Ising distribution as a latent variable model

Adrien Wohrer

Ising distribution as a latent variable model

Adrien Wohrer*

Université Clermont Auvergne, CNRS, SIGMA Clermont,
Institut Pascal, F-63000 Clermont-Ferrand, France.

During the past decades, the Ising distribution has attracted interest in many applied disciplines, as the maximum entropy distribution associated to any set of correlated binary ('spin') variables with observed means and covariances. However, numerically speaking, the Ising distribution is unpractical, so alternative models are often preferred to handle correlated binary data. One popular alternative, especially in life sciences, is the *Cox distribution* (or the closely related *dichotomized Gaussian distribution* and *log-normal Cox point process*), where the spins are generated independently conditioned on the drawing of a latent variable with a multivariate normal distribution. This article explores the conditions for a principled replacement of the Ising distribution by a Cox distribution. It shows that the Ising distribution itself can be treated as a latent variable model, and it explores when this latent variable has a quasi-normal distribution. A variational approach to this question reveals a formal link with classic mean field methods, especially Opper and Winther's *adaptive TAP* approximation. This link is confirmed by weak coupling (Plefka) expansions of the different approximations, and then by numerical tests. Overall, this study suggests that an Ising distribution can be replaced by a Cox distribution in practical applications, precisely when its parameters lie in the 'mean field domain'.

I. INTRODUCTION

During the last decades, the Ising distribution has been used in several disciplines such as statistics (under the name *quadratic exponential model*) [1, 2], machine learning (under the name *Boltzmann machine*) [3], information processing [4–6], biology [7] and neurosciences, where it has been proposed as a natural model for the spike-based activities of interconnected neural populations [8, 9]. In most of these applications, classic assumptions from statistical physics do not hold (e.g., arrangement on a rectangular lattice, uniform couplings, independently distributed couplings, zero external fields, etc.), and even old problems have to be revisited, such as efficiently simulating the Ising distribution [10, 11] or inferring its parameters from data [12–15].

In this article I will consider the Ising probability distribution over a set of spins $\mathbf{s} = (s_1, \ldots, s_N) \in \{-1, 1\}^N$ defined as

$$P(\mathbf{s}|\mathbf{h}, \mathbf{J}) = \frac{1}{Z_I} \exp\Big(\sum_{i=1}^{N} h_i s_i + \frac{1}{2} \sum_{i,j=1}^{N} J_{ij} s_i s_j \Big), \quad (1)$$

with parameters $\mathbf{h} \in \mathbb{R}^N$ (external fields) and $\mathbf{J}$ an $N \times N$ symmetric matrix (coupling weights), $Z_I(\mathbf{h}, \mathbf{J})$ being the corresponding partition function. In this formulation, diagonal elements J_{ii} can be nonzero without influencing the distribution, simply adding a constant term $\sum_i J_{ii}/2$ to both the exponent and $\log(Z_I)$.

I will note $(\mathbf{m}, \mathbf{C})$ the two first centered moments of the distribution, that is, for all indices i, j,

$$m_i = \mathrm{E}(s_i) \quad , \quad C_{ij} = \mathrm{E}(s_i s_j) - \mathrm{E}(s_i)\mathrm{E}(s_j).$$

The essential interest of the Ising distribution, in all disciplines mentioned above, is its *maximum entropy* property : whenever a dataset of N binary variables has measured moments $(\mathbf{m}, \mathbf{C})$, a single distribution of the form (1) is guaranteed to exist which matches these moments, and furthermore it has maximal entropy under this constraint.

Unfortunately, the Ising distribution is numerically unwieldy. The simple act of drawing samples from the distribution already requires to set up lengthy Markov Chain Monte Carlo (MCMC) schemes. Besides, there is no simple analytical link between parameters $(\mathbf{h}, \mathbf{J})$ and resulting moments $(\mathbf{m}, \mathbf{C})$. Given natural parameters $(\mathbf{h}, \mathbf{J})$, the *direct Ising problem* of estimating $(\mathbf{m}, \mathbf{C})$ can only be solved by numerical sampling from MCMC chains. Given $(\mathbf{m}, \mathbf{C})$, the *inverse Ising problem* of retrieving $(\mathbf{h}, \mathbf{J})$ can only be solved by gradient descent based on numerous iterations of the direct problem, a procedure known as *Boltzmann learning*. In practice, this means that the Ising distribution cannot be parametrized easily from observed data.

For this reason, in spite of the Ising model's theoretical attractiveness when dealing with binary variables, alternative models are generally preferred, which are numerically more convenient. In one such family of alternative models, N *latent variables* $\mathbf{r} = (r_1, \ldots, r_N) \in \mathbb{R}^N$ are drawn from a multivariate normal distribution

$$\mathcal{N}(\mathbf{r}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\Big(-\frac{1}{2}(\mathbf{r}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{r}-\boldsymbol{\mu}) \Big)$$

and then used to generate N spins independently. The simplest option, setting $s_i = \mathrm{sign}(r_i)$ deterministically, yields the *dichotomized Gaussian* distribution [16, 17], which has enjoyed recent popularity as a replacement for the Ising distribution when modeling neural spike trains, as it is easy to sample and to parametrize from data [18, 19].

Slightly more generally, each variable r_i can serve as an intensity to draw the corresponding spin s_i following

a Bernoulli distribution :

$$\mathcal{B}(\mathbf{s}|\mathbf{r}) = \frac{1}{Z_B} \exp\Big(\sum_{i=1}^{N} r_i s_i\Big)$$

with partition function

$$Z_B(\mathbf{r}) = \prod_i \big(e^{r_i} + e^{-r_i}\big).$$

In statistics, this is the model underlying logistic regression, as introduced by Cox [20], so I will refer to it as the *Cox distribution* :

$$Q(\mathbf{r}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{r}|\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{2}$$
$$Q(\mathbf{s}|\mathbf{r}) = \mathcal{B}(\mathbf{s}|\mathbf{r}), \tag{3}$$

with $\boldsymbol{\mu}$ any vector in $\mathbb{R}^N$ and $\boldsymbol{\Sigma}$ any $N \times N$ symmetric definite positive matrix. Note that the dichotomized Gaussian corresponds to a limiting case of the Cox distribution, when the scaling of variables r_i tends to $+\infty$ [1].

Both the Ising (eq. (1)) and Cox (eq. (2)-(3)) distributions can be generalized to point processes, by taking a suitable limit when the N indexed variables tend to a continuum [21]. These are respectively known as the *Gibbs process* [21] and *log Gaussian Cox process* [22, 23]. The latter, much simpler to handle in practice, is used in various applied fields such as epidemiology, geostatistics [24] and neurosciences, to model neural spike trains [25, 26].

To summarize, in practical application, Cox models (including the dichotomized Gaussian, and Cox point processes) are often preferred to the corresponding maximum entropy distributions (Ising distribution, Gibbs point process) because they are easier to sample, and to parametrize from a set of observed data. However, to date, we have little analytical insights into the link between the two families of distributions. For example, given some dataset, we cannot tell in advance how similar the Cox and Ising models fitting this data would be. The goal of this article is to investigate this link.

I first show that the Ising distribution itself can be viewed as a latent variable model, which differs from a Cox distribution only because of the non-Gaussian distribution of its latent variable (Section II). This allows to derive simple relations between an Ising distribution and its 'best-fitting' Cox distributions, in two possible senses (Section III). In particular, the variational approach for targeting a best-fitting Cox distribution displays formal

similarities with classic mean-field methods which aim at approximating the Ising moments $(\mathbf{m}, \mathbf{C})$ (Section IV). Numerical simulations reveal that both types of approximations, despite their seemingly different goals, are efficient in roughly the same domain of parameters (Section V). Thus, an Ising distribution can be replaced in practical applications by a Cox distribution, precisely if its parameters lie in the 'mean field domain'.

II. THE ISING LATENT FIELD

Given any vector $\mathbf{h} \in \mathbb{R}^N$ and $N \times N$ symmetric, *definite positive* matrix $\mathbf{J}$, we will consider the following probability distribution over $\mathbf{r} \in \mathbb{R}^N$ and $\mathbf{s} \in \{-1,1\}^N$:

$$P(\mathbf{s},\mathbf{r}) = \frac{1}{Z} \exp\Big(-\frac{1}{2}(\mathbf{r}-\mathbf{h})^\top \mathbf{J}^{-1}(\mathbf{r}-\mathbf{h}) + \mathbf{r}^\top \mathbf{s}\Big), \tag{4}$$

with Z ensuring proper normalization.

Marginalizing out variable $\mathbf{s}$ yields

$$P(\mathbf{r}) = \frac{Z_B(\mathbf{r})}{Z} \exp\Big(-\frac{1}{2}(\mathbf{r}-\mathbf{h})^\top \mathbf{J}^{-1}(\mathbf{r}-\mathbf{h})\Big), \tag{5}$$
$$P(\mathbf{s}|\mathbf{r}) = \mathcal{B}(\mathbf{s}|\mathbf{r}). \tag{6}$$

Conversely, completing the square in eq. (4) and marginalizing out variable $\mathbf{r}$ yields

$$P(\mathbf{s}) = \frac{|2\pi\mathbf{J}|^{1/2}}{Z} \exp\Big(\mathbf{h}^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{J}\mathbf{s}\Big), \tag{7}$$
$$P(\mathbf{r}|\mathbf{s}) = \mathcal{N}(\mathbf{r}|\mathbf{h} + \mathbf{J}\mathbf{s}, \mathbf{J}). \tag{8}$$

From eq. (7), the resulting spins $\mathbf{s}$ are distributed according to the Ising distribution of parameters $(\mathbf{h}, \mathbf{J})$.

