tag{39}$$

instead of the original expression for E (3) (with $\hat{p} = 0$). Obviously, when the identities (38) are fulfilled, both energies are equal (up to a $\boldsymbol{\sigma}$-independent additive term) and define the same likelihood function (2).

We unravel the squared terms in the partition function (35) through a set of p auxiliary Gaussian variables $\mathbf{x} = (x^1, \ldots, x^p)$, and carry out the summation over the spin configurations. We obtain

$$\begin{aligned} Z \;=\; \int\prod_{\mu}\frac{dx^\mu}{\sqrt{2\pi}}\;\exp\Big[-\frac{1}{2}\sum_{\mu}(x^\mu)^2 - \sum_{i,\mu}\frac{x^\mu\,\xi_i^\mu\,t_i}{\sqrt{N}} \\ + \sum_i \log 2\cosh\left(\tanh^{-1} t_i + \sum_\mu \frac{x^\mu\,\xi_i^\mu}{\sqrt{N}}\right)\Big] \;. \end{aligned} \tag{40}$$

If N is large enough the dominant contribution to the integral will come from $\mathbf{x}^*$, the value of $\mathbf{x}$ maximizing the argument of the exponential above. We obtain the following saddle point equation for $\mathbf{x}$,

$$(x^\mu)^* = \frac{1}{\sqrt{N}}\sum_i \xi_i^\mu (T_i - t_i) \;, \tag{41}$$

where

$$T_i \equiv \tanh\left(\tanh^{-1} t_i + \sum_\mu \frac{(x^\mu)^*\,\xi_i^\mu}{\sqrt{N}}\right) \tag{42}$$

We then write $x^\mu = (x^\mu)^* + y^\mu$ and expand the hyperbolic cosine function in powers of y^μ. The change of variable (38) is such that the linear term in y^μ in the expansion of the hyperbolic cosine function cancels out with the linear term in the exponential, $-\sum_{i,\mu}\frac{y^\mu\,\xi_i^\mu\,t_i}{\sqrt{N}}$, independently of the value of $(x^\mu)^*$. Expanding the hyperbolic cosine up to the second order in y^μ we find our lowest order approximation to the partition function,

$$\begin{aligned} Z^0 \;&=\; e^{F^*}\int\prod_\mu\frac{dy^\mu}{\sqrt{2\pi}}\;\exp\Big[-\frac{1}{2}\sum_\mu (y^\mu)^2 \\ &+\; \frac{1}{2\,N}\sum_i\sum_{\mu,\nu}\xi_i^\mu\xi_i^\nu y^\mu y^\nu\left(1 - T_i^2\right)\Big] = \frac{e^{F^*}}{\sqrt{\det\,\mathrm{A}}} \end{aligned} \tag{43}$$

where F^* is the the argument of the exponential in (40) calculated in $x^{\mu *}$,

$$F^* = N\log 2 + \frac{1}{2}\sum_i \log(1 - T_i^2) - \sum_{\mu,ij}\xi_i^\mu\xi_j^\mu(T_i\,T_j - t_i\,t_j) \;, \tag{44}$$

and A is the $p \times p$ matrix with entries,

$$A^{\mu\nu} = \delta^{\mu\nu} - \frac{1}{N}\sum_i \xi_i^\mu\xi_i^\nu\left(1 - T_i^2\right) \;. \tag{45}$$

We then compute the average energy U (36),

$$\begin{aligned} U \;&=\; -\sum_i m_i \tanh^{-1}\, t_i \\ &-\; \frac{1}{2N}\sum_{\mu,ij}\xi_i^\mu\xi_j^\mu\,(c_{ij} - m_i\,t_j - t_i m_j + t_i\,t_j) \;. \end{aligned} \tag{46}$$

Our lowest order approximation for the cross entropy is, according to (34), (44) and (46):

$$\begin{aligned} \Phi^0 \;&=\; -\sum_{i=1}^{N} m_i \tanh^{-1} T_i + N\log 2 + \frac{1}{2}\sum_i \log(1 - T_i^2) \\ &-\; \frac{1}{2N}\sum_{\mu,ij}\xi_i^\mu\big(c_{ij} - m_i\,m_j\big)\xi_j^\mu - \frac{1}{2}\log\det A \\ &+\; \frac{1}{2N}\sum_\mu\left[\sum_i \xi_i^\mu\big(T_i - m_i\big)\right]^2 \;. \end{aligned} \tag{47}$$

The first order contribution to the cross entropy, Φ^1 in (37), is obtained by retaining the fourth order in y^μ in the expansion of the hyperbolic cosine function in (40),

$$\Phi^1 = \frac{1}{4N^2}\sum_i(1 - 4T_i^2 + 3T_i^4)\left(\sum_{\mu,\nu}\xi_i^\mu(A^{-1})^{\mu\nu}\xi_i^\nu\right)^2 \;. \tag{48}$$

We expect the differences $\Phi - \Phi^0$ and $\Phi - (\Phi^0 + \Phi^1)$ between, respectively, the true and the lowest order cross entropies and the true and the first order cross entropies to be of the order of, respectively, R^2 and R^3, where

$$R = \frac{p}{N}\,\xi^2\,(1 - m^2)\,\Lambda \;. \tag{49}$$

Here, ξ^2 is the order of magnitude of the pattern components, which can range from 1 if the patterns are extended over the whole system to $\sim\sqrt{N}$ for highly sparse

patterns, m is the typical value of the local magnetization, and Λ is the order of magnitude of the eigenvalues of A^{-1}, which can range from 1 to N. The value of R fixes the instrinsic error ϵ on the inferred patterns discussed in Section II G, $\epsilon \sim R$ for the lowest order approximation and $\epsilon \sim R^2$ with the first order corrections.

The above calculation can be straightforwardly extended to the case of the generalized Hopfield model by considering the $\hat{p}$ repulsive patterns as patterns with purely imaginary components, $\boldsymbol{\xi}^\mu = i\,\hat{\boldsymbol{\xi}}^\mu$, with $i^2 = -1$. For instance the general lowest order expression for the cross entropy is

$$\begin{aligned}
\Phi^0 &= -\sum_{i=1}^{N} m_i \tanh^{-1} T_i + N\log 2 + \frac{1}{2}\sum_i \log(1-T_i^2) \\
&- \frac{1}{2N}\sum_{ij}\left(c_{ij} - m_i\, m_j\right)\left(\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu - \sum_{\mu=1}^{\hat{p}}\hat{\xi}_i^\mu\hat{\xi}_j^\mu\right) \\
&+ \frac{1}{2N}\sum_{\mu=1}^{p}\left[\sum_i \xi_i^\mu\left(T_i - m_i\right)\right]^2 \\
&- \frac{1}{2N}\sum_{\mu=1}^{\hat{p}}\left[\sum_i \hat{\xi}_i^\mu\left(T_i - m_i\right)\right]^2 \\
&- \frac{1}{2}\log\det\begin{pmatrix} A & i\hat{A} \\ -i\hat{A}^T & \hat{\hat{A}}\end{pmatrix}\,, \qquad (50)
\end{aligned}$$

where

$$\begin{aligned}
T_i &= \tanh\left(\tanh^{-1} t_i + \sum_{\mu=1}^{p}\frac{(x^\mu)^*\xi_i^\mu}{\sqrt{N}} - \sum_{\mu=1}^{\hat{p}}\frac{(\hat{x}^\mu)^*\hat{\xi}_i^\mu}{\sqrt{N}}\right)\,, \\
(\hat{x}^\mu)^* &= \frac{1}{\sqrt{N}}\sum_i \hat{\xi}_i^\mu(T_i - t_i)\,, \\
\hat{A}^{\mu\nu} &= \frac{1}{N}\sum_i \xi_i^\mu\hat{\xi}_i^\nu\left(1-T_i^2\right)\,, \\
\hat{\hat{A}}^{\mu\nu} &= \delta^{\mu\nu} + \frac{1}{N}\sum_i \hat{\xi}_i^\mu\hat{\xi}_i^\nu\left(1-T_i^2\right)\,. \qquad (51)
\end{aligned}$$

The first order correction (48) can be easily written for the case of repulsive patterns, too.

B. Are the physical properties of the system relevant for the inference?

The Hopfield model was first introduced as a model for which a set of p desired ground states $\boldsymbol{\xi}^\mu$ (or fixed points of the zero temperature Glauber dynamics) could be programmed through an adequate choice of the interactions. Each fixed point has a basin of attraction in the configuration space, corresponding to a phase of the system. The order parameters are the overlaps

$$q^\mu = \sum_{\boldsymbol{\sigma}} P_H[\boldsymbol{\sigma}|\mathbf{h},\boldsymbol{\xi}]\,\left(\frac{1}{N}\sum_i \xi_i^\mu\,\sigma_i\right)\,, \qquad (52)$$

which quantify how much the configurations are on average aligned along each pattern. The amplitudes and directions of the pattern and the field vectors determine if spin configurations tend to be aligned along the field, or along one or more patterns. In the infinite size limit ($N\to\infty$) the overlaps are the roots of p coupled and self-consistent equations,

$$q^\mu = \lim_{N\to\infty}\frac{1}{N}\sum_i \xi_i^\mu\,\tanh\left(h_i + \sum_\rho q^\rho\xi_i^\rho\right)\,. \qquad (53)$$

Using (38) and the saddle point equation (41) it is easy to check that the overlaps

$$q^\mu = \frac{1}{N}\sum_i \xi_i^\mu\,T_i \qquad (54)$$

are solutions to the set of equations (53). Solutions are in one-to-one correspondance with the saddle points $(x^\mu)^*$.