The introduction of field variables r_i has long been known in statistical physics, as a mathematical construct to express the Ising partition function Z_I in an integral form [5]. Indeed, equating the respective expressions for Z imposed by eq. (5) and (7), we obtain the elegant formula

$$Z_I(\mathbf{h},\mathbf{J}) = \int_{\mathbf{r}\in\mathbb{R}^N} Z_B(\mathbf{r})\mathcal{N}(\mathbf{r}|\mathbf{h},\mathbf{J})\mathrm{d}\mathbf{r} \tag{9}$$

expressing Z_I as the convolution of Z_B with a Gaussian kernel of covariance $\mathbf{J}$. This formula can be used as a justification of classic mean field equations [5], and more generally to derive the diagrammatic (i.e., Taylor) expansion of Z_I as a function of $\mathbf{J}$ [27].

In this work instead, I view the r_i as a set of probabilistic variables in their own right, coupled to the Ising spin variables, through eq. (5)-(8). Given a spin configuration $\mathbf{s}$, variable $\mathbf{r}$ is normally distributed (eq. (8)). Thus, the overall distribution $P(\mathbf{r})$ is a mixture of Gaussians with 2^N components, where the component associated to spin configuration $\mathbf{s}$ has weight $P(\mathbf{s})$. More compactly, $P(\mathbf{r})$ can be expressed with eq. (5).

Given some configuration $\mathbf{r}$, the spins $\mathbf{s}$ can simply be drawn independently following a Bernoulli distribution (eq. (6)), so the Ising distribution $P(\mathbf{s})$ itself can be

---

[1] The dichotomized Gaussian of parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the limit of the Cox distribution of parameters $(\lambda\boldsymbol{\mu}, \lambda^2\boldsymbol{\Sigma})$ when $\lambda \to +\infty$. Conversely, note that a Bernoulli variable $S \sim \mathcal{B}(r)$ can be generated as $S = \mathrm{sign}(r + X)$ where X follows the *logistic distribution* of density function $f(x) = \frac{1}{2}\cosh^{-2}(x)$, which resembles closely the normal distribution $\mathcal{N}(x|0,\kappa^2)$ with $\kappa \simeq 0.85$. As a result, any Cox distribution of parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is decently approximated by a dichotomized Gaussian of parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \kappa^2\mathbf{I})$.

viewed as a latent variable model, based on hidden variables r_i. It departs from a Cox distribution (eq. (2)-(3)) only through the fact that the fields' distribution $P(\mathbf{r})$ is not normal, in general.

III. COX APPROXIMATIONS TO THE ISING DISTRIBUTION

This article investigates the possible replacement of the Ising distribution by a Cox distribution. With the above reformulation, this amounts to approximating the Ising latent field distribution $P(\mathbf{r})$ by a well-chosen multivariate normal $Q(\mathbf{r}) = \mathcal{N}(\mathbf{r}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. I will now discuss two possible choices in this regard.

From here on, I will note $(\mathbf{m}, \mathbf{C})$ for (any approximation of) the first moments of a spin variable $\mathbf{s}$, and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for (any approximation of) the first moments of a field variable $\mathbf{r}$. I will distinguish the true moments of the Ising distribution $P(\mathbf{s})$ with a star : $(\mathbf{m}^\star, \mathbf{C}^\star)$. Likewise, the true moments of the corresponding Ising latent field $P(\mathbf{r})$ follow, from eq. (8) :

$$\boldsymbol{\mu}^\star = \mathbf{h} + \mathbf{Jm}^\star, \qquad (10)$$
$$\boldsymbol{\Sigma}^\star = \mathbf{J} + \mathbf{JC}^\star\mathbf{J}. \qquad (11)$$

Optimal Cox distribution

Arguably, the optimal approximation of $P(\mathbf{r})$ by a normal distribution $Q(\mathbf{r}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is achieved by equating their moments, i.e., setting $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$. I will refer to this choice as the *optimal* Cox distribution. Its qualification as 'optimal' stems from the observation that the following KL divergence

$$\mathrm{KL}_{\mathbf{r}}(P||Q) = \int_{\mathbf{r}\in\mathbb{R}^N} P(\mathbf{r}) \ln \frac{P(\mathbf{r})}{Q(\mathbf{r}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} \mathrm{d}\mathbf{r} \qquad (12)$$

is minimized when $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$. Thus, in terms of information geometry, the resulting distribution $Q(\mathbf{r}|\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$ is the nearest neighbor of the latent Ising distribution $P(\mathbf{r})$ in the family of normal distributions.

Unfortunately, the optimal Cox distribution is unpractical to characterize, as $(\mathbf{m}^\star, \mathbf{C}^\star)$ can only be estimated by lengthy Monte-Carlo simulation. In the scope of this article, its study will only be of theoretical interest : it allows to quantify the intrinsic effect of assuming a Gaussian shape for $P(\mathbf{r})$.

Variational Cox approximation

More practically, one may require to approximate an Ising distribution $P(\mathbf{s})$ by a Cox distribution, assuming only knowledge of its natural parameters $(\mathbf{h}, \mathbf{J})$. The *variational* approach to this problem consists in choosing $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ that minimize the reversed Kullback-Leibler divergence

$$\mathrm{KL}_{\mathbf{r}}(Q||P) = \int_{\mathbf{r}\in\mathbb{R}^N} Q(\mathbf{r}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \frac{Q(\mathbf{r}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{P(\mathbf{r})} \mathrm{d}\mathbf{r}. \qquad (13)$$

With some straightforward algebra, one can establish the derivatives of $\mathrm{KL}_{\mathbf{r}}(Q||P)$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and thus its stationary points $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Given the fundamental relation (10)-(11) between spin and field moments in the Ising distribution, it is natural to reparametrize the Cox parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by the 'spin moment' parameters $(\mathbf{m}, \mathbf{C})$ such that

$$\boldsymbol{\mu} = \mathbf{h} + \mathbf{Jm}, \qquad (14)$$
$$\boldsymbol{\Sigma} = \mathbf{J} + \mathbf{JCJ}. \qquad (15)$$

Then, the values of $(\mathbf{m}, \mathbf{C})$ at the stationary points of $\mathrm{KL}_{\mathbf{r}}(Q||P)$ are characterized by the following, fixed point equation :

$$m_i = \int_{x\in\mathbb{R}} \tanh\left(\mu_i + x\sqrt{\Sigma_{ii}}\right) \phi(x)\mathrm{d}x, \qquad (16)$$

$$d_i = \int_{x\in\mathbb{R}} \left(1 - \tanh^2\left(\mu_i + x\sqrt{\Sigma_{ii}}\right)\right) \phi(x)\mathrm{d}x, \qquad (17)$$

$$(\mathbf{C}^{-1})_{ij} = d_i^{-1}\delta_{ij} - J_{ij}, \qquad (18)$$

with $\phi(x) = \mathcal{N}(x|0, 1)$ the standard one-dimensional normal distribution.

Equations (14)-(18) can be solved by an iterative fixed point method on variables $\{m_i, d_i\}_{i=1\ldots N}$ (see Appendix A). At the solution, the Cox distribution $Q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ provides an approximation to the Ising distribution $P(\mathbf{h}, \mathbf{J})$.

The formulas (16)-(17) are conceptually simple : m_i (resp. d_i) is obtained as the average of $\tanh(r)$ (resp. $1 - \tanh^2(r)$) using a Gaussian kernel, centered around $r = \mu_i$ with variance Σ_{ii}. Their estimation at any required precision is straightforward, using numerical integration. However, when repeated computations are required, it is faster to use approximate formulas, given in Appendix E.

Figure 1 illustrates the nature of the latent field $P(\mathbf{r})$, and of its Cox approximations, on a 2-spin toy model. The optimal Cox distribution is unique by construction, but the variational Cox approximation can have multiple solutions (panel b), a classic feature of variational methods based on minimizing reversed KL divergence [28].

Choice of diag($\mathbf{J}$)

The framework developed above requires strictly positive diagonal coupling values J_{ii}, large enough to ensure that matrix $\mathbf{J}$ is definite positive. While these diagonal values do not influence the Ising distribution $P(\mathbf{s})$ over spins, a different choice of diag($\mathbf{J}$) leads to a different latent field distribution $P(\mathbf{r})$, and thus to a different Cox approximation (see Figure 1). This naturally raises the question of what self-couplings J_{ii} represent in the latent field formalism, and how they should be chosen in
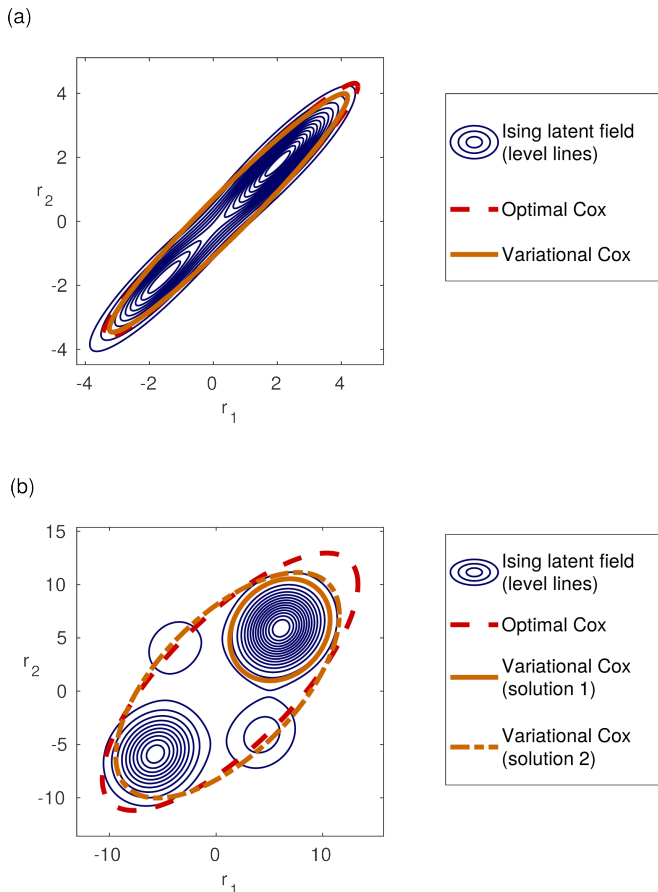
FIG. 1. Ising latent field and its normal approximations on a toy model with $N = 2$. Normal distributions are materialized by their 2-σ level line. Classic Ising parameters $(h_1, h_2, J_{12}) = (0.2, 0, 0.9)$. In (a), diagonal couplings are fixed at $J_{11} = J_{22} = 1$, whereas in (b) they are fixed at $J_{11} = J_{22} = 5$. The Ising spin distribution $P(\mathbf{s})$ is the same in both cases, but not the field distribution $P(\mathbf{r})$ and subsequent Cox approximations. In (b), the variational Cox equation has multiple solutions – two of which are displayed.

practice. Here, I only detail one concrete proposal for this choice, and defer more general considerations to the Discussion.

In the perspective of this work, the choice should be made to optimize the resemblance of $P(\mathbf{r})$ with a normal distribution. From eq. (8), larger values of $\mathbf{J}$ increase the overall separation of the 2^N components in $P(\mathbf{r})$ and thus, its divergence from a normal distribution. In the extreme case where $\mathrm{diag}(\mathbf{J})$ is very large, the 2^N components of $P(\mathbf{r})$ display no overlap at all : see Figure 1(b).

Consequently, the general prescription is that $\mathrm{diag}(\mathbf{J})$ should be kept as small as possible. Given a fixed set of off-diagonal weights $\{J_{ij}\}_{i<j}$, a principled approach is to choose $\mathrm{diag}(\mathbf{J})$ with the following procedure :

$$\underset{\{J_{ii}\}}{\text{minimize}} \quad \sum_i J_{ii} \quad \text{subject to} \quad \mathbf{J} \succeq 0. \qquad (19)$$

This is a well known convex problem, which can be solved

efficiently [29]. Afterwards, a small ridge term $\lambda\mathbf{I}$ can be added to $\mathbf{J}$ to make it *strictly* definite positive.

Moments of the Cox distribution

A note of caution is also required on the interpretation of variables $(\mathbf{m}, \mathbf{C})$ in the variational Cox approximation, eq. (14)-(18). By eq. (2)-(3), the first spin moment of Cox distribution $Q(\mathbf{s}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\mathrm{E}_Q(s_i) = \int_{r\in\mathbb{R}} \tanh(r)\mathcal{N}(r|\mu_i, \Sigma_{ii})\mathrm{d}r, \qquad (20)$$

and we recognize eq. (16). Thus, at the fixed point of eq. (14)-(18), m_i corresponds to the first moment of the approximating Cox distribution.

In contrast, at the fixed point of eq. (14)-(18), $\mathbf{C}$ is *not* the spin covariance of the Cox distribution $Q(\mathbf{s}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. That would be computed (for $i \neq j$) as

$$\mathrm{Cov}_Q(s_i, s_j) = \iint_{(r,t)\in\mathbb{R}^2} \tanh(r)\tanh(t)\mathcal{N}(r,t|\boldsymbol{\mu}^{(ij)}, \boldsymbol{\Sigma}^{(ij)})\mathrm{d}r\mathrm{d}t$$
$$- \mathrm{E}_Q(s_i)\mathrm{E}_Q(s_j), \qquad (21)$$

with $\boldsymbol{\mu}^{(ij)}, \boldsymbol{\Sigma}^{(ij)}$ the two-dimensional restrictions of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ at indices (i, j).

Thus, for given Cox parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, there are two possible predictions for the Ising covariance matrix $\mathbf{C}$: the 'forward' prediction of eq. (21) (spin covariance matrix in the Cox distribution) and the 'backward' prediction of eq. (15) (spin covariance in the Ising model which would give rise to field covariance matrix $\boldsymbol{\Sigma}$). If $Q(\mathbf{r}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a good approximation of $P(\mathbf{r})$, we expect both predictions to be very close. And indeed, the discrepancy between the two predictions of $\mathbf{C}$ is a good indicator of whether the approximation was successful (see Supplementary Material).

IV. COMPARISON WITH MEAN FIELD APPROXIMATIONS

In the variational Cox approximation, eq. (14)-(18), variables $(\mathbf{m}, \mathbf{C})$ constitute an approximation for the moments of the Ising distribution $P(\mathbf{h}, \mathbf{J})$. The goodness of fit wrt. exact Ising moments $(\mathbf{m}^\star, \mathbf{C}^\star)$ constitutes a simple measure of how well distribution P is approximated by the Cox distribution $Q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Deriving an approximation for $(\mathbf{m}, \mathbf{C})$ is also the goal of classic *mean field* methods. In these methods, the magnetizations $\mathbf{m}$ are approximated first, as the solution of some fixed point equation $\mathbf{m} = F(\mathbf{m}|\mathbf{h}, \mathbf{J})$. Then, this equation is differentiated wrt. $\mathbf{h}$, yielding a predicted covariance matrix as

$$C_{ij} = \partial_{h_i} m_j. \qquad (22)$$

Indeed, this so-called *linear response* formula holds true in the exact Ising model ; so it provides a concrete way of estimating $\mathbf{C}$ from the approximation of $\mathbf{m}$.

Informally, we may say that a given set of Ising parameters $(\mathbf{h}, \mathbf{J})$ lies in the 'mean field domain' when *some* mean field method can provide a good estimate of the corresponding moments $(\mathbf{m}^\star, \mathbf{C}^\star)$. In the rest of this article, I will argue that the variational approximation of P by a Cox distribution Q is valid precisely in this 'mean field domain'.

In this section, I briefly remind the nature of different mean field approximations, and compare their respective Plefka expansions in the case of weak couplings, up to order 3 (resp. 4 in the Appendices). In section V, I will proceed to numerical comparisons.

Classic mean field approximations

I considered two classic mean field methods, the *TAP* and *Bethe* approximations, and a more recent generalization called the *adaptive TAP* approximation. For the sake of self-completeness, these approximations are reminded in some detail in Appendices B and C. Here, I only provide essential formulas.

The archetypal mean field method in the Ising model is the *TAP approximation*, where the approximating vector of magnetizations $\mathbf{m}$ is sought as a fixed point of the following Thouless-Anderson-Palmer equation [4, 5, 30] :

$$m_i = \tanh \left(h_i + \sum_{j \neq i} J_{ij} m_j - m_i \sum_{j \neq i} J_{ij}^2 (1 - m_j^2) \right). \quad (23)$$

This equation can arise in different contexts, one of which is the Plefka expansion of the exact Ising model (eq. 29), stopped at order 2. Covariances $\mathbf{C}$ are derived in turn based on the linear response formula (eq. (22)) :

$$(\mathbf{C}_{\mathrm{TAP}}^{-1})_{ij} = \left(1 + \sum_k J_{ik}^2 (1 - m_i^2)(1 - m_k^2) \right) \frac{\delta_{ij}}{1 - m_i^2} - J_{ij} - 2J_{ij}^2 m_i m_j. \quad (24)$$

The *Bethe approximation* is a related mean field method, where the approximating vector of magnetizations $\mathbf{m}$ is sought as a fixed point of the following equation [31, 32] :

$$\theta_i^{(j)} = h_i + \sum_{k \neq j, i} \tanh^{-1} \left(\tanh(J_{ik}) \tanh(\theta_k^{(i)}) \right). \quad (25)$$

Here, the so-called *cavity fields* $\theta_i^{(j)}$ are tractable functions of $\mathbf{m}$, namely, the only numbers such that each 2-spin Ising distribution of natural parameters $(\theta_i^{(j)}, \theta_j^{(i)}, J_{ij})$ have first moments (m_i, m_j). This equation can be justified as the exact solution for $\mathbf{m}$ when the couplings J_{ij} define a tree-like lattice, and its iterative resolution is known as the *belief propagation* algorithm. Covariances $\mathbf{C}$ can be derived in turn based on the linear

response formula (Appendix B, eq. (B10), derived here with an original approach).

Finally, the *adaptive TAP approximation* of Opper and Winther [33] is a generalized mean field approximation, based on the cavity method [4]. The magnetization variables $\{m_i\}$ are joined with a second set of variables $\{V_i\}$, which represent the variance of the cavity field distribution at each spin site, and obey the following fixed point equations :

$$m_i = \tanh \left(h_i + \sum_j J_{ij} m_j - m_i V_i \right), \quad (26)$$

$$(\mathbf{C}_{\mathrm{A}}^{-1})_{ij} = \left(1 + (1 - m_i^2) V_i \right) \frac{\delta_{ij}}{1 - m_i^2} - J_{ij}, \quad (27)$$

$$1 - m_i^2 = (\mathbf{C}_{\mathrm{A}})_{ii}. \quad (28)$$

This derivation is detailed in Appendix C. Briefly, eq. (26) is a generalization of the TAP equation (23) where the variance V_i of the cavity field is left as a free variable, and eq. (27) is the corresponding linear response prediction. Equation (28) imposes coherent predictions for individual variances $\mathrm{Var}(s_i)$, thereby closing the fixed point equation on variables $\{m_i, V_i\}$.