The saddle-point solution $\mathbf{x}^* = 0$ corresponds to $T_i = t_i$. The average interaction term in the energy function (39) vanishes, meaning that configurations tend to be mainly determined by the fields. Such a behaviour corresponds to the paramagnetic phase. The solution $\mathbf{x} = 0$ is locally stable if the eigenvalues of the matrix A are all positive and, thus, if the patterns are weak enough. Solutions with $\mathbf{x}^* \neq 0$ correspond to stronger patterns and interaction terms in (39) having non zero values on average: they correspond to magnetized phases.

The cross entropy Φ depends on the solution $\mathbf{x}^*$ through the variables T_i only. Once the T_i's and the patterns $\boldsymbol{\xi}^\mu$'s are inferred, it is easy to calculate the value of the fields h_i based on equations (38), (41) and (42). One finds that h_i is given by (38) where t_i is substituted with T_i. Hence, the inferred parameters do not explicitely depend on the value of x^*. The procedure followed to infer the patterns and the fields is not affected by the physical phase (paramagnetic or magnetized) of the system, though the values of the data m_i and c_{ij} obviously depend on those physical properties.

It may accidentally happen that equations (41) have different solutions with equal or almost equal contributions to the partition function Z. The most natural illustration is the case of zero field ($t_i = 0$) and one strong pattern, where two ferromagnetic states with opposite overlaps, $(x^1)^*$ and $-(x^1)^*$, coexist. In this latter case both states give equal contributions to the partition function.

C. Maximum Likelihood inference: lowest order

We first infer the patterns and the pseudo-magnetizations from Φ^0. Minimization of Φ^0 (47) over $\mathbf{T}$ immediately shows that, up to $O(R)$ corrections, pseudo- and true magnetizations coincide:

$$(T_i)^* = m_i\,. \qquad (55)$$

Without loss of generality we may write the patterns to infer as

$$\begin{aligned}\left(\xi^0\right)_i^\mu &= \frac{\sqrt{N\,a^\mu}\;v_i^\mu + \sqrt{N}\beta_i^\mu}{\sqrt{1-m_i^2}}\;,\\ \left(\hat{\xi}^0\right)_i^\mu &= \frac{\sqrt{N\,\hat{a}^\mu}\;\hat{v}_i^\mu + \sqrt{N}\hat{\beta}_i^\mu}{\sqrt{1-m_i^2}}\;,\end{aligned} \tag{56}$$

where $\sqrt{a^\mu}, \sqrt{\hat{a}^\mu}$ are real-valued coefficients, and $\mathbf{v}^\mu$ and $\hat{\mathbf{v}}^\mu$ are eigenvectors of Γ. According to identity (55) the conditions (6) are fulfilled in the large N limit if the $(p+\hat{p})$ vectors $\boldsymbol{\beta}^\mu$ and $\hat{\boldsymbol{\beta}}^\nu$ are orthogonal to each other, and to all the patterns $(\boldsymbol{\xi}^0)^\nu$ and $(\hat{\boldsymbol{\xi}^0})^\nu$. The matrices A (45) and $\hat{A}$ (51) are then diagonal, while $\hat{A}$ vanishes. We rewrite the cross entropy (50) as

$$\begin{aligned}\Phi^0 &= -\sum_i \sum_{\sigma=\pm 1} \left(\frac{1+\sigma\, m_i}{2}\right) \log\left(\frac{1+\sigma\, m_i}{2}\right)\\ &- \frac{1}{2}\sum_\mu \lambda^\mu\, a^\mu - \frac{1}{2}\sum_{ij,\mu} \beta_i^\mu\, \Gamma_{ij}^{(r)}\, \beta_j^\mu\;,\\ &+ \frac{1}{2}\sum_\mu \hat{\lambda}^\mu\, \hat{a}^\mu + \frac{1}{2}\sum_{ij,\mu} \hat{\beta}_i^\mu\, \Gamma_{ij}^{(r)}\, \hat{\beta}_j^\mu\;,\\ &- \frac{1}{2}\sum_\mu \log\left[1 - a^\mu - \sum_i (\beta_i^\mu)^2\right]\\ &- \frac{1}{2}\sum_\mu \log\left[1 + \hat{a}^\mu + \sum_i (\hat{\beta}_i^\mu)^2\right]\end{aligned} \tag{57}$$

where $\Gamma^{(r)}$ is the restriction of Γ to the $(N-p-\hat{p})$–dimensional subspace orthogonal to the p largest and $\hat{p}$ smallest eigenvectors:

$$\Gamma_{ij}^{(r)} = \sum_{k=p+1}^{N-\hat{p}} \lambda^k v_i^k v_j^k\;. \tag{58}$$

Minimizing Φ^0 over the coefficients a^μ and the vectors $\boldsymbol{\beta}^\mu$ gives the coupled set of equations

$$\lambda^\mu = \frac{1}{1-a^\mu-b^\mu}\;, \tag{59}$$

$$\sum_j \Gamma_{ij}^{(r)}\, \beta_j^\mu = \frac{\beta_i^\mu}{1-a^\mu-b^\mu}\;, \tag{60}$$

where $b^\mu = (\boldsymbol{\beta}^\mu)^2$ is the squared norm of $\boldsymbol{\beta}^\mu$. If the vector $\boldsymbol{\beta}^\mu$ were non zero, it would be an eigenvector of Γ with eigenvalue λ^μ according to (60). This cannot be true as the largest eigenvalue of $\Gamma^{(r)}$ is smaller than λ^p. Hence, $\boldsymbol{\beta}^\mu = b^\mu = 0$. From (59) we obtain

$$a^\mu = 1 - \frac{1}{\lambda^\mu}\;. \tag{61}$$

We conclude that the maximum likelihood values for the p attractive patterns are given by (9). The minimization of Φ^0 over the coefficients $\hat{a}^\mu$ and the vectors $\hat{\boldsymbol{\beta}}^\mu$ can be done along the same lines. We find

$$\hat{a}^\mu = \frac{1}{\hat{\lambda}^\mu} - 1\;. \tag{62}$$

and $\hat{\beta}^\mu = 0$. The maximum likelihood estimators for the $\hat{p}$ repulsive patterns are given by (9) again. Once the patterns are computed the values of the local fields h_i are obtained from (11).

Notice that $v_i^\mu, \hat{v}_i^\mu$ are typically of the order of $N^{-\frac{1}{2}}$, which entails that the components of the patterns are of the order of unity. Though keeping each $\xi_i, \hat{\xi}_i$ of the order of unity is a natural scaling in the infinite size limit $N \to \infty$, other scalings are possible. Consider a pair of strongly coupled spins, *i.e.* such that the correlation Γ_{ij} is sizeably larger than $\frac{1}{N}$. According to expression (4) for the coupling J_{ij} induced by the patterns between spins i and j, we expect the pattern components to be of the order of $\sqrt{N}$. There is thus no compelling reason to assume that $\frac{\xi_i}{\sqrt{N}}, \frac{\hat{\xi}_i}{\sqrt{N}}$ is vanishingly small for all components i.

To end with we compute the decrease in cross entropy when adding a pattern attached to the eigenvalue λ ($= \lambda^\mu$ or $\hat{\lambda}^\mu$). Inserting expressions (61,62) for $a^\mu, \hat{a}^\mu$ in (57) we obtain

$$\Delta\Phi = -\frac{1}{2}\big(\lambda - 1 - \log\lambda\big)\;, \tag{63}$$

a quantity which is strictly negative for $\lambda \neq 1$. Not surprisingly, adding more parameters to the model allows for a better fit of the data. We will see in Section V E how the values of p and $\hat{p}$ can be determined.

D. Error bars on the patterns and fields

When the sample size B is large the posterior distribution P tends to a Gaussian law centered in the most likely values for the patterns, $\{\boldsymbol{\xi}^\mu\}, \{\hat{\boldsymbol{\xi}}^\mu\}$, and the pseudo-magnetizations, $\mathbf{T}$. For the sake of simplicty we consider below the case of attractive patterns only; repulsive patterns can formally be seen as purely imaginary attractive patterns, see Section V A. Let $\mathbf{H}$ denote the Hessian matrix of Φ^0. We find, to the leading orders,

$$\begin{aligned}(\mathbf{H}^{tt})_{ij} &\equiv \frac{\partial^2\Phi^0}{\partial T_i^1 \partial T_j^1} = \frac{\delta_{ij}}{1-m_i^2} - (J^0)_{ij}\;,\\ (\mathbf{H}^{\xi\xi})_{ij}^{\mu\nu} &\equiv \frac{\partial^2\Phi^0}{\partial (\xi^0)_i^\mu \partial (\xi^0)_j^\nu} = \frac{\delta^{\mu\nu}}{N}\Big[m_i m_j - c_{ij} + (1-m_i^2)\\ &\times\; \lambda^\mu\left(\delta_{ij} + (1-m_j^2)\sum_\rho \frac{\lambda^\rho}{N}(\xi^0)_i^\rho (\xi^0)_j^\rho\right)\Big]\\ &+ \frac{\lambda^\mu\lambda^\nu}{N^2}(1-m_i^2)(1-m_j^2)(\xi^0)_i^\nu (\xi^0)_j^\mu\;,\end{aligned} \tag{64}$$

$$(\mathbf{H}^{t\xi})_{ij}^{\nu} \equiv \frac{\partial^2\Phi^0}{\partial T_i \partial (\xi^0)_j^\nu} \simeq 0\;. \tag{65}$$

Here, δ denotes the Kronecker function and the expression of the lowest order coupling matrix, J^0, is given in (10). The sum over ρ runs over all pattern indices. The cross second derivative, $\mathbf{H}^{t\xi}$, of the order of $\frac{|\xi|}{N}$, is much smaller than the expected order, $\frac{|\xi|}{\sqrt{N}}$, and can be neglected.