The adaptive TAP approximation is a 'universal' mean field method : by letting the variances V_i adapt freely, it can account for any statistical structure of matrix $\mathbf{J}$, whereas the classic TAP equation (eq. (23)) is only true when the individual coupling weights J_{ij} are decorrelated [34]. This is especially welcome in machine learning and neurosciences, where coupling strengths J_{ij} are generally structured (because they represent learned regularities of the outside world).

Equations (26)-(28) bear a striking similarity with the variational Cox equations, eq. (14)-(18). Both can be seen as modifications of the naive mean field equations through N additional variables (the d_i, resp. V_i) associated to the *variance* of the field acting on each spin. This similarity will be confirmed in the subsequent analytical and numerical results.

Weak coupling expansions

In the Ising model, the link between natural parameters $(\mathbf{h}, \mathbf{J})$ and magnetizations $\mathbf{m}$ can be abstractly described as $\mathbf{h} = f(\mathbf{m}, \mathbf{J})$, for an intractable function f. Only when $\mathbf{J} = \mathbf{0}$ does the link become tractable : the Ising model boils down to a Bernoulli distribution, with the obvious $h_i = \tanh^{-1}(m_i)$.

One step further, when couplings are weak but nonzero, one can derive the Taylor expansion of f around $\mathbf{J} = \mathbf{0}$. In practice, the coupling matrix is written $\alpha \mathbf{J}$, α being the small parameter of the expansion, and the result is known as the *Plefka expansion* [35] – although the approach can be traced back to anterior work [27].

The expansion up to order 4 is a classic computation, outlined in Appendix B. Stopping at order 3 for brevity, it reads :

$$h_i = \tanh^{-1}(m_i) - \alpha \sum_{j \neq i} J_{ij} m_j + \alpha^2 m_i \sum_{j \neq i} J_{ij}^2 (1 - m_j^2)$$
$$+ \alpha^3 \left[2m_i \sum_{(jk|i)} J_{ij} J_{jk} J_{ki} (1 - m_j^2)(1 - m_k^2) + 2(m_i^2 - \tfrac{1}{3}) \sum_{j \neq i} J_{ij}^3 m_j (1 - m_j^2) \right] + o(\alpha^3), \qquad (29)$$

where $(jk|i)$ denotes all unordered triplets of the form $\{i, j, k\}$ with j and k distinct, and distinct from i.

This expansion can serve as a first test on the various approximations (TAP, Bethe, adaptive TAP, variational Cox) introduced above. Indeed, these approximations can also be described as $\mathbf{h} = f(\mathbf{m}, \mathbf{J})$ for a different function f, and we can compare its Taylor expansion to that of the true Ising model.

The expansion for the TAP approximation is, by defi-

nition, the exact Ising expansion (eq. (29)) stopped at order 2. The expansion for the Bethe approximation is obtained from the exact Ising expansion by retaining only the sums over spin pairs [36] (see Appendix B). In eq. (29), this means suppressing the sum over $(jk|i)$ in the order 3 term, but keeping the sum over $j \neq i$.

The expansion for the adaptive TAP approximation, derived in Appendix C, writes

$$h_i = \tanh^{-1}(m_i) - \alpha \sum_{j \neq i} J_{ij} m_j + \alpha^2 m_i \sum_{j \neq i} J_{ij}^2 (1 - m_j^2)$$
$$+ \alpha^3 \left[2m_i \sum_{(jk|i)} J_{ij} J_{jk} J_{ki} (1 - m_j^2)(1 - m_k^2) \right] + o(\alpha^3). \qquad (30)$$

The expansion for the variational Cox approximation, derived in Appendix D, writes

$$h_i = \tanh^{-1}(m_i) - \alpha \sum_{j \neq i} J_{ij} m_j + \alpha^2 m_i \sum_{j \neq i} J_{ij}^2 (1 - m_j^2)$$
$$+ \alpha^3 \left[2m_i \sum_{(jk|i)} J_{ij} J_{jk} J_{ki} (1 - m_j^2)(1 - m_k^2) - 2(m_i^2 - \tfrac{1}{3}) J_{ii}^3 m_i (1 - m_i^2) \right] + o(\alpha^3). \qquad (31)$$

At order 2, all approximations considered have the same expansion as the exact Ising solution, meaning that they will perform well in case of weak couplings.

At order 3, discrepancies appear between the exact Ising solution and its various approximations. The first contribution to the order 3 term in eq. (29), the sum over $(jk|i)$, is correctly accounted for by the adaptive TAP and variational Cox approximations. The second contribution to the order 3 term in eq. (29), the sum over $j \neq i$, is correctly accounted for by the Bethe approximation. Of these two sums, that over $(jk|i)$ involves many more terms, so we expect it will generally be the dominant contribution, except for very specific coupling matrices $\mathbf{J}$.

At order 4, the same qualitative features are observed. The 'generally dominant' contribution at order 4 in the true Ising solution writes

$$2m_i \sum_{(jkl|i)} J_{ij} J_{jk} J_{kl} J_{li} (1 - m_j^2)(1 - m_k^2)(1 - m_l^2)$$

(Appendices, eq. (B3)), and this is also the dominant

contribution to the adaptive TAP (eq. (C7)) and variational Cox (eq. (D12)) approximations. Hence, we may expect these two approximations to provide a better fit than the others in case of generic coupling matrices $\mathbf{J}$ – and this will indeed be our observation in numerical tests (Figure 6).

The similar structures of eq. (18) and (27) suggest a proximity between the adaptive TAP and variational Cox approximations, and this is confirmed by their weak coupling expansions : up to order 4, their respective expansions differ only through additional terms involving the diagonal weights J_{ii}, so they would be identical for a classic coupling matrix such that $\text{diag}(\mathbf{J}) = \mathbf{0}$.

V. NUMERICAL TESTS

To gain more insights on the behavior of all approximations above, I turned to numerical exploration : I picked a large number of possible configurations $(\mathbf{h}, \mathbf{J})$, estimated the true Ising moments $(\mathbf{m}^\star, \mathbf{C}^\star)$ in each configuration

with lengthy MCMC sampling, and compared all approximations against this ground truth. The numerical details for computing the approximations are given in Appendix A.

I should stress from the start that the role of these tests is *not* to target the most accurate mean field method for approximating $(\mathbf{m}, \mathbf{C})$ – which turns out to be the adaptive TAP method, in most configurations tested here. Instead, the goal of these tests is to support one main claim of this article : that the Ising distribution is well approximated by a Cox distribution in the same domain of parameters where mean field methods are efficient.

To focus on the most important aspect of the Ising model, I only tested configurations where $\mathbf{h} = \mathbf{0}$, so that magnetizations verify $\mathbf{m} = \mathbf{0}$ – both in the exact Ising solution and in the various approximations. Hence, the efficiency of a given approximation is assessed by its ability to correctly predict the covariance matrix $\mathbf{C}$. For each tested configuration and approximation, the fit performance is summarized by number

$$z = \frac{1}{N^2} \sum_{i,j} |C_{ij} - C_{ij}^\star|,$$

where $C_{ij}^\star$ is the true covariance of spins i and j, and C_{ij} its approximation.

Generative model for couplings $\mathbf{J}$

In this approach, the choice of a generative model for coupling matrix $\mathbf{J}$ is a delicate matter, as the Ising model can exhibit very different behaviors depending on its parameters. To test different regimes with a single formula, I used the following generative model :

$$\mathbf{J} = \left[\frac{J_0}{pN}\mathbf{1} + \frac{J}{\sqrt{\kappa pN}}\mathbf{X}_\kappa \mathbf{X}_\kappa^\top \right] \cdot^* \mathbf{M}_p \quad (32)$$

with $\mathbf{1}$ the $N \times N$ matrix with uniform unit entries and $\mathbf{X}_\kappa$ an $N \times \kappa N$ matrix of independent standard normal entries. Afterwards, element-wise multiplication by a symmetric masking matrix $\mathbf{M}_p$ can randomly set each edge (ij) at 0 with probability $1 - p$.

This model has 4 parameters. The positive numbers (J_0, J) correspond to the standard Sherrington-Kirkpatrick (SK) parameters for spin glasses [37], meaning that the probabilistic distribution of each nonzero off-diagonal term J_{ij} follows

$$J_{ij} \sim \mathcal{N}\left(\frac{J_0}{pN}, \frac{J^2}{pN} \right) \quad (33)$$

to a very good approximation, owing to the central limit theorem applied to the κN samples in matrix $\mathbf{X}_\kappa$.

Parameter $\kappa > 0$ fixes the amount of global correlation between individual couplings J_{ij}. When $\kappa \to +\infty$, all off-diagonal entries J_{ij} constitute independent random variables, and the classic SK model is recovered (see Supplementary Material). When κ is smaller, the
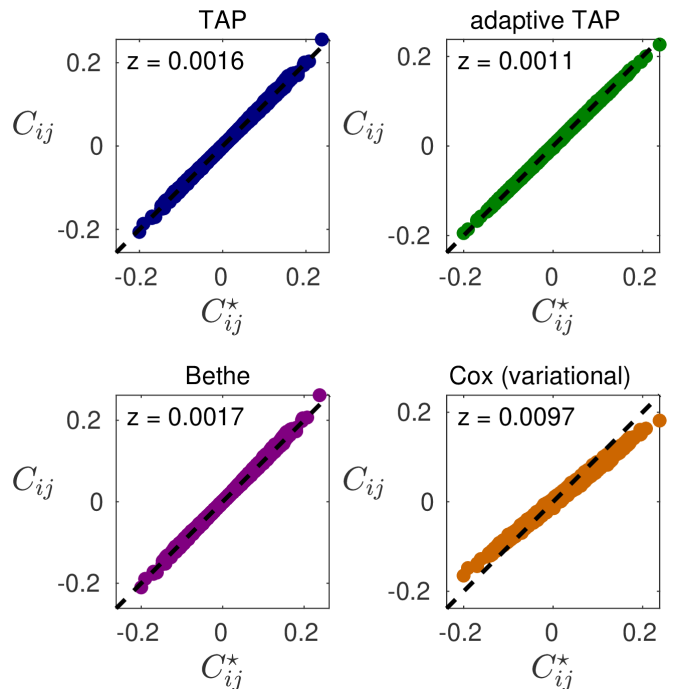

FIG. 2. TAP, Bethe, adaptive TAP and variational Cox approximations on a typical paramagnetic configuration with $N = 100$ spins, and reference parameter values $(J_0, J, \kappa, p) = (0.5, 0.5, \infty, 1)$. The corresponding fit values z are indicated as insets. In each of Figures 3-6, one parameter will be modified, while the three others keep their reference value.

random matrix $\mathbf{X}_\kappa \mathbf{X}_\kappa^\top$ follows a Wishart distribution, as in the Hopfield model of associative memory [4, 38]. The random variables J_{ij} become dependent, and the spectrum of $\mathbf{J}$ differs markedly from the SK case (Wigner vs. Marčenko-Pastur laws). Note that the probabilistic distribution of each element J_{ij} remains virtually unchanged in the process, given by eq. (33) except at very low values of κ.

Finally, parameter $p \in [0, 1]$ allows to dilute the overall connectivity, so that only a proportion p of the couplings are nonzero. Coherently, J_0 and J^2 in eq. (33) are scaled by pN, the effective number of neighbors in the (possibly diluted) model.

Results for the approximations

As the 4 parameters in model (32) prevent from an exhaustive search, I performed a restricted exploration of parameter space, based on a set of reference parameter values :

$$(J, J_0, \kappa, p) = (0.5, 0.5, \infty, 1). \quad (34)$$

Since $(\kappa, p) = (\infty, 1)$, the individual coupling weights J_{ij} are drawn independently and the connectivity matrix is dense. Hence, this is a classic SK model in its paramagnetic phase, because J and J_0 are smaller than 1 [37].

Figure 2 shows the fit performances of the various approximation methods on a typical configuration $(\mathbf{h}, \mathbf{J})$

with these generative parameters. All approximations perform well, as expected, since the TAP equations are exact when $N \to \infty$ in the paramagnetic phase of the SK model [4]. However, the variational Cox approximation (fourth panel) displays a bias : the overall magnitude of its predictions C_{ij} is somewhat underestimated, leading to a slant in the graph of $(C_{ij}^\star, C_{ij})$, and a larger fit value z. This bias is a systematic property of the Cox approximation, which is mainly caused by the presence of nonzero self-coupling terms $J_{ii} > 0$ (see Discussion).

I then explored the approximations' behavior in different departures from the SK paramagnetic situation. In each of Figures 3-6, one parameter in eq. (34) is varied while the three others keep their reference value. Panel (a) shows the various approximations on a typical configuration at the transition out of the 'paramagnetic SK' phase. Panel (b) shows the mean fit performance of each approximation as the concerned parameter is varied. Panel (c) shows when multiple solutions have been detected to each approximation's constitutive fixed point equation.

I first tested the approximations' behavior when transiting into the ferromagnetic (Figure 3) and spin-glass (Figure 4) phases of the SK model. The ferromagnetic phase, corresponding to $J_0 > \max(1, J)$, is characterized by a symmetry breaking into two 'magnetized' states with $s_i = 1$ (resp. -1) for all spins, constituting the stable solutions of the TAP equation [4, 6]. The spin glass phase, corresponding to $J > \max(1, J_0)$, is characterized by the apparition of multiple 'metastable' local minima of the TAP free energy, with limited basins of attraction [4]. All approximations considered have roughly the same behavior at the phase transitions (Figure 3, $J_0 \geq 1$, Figure 4, $J \geq 1$), losing precise fit (panels (b)) concurrently with the apparition of multiple solutions to their respective equations (panels (c)). This confirms the existence of a universal 'mean field' domain for the SK model, corresponding to its paramagnetic phase. [2]

When couplings are made sparser, all approximations again display the same qualitative behavior (Figure 5), maintaining a reasonable precision down to very diluted models. The Bethe approximation is the most efficient in this case, because the rarefaction of loops creates a 'tree-like' structure of connectivity.

The main difference between the approximations is their handling of structured coupling matrices $\mathbf{J}$ (Figure 6). When parameter κ decreases and couplings weights J_{ij} become correlated, the TAP and Bethe approximations deteriorate much faster than the variational Cox

---

[2] Looking in more detail, a notable qualitative difference exists between the different approximations, at least in the ferromagnetic phase. In the TAP and Bethe approximations, the symmetric solution with $\mathbf{m} = \mathbf{0}$ becomes unstable [4, 6], and the linear response prediction for $\mathbf{C}$ at this point diverges – as visible in Figure 3(a). In the adaptive TAP and variational Cox approximations, this symmetric solution remains stable and coexists with the magnetized solutions (as in Figure 1(b)), and the prediction for $\mathbf{C}$ is progressively degraded, rather than totally lost – see Figure 3(a).
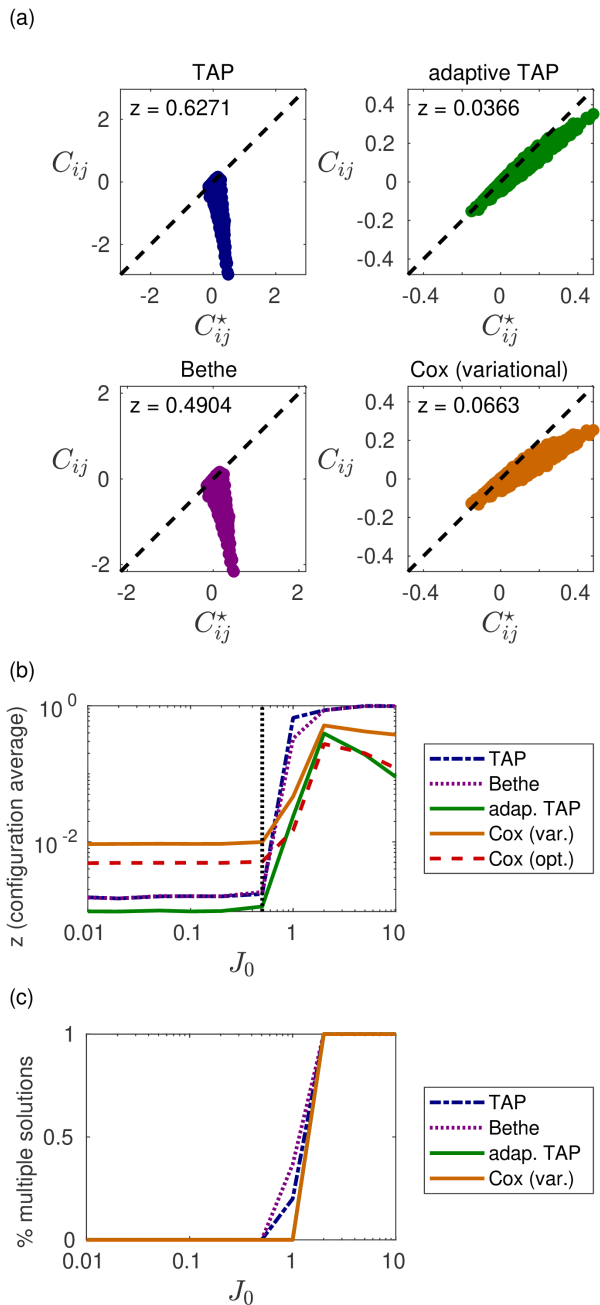

FIG. 3. Transition from paramagnetic to ferromagnetic phase, when parameter J_0 is increased. Other parameters keep their reference value $(J, \kappa, p) = (0.5, \infty, 1)$. (a) TAP, Bethe, adaptive TAP and variational Cox approximations on a typical configuration with $J_0 = 1$. (b) Average fit value z for each of the approximations, assessed over 30 configurations of matrix $\mathbf{J}$ picked according to eq. (32), for each tested value of J_0. Fit measure is also provided for the optimal Cox distribution (see text). The vertical dashed line corresponds to the reference values of Figure 2. (c) Apparition of multiple solutions to the respective fixed point equations. For each value of J_0 and approximation considered, I plot the percentage of the 30 configurations in which multiple solutions were found (see Appendix A, *Numerical procedures*).
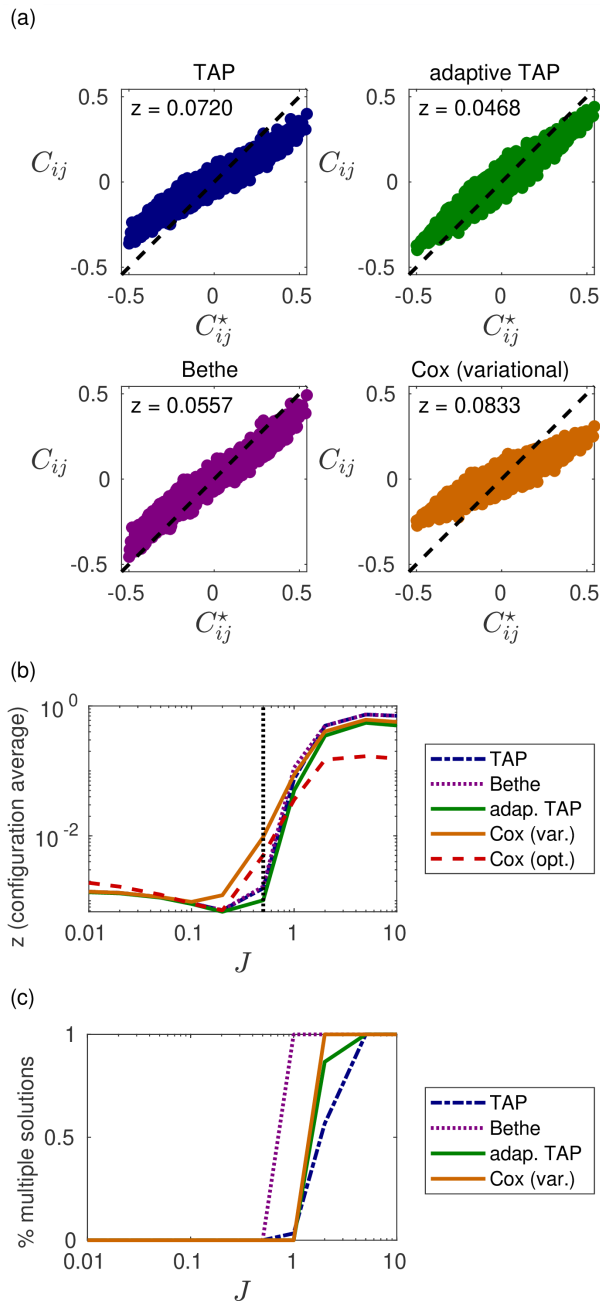
(a)