The covariance matrix of the Gaussian posterior probability P is the inverse matrix of $B\,\mathbf{H}$. The inverse is properly defined in the subspace of dimension $N(p+\hat{p}+1)-\frac{1}{2}(p+\hat{p})(p+\hat{p}-1)$, orthogonal to the modes generating the invariance over the patterns, see Section II A. We write $\tilde{\mathbf{H}} = D\mathbf{H}D$, where D is a diagonal matrix with elements: $D_i = \sqrt{1-m_i^2}$ in the $\mathbf{T}$-sector, and $D_i^\mu = \sqrt{\frac{N}{1-m_i^2}}$ in the $\boldsymbol{\xi}^\mu$-sector. Matrix $\tilde{\mathbf{H}}$ has a particularly simple expression in the eigenbasis of the correlation matrix Γ, and can be diagonalized exactly after some simple algebra. We obtain the following expression for the covariance matrix of the fluctuations:

$$\langle \Delta T_i\, \Delta T_j \rangle = \frac{\sqrt{(1-m_i^2)(1-m_j^2)}}{B}\, \left[\mathbf{M}^{tt}\right]_{ij}\,, \qquad (66)$$

where

$$\left[\mathbf{M}^{tt}\right]_{ij} = \delta_{ij} + \sum_{\rho=1}^{p} (\lambda^\rho - 1)\, v_i^\rho\, v_j^\rho + \sum_{\rho=1}^{\hat{p}} (\hat{\lambda}^\rho - 1)\, \hat{v}_i^\rho\, \hat{v}_j^\rho\,. \quad (67)$$

The expressions for the fluctuations of the pattern components are reported in (13). Note that the cross-term $\langle \Delta T_i\, \Delta \xi_j^\nu \rangle$ vanishes at the expected order of $\frac{\sqrt{N}}{B}$, and is actually of the order of $\frac{1}{B}$ only. Using formula (38) we find that the error over the fields h_i is of the order of $\frac{p}{\sqrt{\alpha}}$, where $\alpha = \frac{B}{N}$.

E. Optimal number of patterns

So far we have assumed that the number of patterns, p, was known. In practice p is often determined based on simple criteria, such as how many eigenvalues 'come out' from the spectrum of the correlation matrix (Section VI B 2). Alternative approaches exist, *e.g.* Bayesian Information Criterion (BIC) [26]. In the BIC the decrease $B\Delta\Phi$ (63) in cross entropy obtained with a new pattern is added a 'cost' $N \log B$, equal to the number of new parameters times the logarithm of the number of data. As the index μ increases the selected eigenvalue λ^μ or $\hat{\lambda}^\mu$ gets closer to one; $B|\Delta\Phi|$ (63) decreases in absolute value, and, eventually, is counterbalanced by the cost term $N \log B$. The value of μ for which the two terms balance each other depends on the size of the data set: the higher B, the more significative are the correlations and the more patterns we need to represent the interactions. However BIC is mathematically justified when B is large compared to N, which is not always the case in real data sets.

Hereafter, we propose a different approach based on Bayesian and geometric considerations. Based on the discussion in Section II D we expect the squared norm b^μ of the transerve fluctuations $\boldsymbol{\beta}^\mu$ to be non vanishing in the $B, N \to \infty$ limits. Let us call a^μ the squared projection of the μ^{th} rescaled pattern onto $\mathbf{v}^\mu$ (16). The same quantities, $\hat{a}^\nu$ and $\hat{b}^\nu$, can be defined for repulsive patterns. We define the marginal probability P_M of the squared projections $a^\mu, \hat{a}^\nu$ and of the squared norms $b^\mu, \hat{b}^\nu$ through

$$\begin{aligned} P_M \;=\; & \int \prod_{\mu,i} \frac{d\beta_i^\mu}{\sqrt{1-m_i^2}} \prod_{\nu,i} \frac{d\hat{\beta}_i^\nu}{\sqrt{1-m_i^2}} \prod_\mu \frac{d\Omega^\mu}{\pi i \alpha N/2} \\ & \times \prod_\nu \frac{d\hat{\Omega}^\nu}{\pi i \alpha N/2}\; \exp\Big[-\frac{\alpha}{2} \sum_\mu \Omega^\mu \left((\boldsymbol{\beta}^\mu)^2 - N\, b^\mu \right) \Big] \\ & \times \exp\Big[-\frac{\alpha}{2} \sum_\nu \hat{\Omega}^\nu \left((\hat{\boldsymbol{\beta}}^\nu)^2 - N\, \hat{b}^\nu \right) \Big] \qquad (68) \\ & \times P\Big[\{ T_i^0, \frac{\sqrt{N\,a^\mu}\; v_i^\mu + \sqrt{N}\beta_i^\mu}{\sqrt{1-m_i^2}}, \frac{\sqrt{N\,\hat{a}^\nu}\; \hat{v}_i^\nu + \sqrt{N}\hat{\beta}_i^\nu}{\sqrt{1-m_i^2}} \} \Big]\,, \end{aligned}$$

where P is the posterior probability (5), and the sums over μ and ν run from 1 to, respectively, p and $\hat{p}$. After carrying out the integrals over the fluctuations $\boldsymbol{\beta}^\mu$ and $\hat{\boldsymbol{\beta}}^\nu$ we obtain

$$\begin{aligned} P_M \;=\; & \frac{1}{Z_1} \int \prod_\mu d\Omega^\mu \prod_\nu d\hat{\Omega}^\nu \qquad (69) \\ & \times \exp\Big[-\frac{B}{2} \sum_\mu \Delta\Phi_M(\Omega^\mu) - \frac{B}{2} \sum_\nu \Delta\hat{\Phi}_M(\hat{\Omega}^\nu) \Big] \end{aligned}$$

where Z_1 is a normalization constant and

$$\begin{aligned} \Delta\Phi_M(\Omega^\mu) \;=\; & \lambda^\mu\, a^\mu + \Omega^\mu\, b^\mu + \log(1 - a^\mu - b^\mu) \qquad (70) \\ & - \frac{1}{B} \log\det\left[\Omega^\mu\, \mathbf{1} - \Gamma^{(r)} \right] + O\big(\frac{\log N}{N}\big)\,, \\ \Delta\hat{\Phi}_M(\hat{\Omega}^\nu) \;=\; & -\hat{\lambda}^\nu\, \hat{a}^\nu + \hat{\Omega}^\nu\, \hat{b}^\nu + \log\Big(1 + \hat{a}^\nu + \hat{b}^\nu\Big) \qquad (71) \\ & - \frac{1}{B} \log\det\left[\hat{\Omega}^\nu\, \mathbf{1} + \Gamma^{(r)} \right] + O\big(\frac{\log N}{N}\big)\,, \end{aligned}$$

Here $\mathbf{1}$ denotes the N-dimensional identity matrix. When B is large the integrals in (69) are dominated by the contributions coming from the vicinity of the roots of

$$\frac{\partial \Delta\Phi_M}{\partial \Omega^\mu} = \frac{\partial \Delta\hat{\Phi}_M}{\partial \hat{\Omega}^\nu} = 0\,. \qquad (72)$$

Maximimization of $\Delta\Phi_M$ with respect to the a^μ, b^μ's gives equations (59) and

$$\Omega^\mu = \lambda^\mu\,, \qquad (73)$$

for each $\mu = 1, \ldots, p$. We then compute the squared norm b^μ from the extremization condition (72) and obtain

$$b^\mu \;=\; \frac{1}{B} \sum_{k=p+1}^{N-\hat{p}} \frac{1}{\lambda^\mu - \lambda^k}\,, \qquad (74)$$

$$a^\mu \;=\; 1 - \frac{1}{\lambda^\mu} - b^\mu\,. \qquad (75)$$

Repeating the same procedure to maximize $\Delta\hat{\Phi}_M$ gives

$$\hat{b}^\nu = \frac{1}{B}\sum_{k=p+1}^{N-\hat{p}} \frac{1}{\lambda^k - \hat{\lambda}^\nu} ,$$
$$\hat{a}^\nu = \frac{1}{\hat{\lambda}^\nu} - 1 - \hat{b}^\nu . \qquad (76)$$

The difference between expressions (61) and (75) for the coefficients a^μ must be emphasized. P defined in (5) is a probability density over pN pattern components, once the pseudo-magnetizations have been inferred. Maximization of P, or, equivalently, of Φ over this large-dimensional space gives expression (61) for the projection a^μ of the pattern $\boldsymbol{\xi}^\mu$ onto the μ^{th} largest eigenvector of Γ, $\mathbf{v}^\mu$. Instead of directly maximizing P, we may first integrate out the orthogonal fluctuations to $\mathbf{v}^\mu$ in P, and obtain the marginal probability density P_M for $2p$ parameters only, namely the squared projections on the eigenvectors, a^μ, and the squared norms of the orthogonal fluctuations, b^μ. Maximizing the marginal probability density P_M or, equivalently, minimizing Φ_M shows that b^μ (75) does not vanish, and that the value of the squared projection a^μ (75) is smaller than (61). Figure 1 sketches the geometrical meaning of the coefficient $\sqrt{a^\mu}$ and the fluctuations $\boldsymbol{\beta}^\mu$, see (16). Small values of the angle θ^μ are expected for reliable patterns. A similar picture can be drawn for repulsive patterns. We will see how expression (75) for the squared norm b^μ naturally arises in the context of random matrix theory.