(a)

(b)

(b)

(c)

(c)

FIG. 4. Transition from paramagnetic to spin glass phase, when parameter J is increased. Other parameters keep their reference value $(J_0, \kappa, p) = (0.5, \infty, 1)$. (a) TAP, Bethe, adaptive TAP and variational Cox approximations on a typical configuration with $J = 1$. (b), (c) : same as Figure 3.

FIG. 5. Introduction of sparseness in the coupling weights, when parameter p is decreased. Other parameters keep their reference value $(J, J_0, \kappa) = (0.5, 0.5, \infty)$. (a) TAP, Bethe, adaptive TAP and variational Cox approximations on a typical configuration with $p = 0.01$. (b), (c) : same as Figure 3.

and adaptive TAP approximation – which was designed precisely for this purpose [33].

This numerical study confirms the adaptive TAP's interest as a 'universal' mean-field method, the most efficient in all tested regimes with dense couplings, and second most efficient in case of sparse couplings (the Bethe approximation performing marginally better). It also reveals similar domains of validity for the adaptive TAP and variational Cox approximations – notwithstanding
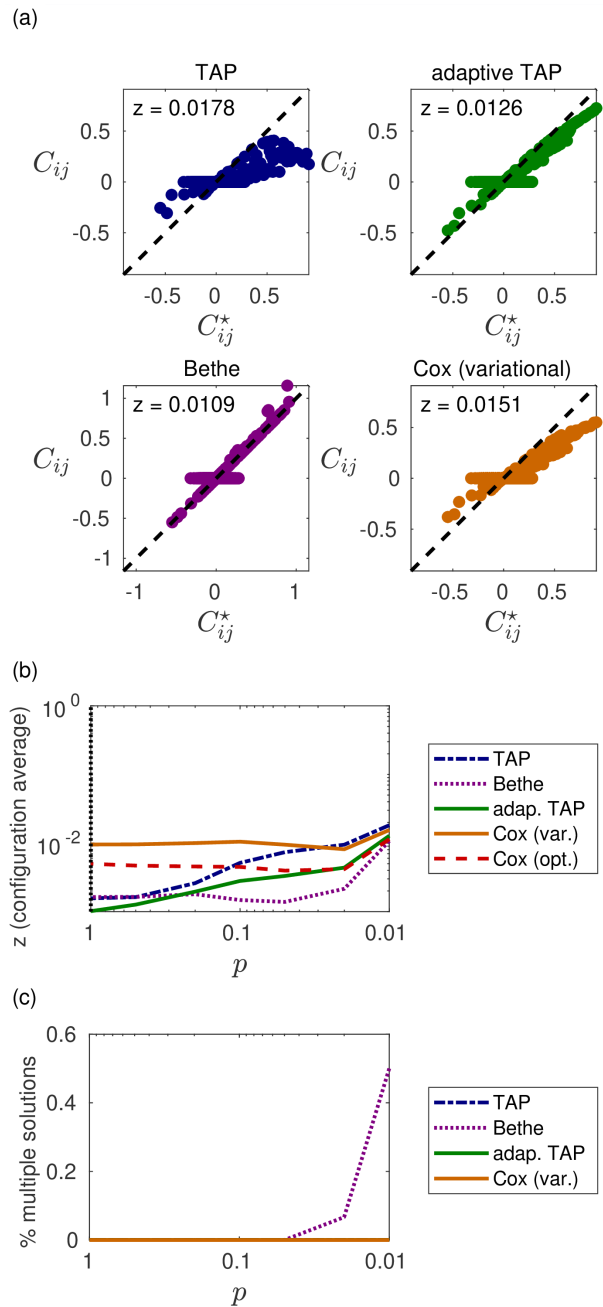
the latter's systematic bias in 'easy' configurations, leading to higher fit values z. In summary, a 'mean field domain' can be defined as the ensemble of parameters $(\mathbf{h}, \mathbf{J})$ for which the adaptive TAP method efficiently predicts the spin moments $(\mathbf{m}, \mathbf{C})$, and this is also the domain where the Ising distribution can be easily replaced by a Cox approximation, thanks to a variational principle.
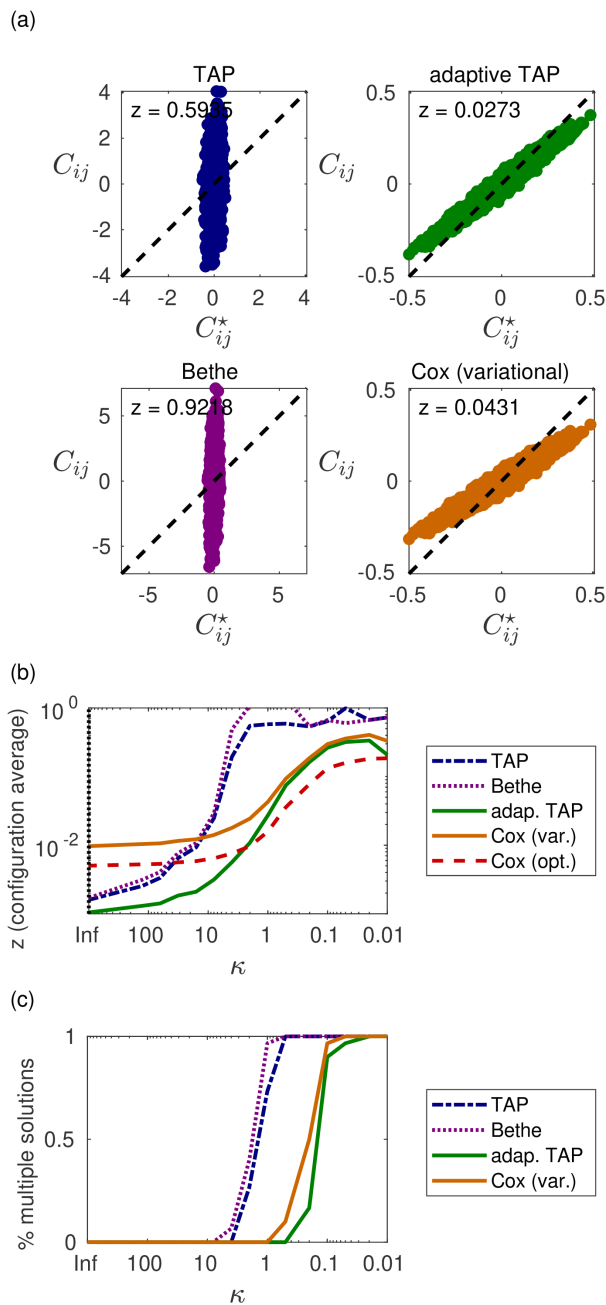
(a)

(b)

(c)

FIG. 6. Introduction of stochastic dependence between the coupling weights, when parameter κ is decreased. Other parameters keep their reference value $(J, J_0, p) = (0.5, 0.5, 1)$. (a) TAP, Bethe, adaptive TAP and variational Cox approximations on a typical configuration with $\kappa = 1$. (b), (c) : same as Figure 3.

Optimal Cox distribution

As such, the above results do not explicitly tell whether the Ising distribution can be approximated by a Cox distribution outside of the 'mean field domain'. It may be the case that a decent Cox approximation exists, but cannot be retrieved by a variational principle anymore. To clarify this point, I also considered the fit performance of the *optimal* Cox distribution $Q(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$ defined above. My measure of fit in this case consisted in comparing the true spin covariances $C_{ij}^\star$ to their values in the optimal Cox distribution, as given by eq. (21). Thus, a successful fit indicates when assuming a Gaussian shape for the latent field distribution $P(\mathbf{r})$, without modifying its moments, does not modify much the resulting spin moments.

Figures 3-6(b) show that this measure globally correlates with the efficiency of the adaptive TAP and variational Cox approximations, i.e, it also deteriorates outside of the 'mean field domain'. Hence, the increased discrepancy between Ising and Cox distributions outside of the 'mean field domain' seems intrinsically related to the Ising latent field $P(\mathbf{r})$ becoming non-Gaussian. This is also coherent with the fact that the adaptive TAP approximation is bound to fail precisely when the instantaneous field acting on each spin cannot be considered Gaussian (see Appendix C).

Nonetheless, in quantitative terms, the loss of fit by the optimal Cox distribution is never total. Figure 7 shows examples of fit performance for the optimal Cox distribution in various configurations at the boundary (first row, compare to Figures 3-6(a)) and far outside (second row) of the mean field domain. It reveals that, even when the Ising latent field $P(\mathbf{r})$ is far from being Gaussian-distributed, its replacement by a Gaussian preserves the overall pattern of spin correlations (as visible in Figure 1(b)). This global correctness is hardly reflected in the magnitude of error z, yet it does imply that an Ising model is never 'too far' away from its optimal Cox distribution.

VI. DISCUSSION

I have proposed a reformulation of the Ising distribution as a latent variable model, and used it to derive principled approximations by the simpler Cox distribution. In practical applications, Cox models (including the dichotomized Gaussian, and Cox point processes) are often preferred to the corresponding maximum entropy distributions (Ising distribution, Gibbs point process) because they are easier to sample, and to parametrize from a set of observed moments. This article establishes a simple analytical connection between the two families of models, and investigates under what conditions they can be used interchangeably.

The most natural connection between an Ising and a Cox distribution is obtained by equating the two first moments of their latent variables, eq. (10)-(11), a simple but fundamental result of the article. The resulting 'optimal' Cox approximation holds well in paramagnetic conditions, and even beyond, as far as global trends are concerned (Figure 7). However, eq. (10)-(11) involve both the natural parameters $(\mathbf{h}, \mathbf{J})$ and the resulting moments $(\mathbf{m}^\star, \mathbf{C}^\star)$ of the Ising distribution, which makes them unpractical in most concrete situations.

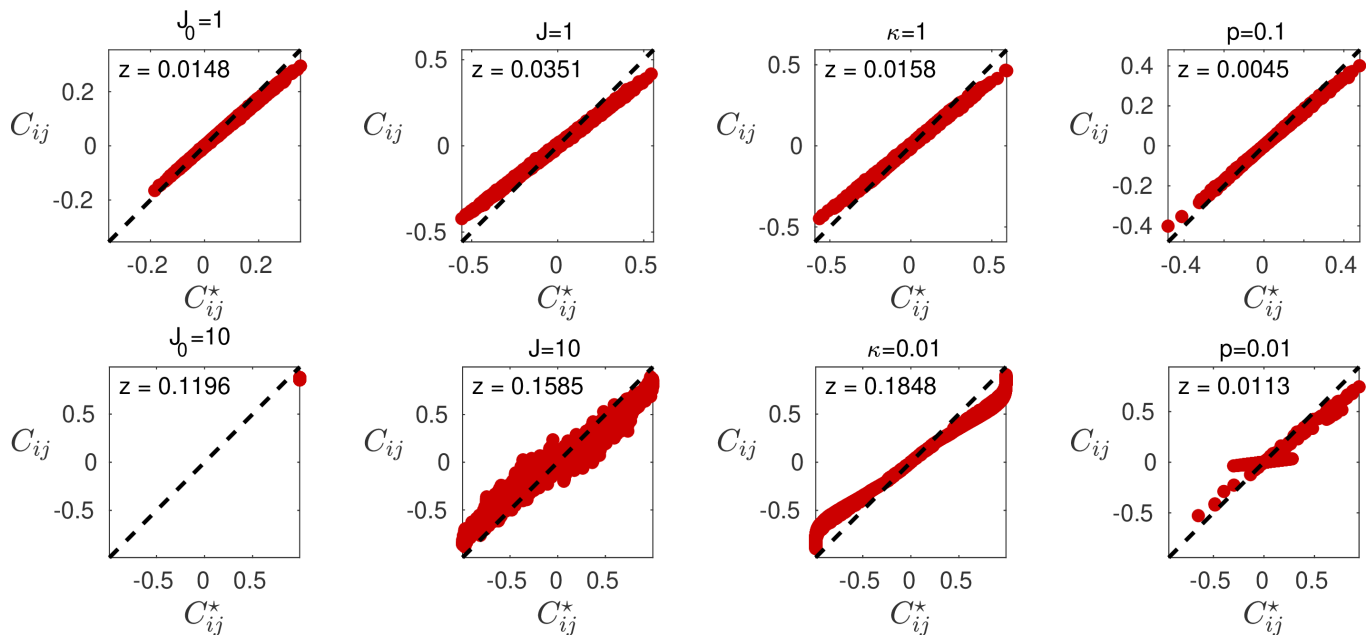To target a Cox approximation given only some natu-

11

FIG. 7. Fit for the optimal Cox distribution on typical configurations at the boundary (first row) and far outside (second row) of the paramagnetic SK phase. Each panel plots the pairwise spin covariances in the true Ising distribution ($C_{ij}^\star$) against their values in the optimal Cox distribution (C_{ij}), computed with eq. (21). Starting from reference parameter values $(J_0, J, \kappa, p) = (0.5, 0.5, \infty, 1)$, the approximation is assessed at different values of J_0 (first column, transition to ferromagnetic SK model), J (second column, transition to spin glass SK model), κ (third column, introduction of correlations between coupling weights J_{ij}), and p (fourth column, dilution of connectivity).

ral parameters $(\mathbf{h}, \mathbf{J})$, I have explored a classic variational approach, leading to eq. (14)-(18). These equations are not particularly interesting as a mean-field method for predicting $(\mathbf{m}, \mathbf{C})$, since they globally behave as a biased version of Opper and Winther's adaptive TAP method. However, their analytical (Section IV) and numerical (Section V) analysis allows to formulate the key conclusion of this article : mean-field methods are efficient precisely when the Ising distribution can be associated to a quasi-normal latent field distribution.

If the practical goal is to establish a Cox approximation for the Ising distribution of parameters $(\mathbf{h}, \mathbf{J})$, the moments $(\mathbf{m}, \mathbf{C})$ may as well be estimated with any other choice of mean field method, and then input to eq. (14)-(15) to produce the corresponding Cox approximation.

Naturally, the relation to mean field methods is not accidental. From eq. (8), given a spin configuration $\mathbf{s}$, the latent field r_i is distributed as $\mathcal{N}(h_i + \sum_j J_{ij}s_j, J_{ii})$. In particular, if self-couplings J_{ii} are zero as in the classic Ising model, we simply recover

$$r_i = h_i + \sum_{j \neq i} J_{ij}s_j,$$

that is, the instantaneous field variable considered in the cavity method (Appendix C, eq. (C1)) and adaptive TAP equations. Thus, the latent field formalism differs from the classic cavity approach only through the role of nonzero self-couplings J_{ii}. Coherently, the weak coupling expansions of the variational Cox and adaptive TAP approximations up to order 4 differ only because of nonzero

J_{ii}.

This suggests that nonzero self-couplings J_{ii} may be responsible for the systematic bias of the Cox approximation in its prediction of $(\mathbf{m}, \mathbf{C})$, compared to the adaptive TAP approximation. As a direct confirmation, I observed that when matrix $\mathbf{J}$ is given a zero diagonal, the variational Cox equation (14)-(18) generally retains a solution, and it is then remarkably close to the adaptive TAP solution (see Supplementary Material). Unfortunately, nonzero weights J_{ii} are required to endow the field variables r_i with a true, multivariate distribution $P(\mathbf{r})$ (this is not the case in the cavity method), and thus produce a *concrete* approximation of the Ising distribution by a simpler latent variable distribution.

A disturbing consequence is that there is not one latent field distribution associated to the Ising distribution, but *many* different distributions, depending on the value given to self-couplings J_{ii}. When all values J_{ii} are taken very large, the latent field $P(\mathbf{r})$ is 'useless' : it is the mere mixture of 2^N Gaussian bumps with no overlap, located on the 2^N summits of a hypercube, and the bump at summit $\mathbf{s}$ is simply associated to weight $P(\mathbf{s})$ (as in Figure 1(b)). Then, as the J_{ii} become smaller, some of the 2^N bumps start overlapping, and $P(\mathbf{r})$ acquires a less trivial overall distribution. In some cases, when the J_{ii} are made small enough, the overall shape of $P(\mathbf{r})$ becomes quasi-Gaussian (as in Figure 1(a)). In other cases, $P(\mathbf{r})$ never becomes quasi-Gaussian, because a lower bound is reached where the J_{ii} cannot be made smaller while ensuring that matrix $\mathbf{J}$ remains definite positive.

Given some classic Ising parameters $\{J_{ij}\}_{i<j}$, the set of self-couplings J_{ii} obtained with eq. (19) represents the "smallest" diagonal elements that can be used, and we could suppose that if the latent field $P(\mathbf{r})$ is still not Gaussian for these values, it won't be either for other values of J_{ii}. Interestingly, the optimization problem in eq. (19) has an intrinsic significance for the 'classic' Ising distribution (without self-couplings) : the resulting value of $-\sum_i J_{ii}$ constitutes the Lagrangian dual approximation for the minimum of $\sum_{i<j} J_{ij} s_i s_j$ over all 2^N spin configurations [29], that is, the log-likelihood of the most unlikely spin configuration in the model. Arguments of this type suggests that, while self-couplings J_{ii} are extraneous elements to the standard Ising model, the *set of admissible values* for the J_{ii} may have theoretical links with the nature and global difficulty of the considered Ising model.