F. Maximum likelihood inference: first corrections

We now look for the corrections to the lowest order expressions of the patterns and the fields (9,55), encoded in expressions (8) and $T_i = T_i^0 + T_i^1$. The first order contribution to the cross entropy, Φ^1, can be seen as a perturbation to the lowest order cross entropy, Φ^0, according to (37). Within linear response theory this perturbation will shift the maximum likelihood estimators by

$$\begin{pmatrix} \mathbf{T^1} \\ \{(\boldsymbol{\xi}^1)^\mu\} \\ \{(\hat{\boldsymbol{\xi}}^1)^\mu\} \end{pmatrix} = -\big(\mathbf{H}\big)^{-1} \begin{pmatrix} \frac{\partial\Phi^1}{\partial\mathbf{T}} \\ \{\frac{\partial\Phi^1}{\partial\boldsymbol{\xi}^\mu}\} \\ \{\frac{\partial\Phi^1}{\partial\hat{\boldsymbol{\xi}}^\mu}\} \end{pmatrix} , \qquad (77)$$

where the inverse of the Hessian matrix of Φ^0, $\mathbf{H}$, was given in Section V D. The calculation of the gradient of Φ^1 does not present any particular difficulty. The resulting corrections to the patterns are given in eqn (23). The expression for the shift in the pseudo-magnetization is

$$T_i^1 = \sum_{\mu=1}^{p} (\lambda^\mu - 1)\left[C^\mu\, v_i^\mu \sqrt{1-m_i^2} + m_i\,(v_i^\mu)^2\right] \qquad (78)$$
$$+ \sum_{\mu=1}^{\hat{p}} (\hat{\lambda}^\mu - 1)\left[C^{N+1-\mu}\, \hat{v}_i^\mu \sqrt{1-m_i^2} + m_i\,(\hat{v}_i^\mu)^2\right] .$$

where C^k is given in (26). Notice that, if the magnetizations m_i vanish, so do the dominant and first-order contributions to the pseudo-magnetizations.

VI. RELIABILITY OF THE INFERENCE

An important issue is to determine how many configurations should be sampled in order to ensure that the inference of the patterns is accurate. To do so, we assume that the examples $\boldsymbol{\sigma}^b$ are drawn independently and at random from the equilibrium probability P_H (2) of a Hopfield model, with fixed fields $\tilde{\mathbf{h}}$ and patterns $\tilde{\boldsymbol{\xi}}$. We call $S[\{\boldsymbol{\sigma}^b\}]$ the entropy of the posterior distribution P (5) for the fields $\mathbf{h}$ and patterns $\boldsymbol{\xi}$. In the large N limit, we expect this entropy to be self-averaging, that is, to depend on the set of examples only through their number B. We want to determine how fast S decays with B. To do so it is instructive to first consider the simple case where the local fields are known, and only one pattern has to be inferred. This specific situation is treated in great analytical details in Section VI A. The general (and harder) case where both fields and patterns have to inferred is treated in Section VI B.

A. Case of one unknown pattern and known fields

Throughout this Section, we assume that the local fields vanish, $\tilde{\mathbf{h}} = 0$ and that the number of patterns to be inferred is $p = 1$. The posterior entropy,

$$S[\{\boldsymbol{\sigma}^b\}] = -\sum_{\{\xi_i = \pm\tilde{\xi}\}} P[0,\boldsymbol{\xi}|\{\boldsymbol{\sigma}^b\}]\, \log P[0,\boldsymbol{\xi}|\{\boldsymbol{\sigma}^b\}] , \qquad (79)$$

therefore measures the uncertainty about this unique pattern given a set B sampled configurations. Intuitively, the dependence of S on B is closely related to the physics of the Hopfield model (with pattern $\tilde{\boldsymbol{\xi}}$ and zero fields) used to generate the examples. If the model is in the paramagnetic phase, *i.e.* if the components of the pattern are weak [27], the examples $\boldsymbol{\sigma}^b$ have vanishingly small overlap (52) with the pattern. We expect that a large number B (diverging with N) of examples is necessary to convey reliable information about the pattern. Conversely, few configurations sampled in a ferromagnetic state around a strong pattern (or its opposite) should be sufficient to reconstruct the pattern.

We now make this scenario quantitative in various cases. An important simplication arises when the pattern is restricted to have binary components, $\tilde{\boldsymbol{\xi}} = \{\tilde{\xi}_i = \pm\tilde{\xi}\}$, with $\tilde{\xi} > 0$. Hamiltonian (3) with $p = 1$ pattern is invariant under the exchange of the spin configuration and the pattern: $E[\boldsymbol{\sigma}, 0, \boldsymbol{\xi}] = E[\boldsymbol{\xi}, 0, \boldsymbol{\sigma}]$. Our inference problem can thus be mapped onto a dual Hopfield model, where the normalized inferred pattern, $\boldsymbol{\xi}/\tilde{\xi}$, plays the role of the dual spin configuration and the sampled spin configurations, $\boldsymbol{\sigma}^b$, $b = 1, \ldots, B$ correspond to the B dual

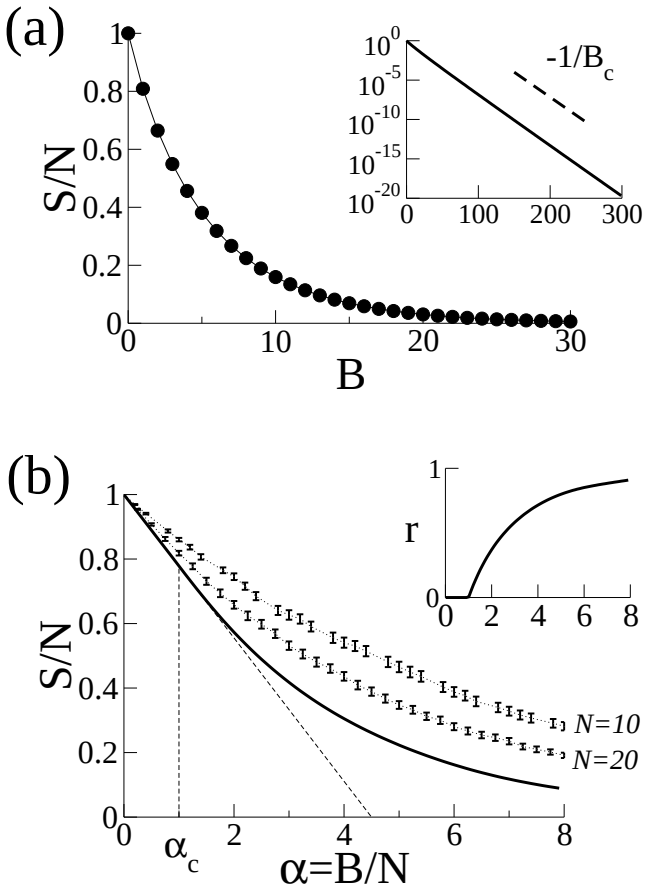

FIG. 17: Entropy of the posterior distribution for the patterns, S (in bits and per component), as a function of the number of sampled configurations, B, when the local fields h_i are known to vanish. **(a).** Ferromagnetic regime ($\tilde{\xi}^2 = 1.1$): the entropy decays exponentially with B. Inset: comparison with the theoretical prediction $\exp(-B/B_c)$ (dashed line), with $B_c \simeq 6.85$, in semi-log scale. **(b).** Paramagnetic regime ($\tilde{\xi}^2 = .5$): S (86) is a decreasing function of $\alpha = B/N$. The entropies calculated from numerical calculations are shown for $N = 10$ and $N = 20$. Inset: the overlap r (83) between the inferred and true patterns is positive when α exceeds $\alpha_c = 1$ (87).

patterns. In particular, the posterior entropy S is equal to the entropy of the dual Hopfield model at inverse temperature

$$\beta = \tilde{\xi}^2 \ . \tag{80}$$

The duality property allows us to exploit the well-understood physics of the Hopfield model [27] to simplify the study of our inference problem.

1. Strong components

In the ferromagnetic regime ($\tilde{\xi} > 1$), the dual spin configuration is strongly magnetized along the dual patterns. Going back to the inference problem, we find that the overlap between the inferred pattern and a sampled configuration,

$$q^b = \sum_{\{\boldsymbol{\sigma}^b\},\boldsymbol{\xi}} P[0,\boldsymbol{\xi}|\{\boldsymbol{\sigma}^b\}] \prod_b P_H[\boldsymbol{\sigma}^b,\tilde{\boldsymbol{\xi}}] \ \frac{1}{N}\sum_i \xi_i \sigma_i^1 \ , \tag{81}$$

may take values $+q$ or $-q$, where q is the positive root of $q = \tanh(q\,\tilde{\xi}^2)$. The sign of the overlap q^b is random, depending on which one of the two states with opposite magnetizations the configuration $\boldsymbol{\sigma}^b$ in sampled in; it is equal to $+$ or $-$ with equal probabilities $\frac{1}{2}$. These statements hold if the thermodynamical limit, $N \to \infty$, is taken while B is kept fixed. We find that S is equal to the entropy of a single spin at inverse temperature β, interacting with B other spins of magnetization q,

$$S = \sum_{b=0}^{B} \binom{B}{b} \left(\frac{1+q}{2}\right)^b \left(\frac{1-q}{2}\right)^{B-b} \mathcal{S}\big((B-2b)q\tilde{\xi}^2\big) \ , \tag{82}$$

where $\mathcal{S}(u) = \log(2\cosh u) - u\tanh u$. Figure 17A shows that the entropy is almost a pure exponential: $\log S \simeq -B/B_c$ where the decay constant, $B_c = 1/\log\cosh(q\,\tilde{\xi}^2)$, is finite (compared to N). In the ferromagnetic regime few sampled configurations are sufficient to determine $\tilde{\boldsymbol{\xi}}$ accurately.

This result also applies to the case of a single ferromagnetic state. If the field $\mathbf{h}$ does not strictly vanish and explicitely breaks the reversal symmetry between the two states, all configurations are sampled from the same state, with probability $1 - \exp(-O(N))$. Remarkably, expression (82) for the entropy still holds. Again we find that $B = O(1)$ configurations are sufficient to infer the pattern. We will discuss in more details the inference in the ferromagnetic regime in Sections VI B 1 and VI B 3.

2. Weak components

In the paramagnetic phase ($\tilde{\xi} < 1$), the overlap (81) between the inferred pattern and an example is typically very small, $q \sim N^{-1/2}$. No inference is possible unless the number of examples, B, scales linearly with N; we denote $\alpha = B/N$. In this regime, we expect the entropy to be self-averaging: $S[\{\boldsymbol{\sigma}^b\}]$ does not depend on the detailed composition of the data set and is a function of the value of the macroscopic parameters, *e.g.* the ratio α, only. To calculate this function S we use the replica method [16, 27]. We report below the results of the replica symmetric calculation; technical details can be found in Appendix. The order parameter is the average overlap r between the inferred and the true patterns,

$$r = \sum_{\{\boldsymbol{\sigma}^b\},\boldsymbol{\xi}} P[0,\boldsymbol{\xi}|\{\boldsymbol{\sigma}^b\}] \prod_b P_H[\boldsymbol{\sigma}^b,\tilde{\boldsymbol{\xi}}] \ \frac{1}{N}\sum_i \xi_i \tilde{\xi}_i \ . \tag{83}$$

which is solution of the self-consistent equation

$$r = \int_{-\infty}^{\infty} Dz \ \tanh(z\sqrt{\gamma} + \gamma) \ , \tag{84}$$

where $Dz = \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}$ is the Gaussian measure, and

$$\gamma = \frac{\alpha\beta^2 r}{(1-\beta)(1-\beta+\beta r)} \ . \tag{85}$$

The posterior entropy is equal to

$$\begin{aligned} S &= \int_{-\infty}^{\infty} Dz \, \log 2\cosh(z\sqrt{\gamma}+\gamma) - \frac{\alpha}{2}\log(1-\beta+\beta r) \\ &- \frac{\alpha\beta(1-\beta-r+3\beta r)}{2(1-\beta)(1-\beta+\beta r)} \ , \end{aligned} \tag{86}$$

and is plotted in Fig. 17B. To check this analytical prediction we have run extensive numerical simulations on small-size systems ($N = 10, 20$). The numerical procedure follows three steps: 1. evaluate the partition function Z in (2) through an exact enumeration; 2. generate a data set of $B = \alpha N$ configurations $\{\sigma_i^b\}$ according to the Hopfield measure P_H by rejection sampling; 3. evaluate P_1 in (5) and S in (79) through exact enumerations. The resulting entropy, averaged over one hundred data sets, is compatible with the analytical prediction and the existence of $\frac{1}{N}$ finite-size effects.

Inset of Fig. 17B shows that the overlap r remains null until α reaches the critical value

$$\alpha_c = \left(\frac{1}{\tilde{\xi}^2} - 1\right)^2 \ . \tag{87}$$

Hence, in the range $[0; \alpha_c]$, the posterior probability becomes more concentrated (S decreases), but not around the true pattern $\tilde{\boldsymbol{\xi}}$. The existence of a lagging phase before any meaningful inference is possible is similar to the 'retarded learning' phenomenon discovered in the field of unsupervised learning, where the variables to be learned are real-valued [28–30]. In the present case of binary spins we expect the replica symmetric assumption to break down at large α. The entropy (86) indeed becomes negative when $\alpha > \alpha_0 \simeq 42$ for the case studied in Fig. 17B. Nevertheless we may conjecture that the entropy decays as $S \sim \frac{1}{\alpha}$ when $\alpha \to \infty$. The dual Hopfield model has random couplings J_{ij}, with second moment equal to $\langle J_{ij}^2 \rangle - \langle J_{ij} \rangle^2 = \frac{\alpha}{N}$. Hence $T = \frac{1}{\sqrt{\alpha}}$ sets the temperature scale of the dual model. The low temperature scaling of the entropy of the Sherrington-Kirkpatrick (SK) model suggests that $S \propto T^2$ [31]; this scaling is compatible with the small–N results of Fig. 17B. However the dual and SK models are not strictly identical when $\alpha \to \infty$: the coupling matrix $\mathbf{J}$ of the dual model is guaranteed to be semidefinite positive, while the entries of $\mathbf{J}$ are independent in the SK model. A complete calculation of the entropy valid for any (large) α would require a replica symmetry broken Ansatz for the order parameters [32], and is beyond the scope of this article.

Note that the calculations above can be extended to real patterns; β in (80) is then replaced with $\langle \xi^2 \rangle$, where the average is taken over the pattern components. The entropy is not constrained to be positive as in the binary case. The distinction between the strong- and weak-component regimes remains qualitatively unchanged, and so does the value of the critical ratio α_c (87), which does not depend on the third and higher moments of $\tilde{\xi}_i$.

B. General case of unknown patterns and fields

In this Section, we first interpret the above results. We show that, while $B = O(1)$ configurations can be sufficient in a particular context, $B = O(N)$ data are generally necessary for the inference to be sucessful. The connection between the results of Section VI A and random matrix theory are emphasized.

1. *Inference from the magnetizations*

Consider first the case where a single state exists, *i.e.* equations (53) admit a single solution $\{q^\mu\}$; the case where states coexist will be discussed in Section VI B 3. For large N, the average value of spin i with the measure P_H (2) is

$$m_i = \tanh\left(h_i + \sum_\mu q^\mu \, \xi_i^\mu\right) \ . \tag{88}$$

As the error on the estimate of m_i decreases as $\sim \sqrt{\frac{1-m_i^2}{B}}$ with B, $O(1)$ configurations are sufficient to sample the magnetizations accurately. Few sampled configurations therefore give access to the knowledge of a linear combination of the field vector and pattern vectors with non zero-overlaps q^μ. This linear combination is simply T_i^0, and equation (88) coincides with (55).

When the fields h_i are known and the model consists of a single strong pattern ($p = 1$) the pattern components ξ_i^1 can be readily calculated from the magnetizations (88) through

$$\xi_i^1 = \frac{1}{q} \, \tanh^{-1} m_i \quad \text{where} \quad q^2 = \frac{1}{N}\sum_j m_j \tanh^{-1} m_j \ . \tag{89}$$

This particular case was encountered at the end of Section VI A 1, when the fields h_i are sent to zero after having broken the reversal symmetry of the system to avoid state coexistence. In the generic situation of unknown fields and patterns, knowledge of the magnetizations does not suffice to determine the field and the patterns, and must be supplemented with the information coming from the correlation matrix Γ_{ij}.

2. *Inference from the correlations: relationship with random matrix theory*

What is the order of magnitude of Γ_{ij}? We first consider the ideal case of perfect sampling ($B \to \infty$ while N is large but finite). As a result of the presence of the

patterns in the energy (3) the spins are correlated. The entries of the correlation matrix are, for large N [42],

$$\Gamma_{ij} = \delta_{ij} + \frac{1}{N} \frac{\xi_i \xi_j \sqrt{(1-m_i^2)(1-m_j^2)}}{1 - \frac{1}{N}\sum_k \xi_k^2 (1-m_k)^2} \tag{90}$$

where we have considered the case of a single pattern ($p = 1, \hat{p} = 0$) to lighten notations. Though the pattern affects each correlation Γ_{ij} by $O(\frac{1}{N})$ only, these small contributions add up to boost the largest eigenvalue from one (in the absence of pattern) to

$$L = \frac{1}{1 - \frac{1}{N}\sum_k \xi_k^2 (1-m_k)^2} \ . \tag{91}$$

The eigenvector attached to L has components $v_i \propto \xi_i \sqrt{1-m_i^2}$ and ML inference perfectly recovers the pattern.

In the presence of sampling noise (finite B), each correlation (90) is corrupted by a stochastic term of the order of $x = \frac{1}{\sqrt{B}}$. This stochastic term will, in turn, produce an overall contribution of the order of $x\sqrt{N} = \frac{1}{\sqrt{\alpha}}$ to the largest eigenvalue. Intuitively, whether α is large or small compared to L^{-2} should tell us how hard or easy it is to extract the pattern $\boldsymbol{\xi}$ from Γ. Several studies in the physics [33, 34] and in the mathematics [35] literatures have indeed found that an abrupt phase transition takes place at the critical ratio

$$\alpha_c = \frac{1}{(L-1)^2} \ . \tag{92}$$

It is a simple check that α_c coincides with the ratio (87) for the retarded learning transition calculated in Sections VI A 2.

In the strong noise regime ($\alpha < \alpha_c$) the largest eigenvector $\mathbf{v}^1$ of Γ is uncorrelated with (orthogonal to) the pattern $\boldsymbol{\xi}$, and the spectrum of Γ is identical to the one of the sample correlation matrix of independent spins, whose density of eigenvalues is given by the Marcenko-Pastur (MP) law,

$$\rho_{MP}(\lambda') = v(1-\alpha)\,\delta(\lambda') + \frac{\alpha}{2\pi\lambda'} \sqrt{v\big((\lambda_+ - \lambda')(\lambda' - \lambda_-)\big)} \tag{93}$$

with $v(u) = \max(u, 0)$ [19]. The edges of the continuous component of the MP spectrum are given by

$$\lambda_\pm = \left(1 \pm \frac{1}{\sqrt{\alpha}}\right)^2 \ . \tag{94}$$

The largest eigenvalue of Γ, λ_+, is not related to the value of L.

In the weak noise regime ($\alpha > \alpha_c$) the largest eigenvalue of Γ is [35]

$$\lambda^1 = L\,\left(1 + \frac{1}{\alpha\,(L-1)}\right) \ . \tag{95}$$

It exceeds L for any finite α, and converges to L when $\alpha \to \infty$. The rest of the spectrum is described by the MP density (93). Expression (74) for the squared norm b^1 of the orthogonal fluctuations leads to the analytical formula

$$b^1 = \frac{1}{\alpha} \int_{\lambda_-}^{\lambda_+} d\lambda'\ \frac{\rho_{MP}(\lambda')}{\lambda^1 - \lambda'} = \frac{\lambda^1 - L}{\lambda^1} \ , \tag{96}$$

where we have used the analytical expression of the Stieltjes transform of ρ_{MP} [13]. Using (75) we deduce the value of the squared projection of the inferred rescaled pattern $(\boldsymbol{\xi}^1)'$ onto $\mathbf{v}^1$,

$$a^1 = \frac{L-1}{\lambda^1} \ . \tag{97}$$

Identities (96) and (97) are graphically interpreted in Fig. 1: b^1 is the squared norm of the orthogonal fluctuations $\boldsymbol{\beta}$, while a^1 is the squared projection of the rescaled pattern $\boldsymbol{\xi}$ onto $\mathbf{v}^1$.

The above discussion is illustrated on the simple case of a Hopfield model with $p = 1, \hat{p} = 0$ patterns in Fig. 18, see caption for the description of the model. Using formula (91) we compute the largest eigenvalue of the correlation matrix for perfect sampling, $L = 2$. Figure 18 shows that a large eigenvalue clearly pulls out from the bulk spectrum for the ratio $\alpha = 4$ (top spectrum), larger than the critical ratio $\alpha_c = 1$ according to (92) (bottom). For $\alpha = 4$, the infinite–N predicted values for the largest eigenvalue, $\lambda_1 = 2.5$ (95), and for the edges of the MP spectrum, $\lambda_- = .25, \lambda_+ = 2.25$ (94), are in good agreement with the numerical results for $N = 100$.

Formulae (96) and (97) hold for each pattern μ when $p \geq 2$ patterns are present, provided that p remains finite when $N \to \infty$. The case of $p = 2$ patterns, where one pattern is strong and has overlap $q > 0$ (81) with the sampled configurations, and the second pattern has weak components, is of particular interest. Again, we assume that the fields vanish. Repeating the calculation of Section VI A 2 and Appendix A we find that the entropy S/N quickly decreases with B from 2 bits down to 1 for $B = O(1)$. When $B \propto N$, the entropy decreases from 1 down to 0; the expression of S coincides with (86) where β is replaced with $\beta(1-q^2)$. Hence we have a two-step behaviour: the strong pattern is determined with $O(1)$ examples, the weak pattern requires $O(N)$ sampled configurations. Learning of the weak pattern is possible if

$$\alpha \geq \left(\frac{1}{\tilde{\xi}^2(1-q^2)} - 1\right)^2 \ , \tag{98}$$

according to (87). The two-step behaviour agrees with the discussion of Section VI B 1.

3. *Coexistence of ferromagnetic states*

Consider now the case of the coexistence of two ferromagnetic states exposed in Section V B. Data are generated from a Hopfield model, with zero fields and one

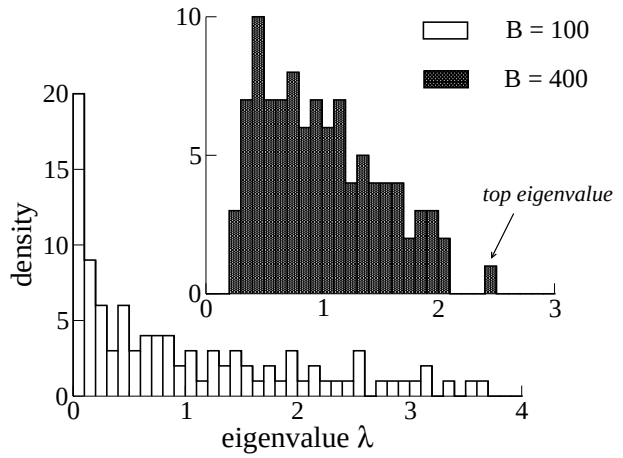

FIG. 18: Spectrum of the correlation matrix for a Hopfield model with $p = 1$ pattern, $N = 100$ spins, and for $B = 100$ (bottom) and 400 (top) randomly sampled configurations at equilibrium. The bulk parts of the spectra coincide with the Marcenko-Pastur law for random correlation matrices. When B is large the top eigenvalue clearly comes out from the noisy bulk and the corresponding eigenvector approximately corresponds to the pattern. The pattern components are i.i.d. Gaussian variables, of zero mean and variance $\xi^2 = .5$; local fields h_i have zero values.

strong pattern $\boldsymbol{\xi}$, as in Fig. 4. In the up-state the spins are magnetized with $m_i^+ = \tanh(q\,\xi_i)$. In the down-state the local magnetization is $m_i^- = -m_i^+$. On the overall the local magnetization is $m_i = \frac{1}{2}\,m_i^+ + \frac{1}{2}\,m_i^- = 0$, up to $O(\frac{1}{\sqrt{B}})$ fluctuations. The discrepancy between the Gibbs magnetizations, $m_i = 0$, and the state magnetizations, $m_i^\pm$, results in a $O(1)$ contribution $m_i^+ m_j^+ (= m_i^- m_j^-)$ to the correlation matrix entry Γ_{ij}, dominating the $O(\frac{1}{N})$ contributions due to the interactions between spins. The largest eigenvalue of Γ,

$$\lambda^1 = \sum_i (m_i^+)^2 \ , \tag{99}$$

is of the order of N; the corresponding eigenvector is $\mathbf{v}^1 = (m_1^+, m_2^+, \ldots, m_N^+)/\sqrt{\lambda^1}$. Informally speaking, the information about the state magnetizations is not conveyed by the Gibbs magnetizations (as in Section VI B 1) but by the correlation matrix [36]. According to formula (55) the pseudo-magnetization T_i vanishes; hence we correctly infer that the fields h_i have zero values. Using formula (9) we obtain

$$(\xi^0)_i \simeq \sqrt{\frac{N}{\lambda^1}}\, m_i^+ \ . \tag{100}$$

Therefore, the inferred pattern component is not equal to the true pattern component, but is proportional to its hyperbolic tangent. This non linear transform is clearly seen in Fig. 4. The discrepancy between the true and inferred components is a nice illustration of the claimed scaling for the higher order corrections in (49) (recall that the eigenvalues of A^{-1} are the p largest eigenvalues of Γ). In the presence of coexistent states, while ξ^2 is small compared to N, λ^1 is of the order of N, making the ratio $\frac{\lambda^1 \boldsymbol{\xi}^2}{N}$ of the order of unity. Corrections are required and shown to improve the quality of the inferred pattern in Fig. 11.

VII. CONCLUSION

In this paper we have studied how to infer a small-rank interaction matrix between N binary variables given the average values and pairwise correlations of those variables. We have seen that the generalized Hopfield model, where the interactions are encoded into a set of attractive and repulsive patterns $\boldsymbol{\xi}$, is a natural framework for Maximum Likelihood (ML) inference. Using techniques from the statistical physics of disordered systems, we have presented a systematic expansion of the log-likelihood in powers of $\lambda\frac{\boldsymbol{\xi}^2}{N}$, where λ is the largest eigenvalue of the correlation matrix Γ (1). We have then calculated the ML estimators for the patterns and the fields to the lowest and first order in this expansion in a variety of physical regimes. The lowest order is a simple extension of Principal Component Analysis, where not only the largest but also the smallest eigenmodes build in the interactions. First order corrections involve non-linear combinations of the eigenvalues and eigenvectors of Γ. We have validated our ML expressions for the patterns on synthetic data generated by Hopfield models with known patterns and fields, and by Ising models with sparse interactions. We have also presented a simple geometrical criterion for deciding the number of patterns. Those results have been discussed and compared to previous studies in the unsupervised learning and random matrix literatures.

The quality of the inference strongly depends on the number of sampled configurations, B. The sampling error on each magnetization, m_i, and pairwise correlation, c_{ij}, is of the order of $B^{-1/2}$. Elementary insights from random matrix theory suggest that the resulting errors on the eigenvectors of the matrix Γ are $\sqrt{N}$ times larger. The error on the inferred patterns, ϵ, picks up a contribution $\sim \left(\frac{N}{B}\right)^{1/2}$ due to finite sampling, as found in Section II C. This scaling has several important consequences. First, inference is retarded: no information about the true couplings can be obtained unless the ratio $\frac{B}{N}$ exceeds a critical value (Sections VI A 2 and VI B 2). Secondly, for larger B, ϵ decreases as $B^{-1/2}$, which is confirmed by the simulations presented in Fig. 12, and then saturates to the intrinsic error resulting from our approximate expressions for the patterns. The intrinsic error depends on the order in the expansion used for the calculation of the cross-entropy in Section V. Note that other inference methods, looking for the local structure of the interaction network [11, 12], may unveil strong cou-

plings $J = O(1)$ from a much smaller number of sampled configurations, $B = O(\log N)$, and do not suffer from the retarded learning transition.

Our study could be extended in several directions. It would be particularly interesting to consider the case of spins taking $Q > 2$ values (Potts model), *e.g.* for applications to the study of coevolution between residues in protein sequences [23, 25, 37]. Mean-field inference methods provide a simple and efficient way to get interactions from correlations [38]. Knowing how MF interactions are modified when some eigenmodes are rejected (using the criterion of Section II D) or first-order corrections are taken into account would be of interest. However the linear increase in the number of possible symbols with Q ($= 20$ for amino-acids) may make the effective size of the problem, $N \times Q$, larger than the number of configurations, B, in practical applications. A large number of vanishing eigenvalues is expected in those cases, and extracting repulsive patterns may become a difficult task.

Appropriate priors P_0 could also be used to force many pattern components to identically vanish, instead of acquiring small values as in Section II E. This can be particularly useful when the true patterns are known to be highly sparse and few data are available. Inspired by the so-called Lasso regression method [39], a natural prior is

$$P_0 \propto \exp\left[-\gamma \sum_{i=1}^{N} \sqrt{1-m_i^2}\left(\sum_{\mu=1}^{p} |\xi_i^\mu| + \sum_{\mu=1}^{\hat{p}} |\hat{\xi}_i^\mu|\right)\right] . \tag{101}$$

Contrary to the case of the quadratic penalty (21) the most likely values for the patterns cannot be expressed by means of simple analytical formulae. However, they could be efficiently obtained using convex optimization algorithms minimizing the sum of the cross entropy and of the penalty term (101).

Last of all, we have considered in this work that the configurations were sampled at equilibrium. In practice, when more than one state exist, the equilibration time may be prohibitive and a reasonable assumption would be to sample from one state only. To what extent ergodicity breaking in the sampling affects the quality of inference is an interesting question.

Acknowledgments: We thank S. Leibler for numerous discussions. V.S. thanks the Simons Center for Systems Biology for its hospitality. This work was partially funded by the ANR contract 06-JC-JC-051.

APPENDIX A: REPLICA CALCULATION OF THE ENTROPY S FOR WEAK PATTERNS

When the pattern has binary components $\tilde{\xi}_i = \pm\tilde{\xi}$ we make the change of variables $\sigma_i' = \xi_i \sigma_i$ to rewrite the partition function (35) of the Hopfield model through

$$Z = \sum_{\{\sigma'\}} \exp\left[\frac{\beta}{N}\sum_{i<j}\sigma_i'\sigma_j' + \frac{\beta}{2N}\right] , \tag{A1}$$

where the inverse temperature β is defined in (80). The partition function is thus independent of the pattern direction, which makes the calculation considerably simpler. The posterior entropy (79) can be written as

$$S[\{\boldsymbol{\sigma}^b\}] = \left(1 - \beta\,\frac{\partial}{\partial\beta}\right)\log\tilde{N}[\{\boldsymbol{\sigma}^b\},\beta] . \tag{A2}$$

where

$$\tilde{N}[\{\boldsymbol{\sigma}^b\},\beta] = \sum_{\{\boldsymbol{\xi}\}}\exp\left(\frac{\beta}{N}\sum_{b=0}^{B}\sum_{i<j}\xi_i\xi_j\sigma_i^b\sigma_j^b\right) , \tag{A3}$$

Thus, we are left with the calculation of $\tilde{N}[\{\boldsymbol{\sigma}^b\}]$. The expression for $\tilde{N}$ is formally identical to the partition function of a dual Hopfield model where the B measured configurations $\boldsymbol{\sigma}^b$ play the role of the dual patterns and $\boldsymbol{\xi}$ plays the role of the dual spin variables. The posterior entropy S is simply the entropy of this dual Hopfield model.

Equation (A2) gives the entropy of the system for a particular set of measures $\{\boldsymbol{\sigma}^b\}$. It is natural to expect the entropy to be reproducible across different sets of measurements. In this context, we are interested in evaluating the average of the entropy with respect to all possible measurements. Assuming that the configurations $\{\boldsymbol{\sigma}^b\}$ are sampled from the equilibrium measure of a Hopfield model with one pattern $\tilde{\boldsymbol{\xi}}$, we write the average entropy as

$$S = \left(1 - \beta\,\frac{\partial}{\partial\beta}\right)\,\langle\log\tilde{N}\rangle(\tilde{\beta},\beta)\Big|_{\tilde{\beta}=\beta} . \tag{A4}$$

where

$$\begin{aligned}\langle\log\tilde{N}\rangle(\tilde{\beta},\beta) \;&=\; \frac{1}{Z^B}\sum_{\{\boldsymbol{\sigma}^b\}}\exp\left(\frac{\tilde{\beta}}{N}\sum_{b=0}^{B}\sum_{i<j}\tilde{\xi}_i\tilde{\xi}_j\sigma_i^b\sigma_j^b\right)\\ &\times\;\log\tilde{N}[\{\boldsymbol{\sigma}^b\},\beta] ,\end{aligned} \tag{A5}$$

where we have introduced a new variable $\tilde{\beta}$ since we should not take the derivative only with respect to β in (A4).

To calculate the average value of the logarithm of $\tilde{N}$ in (A5) we use the replica trick [27] and estimate the n^{th}

moment of $\tilde{N}$,

$$\langle \tilde{N}^n \rangle = e^{-\beta B n/2} \sum_{\{\boldsymbol{\xi}^\rho\},\tilde{\boldsymbol{\xi}},\{\boldsymbol{\sigma}^b\}} \int \prod_{b=1}^{B} \prod_{\rho=1}^{n} \frac{dm_b^\rho}{\sqrt{2\pi}} \times \exp\left[-\frac{\beta N}{2} \sum_{b,\rho} (m_b^\rho)^2 + \beta \sum_{b,\rho,i} m_b^\rho \xi_i^\rho \sigma_i^b + \frac{\tilde{\beta}}{N} \sum_b \sum_{i<j} \sigma_i^b \sigma_j^b \tilde{\xi}_i \tilde{\xi}_j \right] . \quad \text{(A6)}$$

We introduce auxiliary Gaussian variables, denoted by $\tilde{m}_b$, to linearize the quadratic term in the spins σ_i^b. We obtain, after summation over the spins,

$$\langle \tilde{N}^n \rangle = e^{-\beta B n/2} \sum_{\{\boldsymbol{\xi}^\rho\},\tilde{\boldsymbol{\xi}}} \int \prod_{b,\rho} \frac{dm_b^\rho}{\sqrt{2\pi}} \prod_b \frac{d\tilde{m}_b}{\sqrt{2\pi}} \quad \text{(A7)}$$
$$\times \exp\left[-\frac{\beta N}{2} \sum_{b,\rho} (m_b^\rho)^2 - \frac{\beta N}{2} \sum_b (\tilde{m}_b)^2 + \sum_{i,b} \ln 2\cosh\left(\beta \sum_\rho m_b^\rho \xi_i^\rho + \tilde{\beta}\, \tilde{m}_b\, \tilde{\xi}_i \right) \right] .$$

In the paramagnetic phase we expect the variables m_b^ρ and $\tilde{m}_b$ to be of the order of $\frac{1}{\sqrt{N}}$. Expanding the hyperbolic cosine to the second order in those variables and carrying out the resulting Gaussian integral we obtain

$$\langle \tilde{N}^n \rangle \simeq e^{-\beta B n/2} \sum_{\{\boldsymbol{\xi}^\rho\},\tilde{\boldsymbol{\xi}}} [\det M]^{-B/2} . \quad \text{(A8)}$$

Here, M is the $(n+1)\times(n+1)$ matrix with elements

$$M_{\rho\sigma} = \begin{cases} 1-\beta & \text{if} \quad \rho = \sigma \le p , \\ 1-\tilde{\beta} & \text{if} \quad \rho = \sigma = p+1 , \\ -\sqrt{\beta\tilde{\beta}}\, t_\sigma & \text{if} \quad \rho = p+1,\ \sigma \le p , \\ -\sqrt{\beta\tilde{\beta}}\, t_\rho & \text{if} \quad \rho \le p,\ \sigma = p+1 , \\ -\beta\, r_{\rho\sigma} & \text{if} \quad \rho \le p,\ \sigma \le p . \end{cases} \quad \text{(A9)}$$

with the overlaps defined through $r_{\rho\sigma} = \frac{1}{N}\sum_i \xi_i^\rho \xi_i^\sigma$ and $t_\rho = \frac{1}{N}\sum_i \xi_i^\rho \tilde{\xi}_i$. We now enforce the definitions of the overlaps using conjugated Lagrange multipliers, $\hat{r}_{\rho\sigma}$ and $\hat{t}_\rho$, and obtain

$$\langle \tilde{N}^n \rangle = \int \prod_{\rho<\sigma} \frac{dr_{\rho\sigma}\, d\hat{r}_{\rho\sigma}}{2\pi} \prod_\rho \frac{dt_\rho\, d\hat{t}_\rho}{2\pi}\ \Xi^N , \quad \text{(A10)}$$

where Ξ is given by

$$\Xi = \sum_{\{\xi^\rho,\tilde{\xi}\}} \exp\left[-\frac{\alpha}{2} \log\det M - \sum_{\rho<\sigma} \hat{r}_{\rho\sigma} r_{\rho\sigma} - \frac{\alpha\beta n}{2} - \sum_\rho \hat{t}_\rho t_\rho + \sum_{\rho<\sigma} \hat{r}_{\rho\sigma} \xi^\rho \xi^\sigma + \sum_\rho \hat{t}_\rho\, \tilde{\xi} \xi^\rho \right] . \quad \text{(A11)}$$

We look for a replica-symmetric saddle point of Ξ: $r_{\rho\sigma} = r$, $t_\rho = t$, $\hat{r}_{\rho\sigma} = \hat{r}$ and $\hat{t}_\rho = \hat{t}$. We obtain, after some elementary algebra,

$$\Xi = \int_{-\infty}^{\infty} Dz \exp\left\{ -\frac{\alpha}{2} \log\det M - \frac{n(n-1)}{2}\, \hat{r}\, r - n\, \hat{t}\, t + n \log\left[2\cosh\left(\hat{t} + z\sqrt{\hat{r}} \right) \right] - \frac{\alpha\beta n}{2} \right\} . \quad \text{(A12)}$$

where $Dz = dz\, e^{-z^2/2}/\sqrt{2\pi}$ is the Gaussian measure and

$$\det M = (1-\beta+\beta r)^{n-1} \left[(1-\tilde{\beta})(1-\beta) - (n-1)(1-\tilde{\beta})\beta r - n\beta\tilde{\beta}\, t^2 \right] . \quad \text{(A13)}$$

We now send n to zero. The saddle-point equations show that $t = r$; this result was expected from the fact that, if $\tilde{\beta} = \beta$, the true pattern $\tilde{\xi}$ plays the role of an extra replicated pattern ξ. In addition, $\hat{t} = \hat{r} \equiv \gamma$, where γ is defined in (85). The self-consistent equations for r and the entropy S are given by, respectively eqns (84) and (86).

[1] I.T. Jolliffe, *Principal Component Analysis*, Springer Verlag (2002).
[2] A.K. Seth, G.M. Edelman, *Neural. Comput.* **19**, 910 (2007).
[3] E.T. Jaynes, *Proc. IEEE* **70**, 939 (1982).
[4] T. Hastie, R. Tibshirani, J. Friedman, *Elements of Statistical Learning: Data Mining, Inference and Prediction (Second Edition)*, Springer-Verlag, New York (2009).
[5] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, *Cognitive Science* **9**, 147 (1985).
[6] M. Opper, D. Saad (eds), *Advanced Mean Field Methods: Theory and Practice*, MIT Press (2001).
[7] Y. Roudi, J.Tyrcha, J. Hertz, *Phys. Rev. E* **79**, 051915 (2009).
[8] M. Mézard, T. Mora, *J. Physiol. Paris* **103**, 107 (2009); E. Marinari, V. Van Kerrebroeck, *J. Stat. Mech.* P02008 (2010).
[9] H.P. Huang, *Phys. Rev. E* **82**, 056111 (2010).
[10] S. Cocco, S. Leibler, R. Monasson, *Proc. Nat. Acad. Sci.* **106**, 14058 (2009).
[11] S. Cocco, R. Monasson, *Phys. Rev. Lett.* **106**, 090601 (2011).
[12] P. Ravikumar, M.J. Wainwright, J. Lafferty, *Annals of Statistics* **38**, 1287 (2010).
[13] Z. Bai, J.W. Silverstein, *Spectral analysis of large dimensional random matrices*, Springer (2009).
[14] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, G.R.G. Lanckriet, *SIAM Review* **49**, 434 (2007).
[15] J.J. Hopfield, *Proc. Nat. Acad. Sci. (USA)* **79**, 2554 (1982).

[16] D.J. Amit, *Modelling Brain Function: the World of Attractor Neural Networks*, Cambridge University Press (1992).

[17] K. Nokura, *J. Phys. A* **31**, 7447 (1998).

[18] A. Engel, C. van den Broeck, *Statistical Mechanics of Learning*, Cambridge University Press (2001).

[19] I.M. Johnstone, *Proc. ICM 2006* **1**, 307 (2006).

[20] D.J.C. MacKay, *Neural Computation* **4**, 415 (1991).

[21] L. Viana, A.J. Bray, *J. Phys. C* **18**, 3037 (1985).

[22] A. Peyrache *et al.*, *Nature Neurosci.* **12**, 919 (2009); A. Peyrache *et al.*, *J. Comput. Neurosci.* **29**, 309 (2009).

[23] S.W. Lockless, R. Ranganathan, *Science* **286**, 295 (1999).

[24] see http://www.hhmi.swmed.edu/Labs/rr/sca.html for a brief description of the SCA approach on PDZ and the definition of the weights D_i.

[25] N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, *Cell* **138**, 774 (2009).

[26] G. Schwarz, *Ann. Stat.* **6**, 461 (1978).

[27] D.J. Amit, H. Gutfreund, H. Sompolinsky, *Phys. Rev. A* **32**, 1007 (1985).

[28] M. Biehl, A. Mietzner, *J. Phys. A* **27**, 1885 (1994).

[29] P. Reimann, C. Van den Broek, G.J. Bex, *J. Phys. A* **29**, 3521 (1996).

[30] T.L.H. Watkin, J.-P. Nadal, *J. Phys. A* **27**, 1899 (1994).

[31] H.J. Sommers, W. Dupont, *J. Phys. C* **17**, 5785 (1984); A Crisanti, T. Rizzo, *Phys. Rev. E* **65**, 046137 (2002).

[32] D.S. Dean, F. Ritort, *Phys. Rev. B* **65**, 224209 (2002).

[33] D.C. Hoyle, M. Rattray, *Europhys. Lett.* **62**, 117 (2003); *Phys. Rev. E* **69**, 026124 (2004); *Phys. Rev. E* **75**, 016101 (2007).

[34] D.C. Hoyle, *J. Stat. Mech.*, P04009 (2010).

[35] J. Baik, G. Ben Arous, S. Péché, *Ann. Probab.* **33**, 1643 (2005).

[36] J. Sinova, G. Canright, A.H. MacDonald, *Phys. Rev. Lett.* **85**, 2609 (2000); J. Sinova, G. Canright, H.E. Castillo, A.H. MacDonald, *Phys. Rev. B* **63**, 104 427 (2001).

[37] M. Weigt *et al.*, *Proc. Nat. Acad. Sci.* **106**, 67 (2009).

[38] M. Weigt, *private communication* (2010).

[39] R. Tibshirani, *J. Royal. Statist. Soc B* **58**, No. 1, p 267 (1996).

[40] As a result of the block structure the energy (3) depends on the N–spin configuration through the four block magnetizations (sums of the $\frac{N}{4}$ spins in each block) only. Hence, the correlations c_{ij} and magnetizations m_i can be calculated in a time growing as N^4 (instead of 2^N), which allows us to reach sizes equal to a few hundreds easily.

[41] The corresponding magnetizations were $\simeq (-.26, .13, .13, .23)$ for $N = 52$ spins.

[42] Formula (90) can be found by inverting identity (12), with $J_{ij} = \frac{1}{N}\xi_i\xi_j$.