Empirically, I observed a certain robustness to the exact choice of diag($\mathbf{J}$), and the values could generally be doubled without affecting much the numerical results. In some cases, increasing all self-couplings J_{ii} a little from the 'optimal' solution of eq. (19) can even improve the fit performance of the variational Cox approximation, presumably because it reduces the condition number of matrix $\mathbf{J}$ (see also Appendix A). Thus, theoretically as well as practically, the significance and optimal choice of diag($\mathbf{J}$) is not totally settled, and could be the subject of future work.

More generally, the present work could give rise to a number of developments. For example, the theoretical study of phase transitions in the variational Cox approximation, as observed in Figures 3-6(c), remains to be done. Intuitively, these phase transitions are related to the apparition of multiple, well-separated modes in the distribution of the latent fields (see Figure 1), which would then play a role similar to the "pure states" of spin glass theory [4, 6].

On a more applied level, the latent field variables r_i could be incorporated into MCMC sampling schemes for the Ising distribution. For example, drawing the initial spins s_i with a Cox approximation $Q(\mathbf{s})$, instead of a classic independent Bernoulli draw, can largely reduce the chain's convergence time to its equilibrium distribution $P(\mathbf{s})$. One step further, one could devise MCMC schemes that directly sample the fields' distribution $P(\mathbf{r})$ and use it to generate the spins.

Finally, the formalism of latent variables could be applied to the inverse problem of retrieving $(\mathbf{h}, \mathbf{J})$ from a set of observed moments $(\mathbf{m}, \mathbf{C})$, which is arguably the biggest obstacle in practical applications of the Ising model listed above [7–9]. On the one hand, advanced methods of mean-field inspiration have been developed in the last decade to tackle the inverse Ising problem [13, 14, 39]. On the other hand, it has been suggested to replace the Ising distribution by a dichotomized Gaussian (a limiting case of Cox distribution) in practical applications, precisely because it offers an easier inverse problem [19]. Hopefully, the latent field formalism can reconcile the two approaches in a unified picture.

Appendix A: Numerical procedures

In this appendix, I give numerical details of the tests presented in Section V. All approximation methods considered are naturally described by a fixed point equation of the form $X = F(X)$. A simple heuristic for solving such equations is a numerical scheme

$$X_{n+1} = (1 - \alpha)X_n + \alpha F(X_n),$$

with α a small, adaptive, update parameter. When $\|X_n - F(X_n)\|$ is found to increase between two successive iterations, α is divided by 10. Else, α is multiplied by 1.05. This heuristic proved sufficient to target a fixed point, in all cases encountered.

In the Cox approximation, the variable was $X = \{m_i, d_i^{-1}\}_{i=1...N}$. In the adaptive TAP approximation, it was $X = \{m_i, V_i\}_{i=1...N}$. The starting point X_0 was chosen with $m_i = 0$, and numbers $\Delta_i = d_i^{-1}$ (resp $\Delta_i = 1 + V_i$) as the minimal values ensuring that matrix $\mathbf{C}^{-1} = \mathrm{diag}(\Delta_i) - \mathbf{J}$ be definite positive (see eq. (18) and eq. (27)).

In the TAP and Bethe approximations, the starting point $X_0 = \mathbf{m} = \mathbf{0}$ was directly a solution of the fixed point equation, and I simply used the linear response prediction for $\mathbf{C}$ at this point. This seemed the fairer choice in the context of these tests, even though in some parameter regimes (e.g., SK ferromagnetic phase) the solution at $\mathbf{m} = \mathbf{0}$ is unstable : the iterative scheme started at any neighboring point does not converge to this solution.

It is well-known that, in certain regimes of parameters, the Ising model can display spontaneous symmetry breaking [4]. First, ergodicity breaking can occur in the MCMC chain used to sample the distribution. To counteract this effect, the true moments $(m_i^\star, C_{ij}^\star)$ were estimated from several independent MCMC chains with simulated annealing.

Second, the various approximations themselves can start displaying multiple solutions to their constitutive equation [4]. To assess this effect, for each tested configuration $(\mathbf{h}, \mathbf{J})$ and approximation formula, I relaunched the numerical search from different starting points, namely, the empirical means found in each individual MCMC chain used during the sampling phase. This procedure only served as a (rudimentary) attempt to detect the presence of multiple solutions. It did not affect the measure of fit for the approximation, which was always based on the solution found from the starting points X_0 listed above.

In the Cox approximations (optimal and variational), for each configuration with off-diagonal elements $\{J_{ij}\}_{i<j}$, the diagonal couplings J_{ii} were chosen as the solution of eq. (19), plus a constant ridge term λ chosen for $\mathbf{J}$ to have a condition number of 10. Indeed, ill-conditioned matrices $\mathbf{J}$ lead to increased errors in some parameter regimes.

Appendix B: Variational mean field approximations

In this appendix, I recall the variational approach of mean field theory, which can be used to recover the Plefka expansion of the exact Ising model, as well as the TAP and Bethe approximations.

Mean field variational approach

As in the main text, let us note $P(\mathbf{s})$ the true Ising distribution of parameters $(\mathbf{h}, \mathbf{J})$, and $(\mathbf{m}^\star, \mathbf{C}^\star)$ its corresponding moments. Let $Q(\mathbf{s})$ any other distribution proposed as an approximation of P. The KL divergence $\mathrm{KL}(Q||P) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln(Q(\mathbf{s})/P(\mathbf{s}))$ can be expressed as

$$\mathrm{KL}(Q||P) = U(Q) - S(Q) + \ln Z_I(\mathbf{h}, \mathbf{J}),$$

where $U(Q)$ is the average Ising energy, and $S(Q)$ the entropy, under distribution Q :

$$U(Q) := -\sum_{\mathbf{s}} Q(\mathbf{s}) \left(\mathbf{h}^\top \mathbf{s} + \frac{1}{2} \mathbf{s}^\top \mathbf{J} \mathbf{s} \right),$$

$$S(Q) := -\sum_{\mathbf{s}} Q(\mathbf{s}) \ln Q(\mathbf{s}).$$

The functional $G(Q) := U(Q) - S(Q)$ is called the *variational (or Gibbs) free energy* of distribution Q as an approximation of P. Smaller values of $G(Q)$ correspond to a lower KL divergence and thus to a better fit, and its minimum is achieved for $Q = P$.

In the Ising distribution, given fixed couplings $\mathbf{J}$, there is a one-to-one correspondence between values of $\mathbf{h}$ and resulting values of $\mathbf{m} = \mathrm{E}(\mathbf{s})$. Thus, in theory, we can apply the variational approach to the family of Ising distributions $Q(\mathbf{m}, \mathbf{J})$, where couplings $\mathbf{J}$ are taken equal to those in P, and $\mathbf{m}$ constitutes the N-dimensional parametrization variable. The natural field parameters of Q, say $\boldsymbol{\theta}$, are then a (generally intractable) function of $(\mathbf{m}, \mathbf{J})$, and the associated free energy writes

$$G(\mathbf{m}, \mathbf{J}) = (\boldsymbol{\theta} - \mathbf{h})^\top \mathbf{m} - \ln Z_I(\boldsymbol{\theta}, \mathbf{J}), \qquad (\text{B1})$$

whose minimum over $\mathbf{m}$ is obtained when $\mathbf{m} = \mathbf{m}^\star$, that is, when $Q = P$. (Note that G also depends on the field parameter $\mathbf{h}$ of distribution P, but I omit it for lighter notations.)

Using the classic conjugacy relation $\partial_{\boldsymbol{\theta}} \ln Z_I = \mathbf{m}$ in the Ising model, one can note that

$$\partial_{\mathbf{m}} G(\mathbf{m}, \mathbf{J}) = \boldsymbol{\theta} - \mathbf{h}, \qquad (\text{B2})$$

so function $G(\mathbf{m}, \mathbf{J}) + \mathbf{h}^\top \mathbf{m}$ corresponds to the Legendre transform of $\ln Z_I$ in its first variable.

Function $G(\mathbf{m}, \mathbf{J})$ is not tractable in general, but it can be approximated – and the resulting minimum will yield an approximation of the true moments $\mathbf{m}^\star$. This approach is known as the *mean field* variational method, of which the TAP and Bethe approximations are two prominent examples.

TAP and Plefka approximations

In so-called *Plefka expansions*, one approximates $G(\mathbf{m}, \mathbf{J})$ by its Taylor expansion in $\mathbf{J}$ around $\mathbf{J} = \mathbf{0}$:

$$G_n(\mathbf{m}, \mathbf{J}) := G(\mathbf{m}, \mathbf{0}) + [\partial_{\mathbf{J}} G(\mathbf{m}, \mathbf{0})](\mathbf{J}) + \frac{1}{2}[\partial_{\mathbf{J}}^2 G(\mathbf{m}, \mathbf{0})](\mathbf{J}, \mathbf{J})$$

$$+ \cdots + \frac{1}{n!}[\partial_{\mathbf{J}}^n G(\mathbf{m}, \mathbf{0})](\mathbf{J}, \ldots, \mathbf{J}),$$

where $\partial_{\mathbf{J}}^n G(\mathbf{m}, \mathbf{0})$ is the n-th derivative of G wrt $\mathbf{J}$ (a symmetric tensor of order n) evaluated at point $(\mathbf{m}, \mathbf{0})$. All these terms can be evaluated, albeit laboriously. First, the fundamental relation $\partial_{J_{ij}} \ln Z_I = m_i m_j + \partial_{\theta_i, \theta_j}^2 \ln Z_I$ allows to replace derivatives wrt. $\mathbf{J}$ by derivatives wrt. $\boldsymbol{\theta}$. Second, at $\mathbf{J} = \mathbf{0}$, the Ising distribution boils down to a Bernoulli distribution, where all derivatives wrt. $\boldsymbol{\theta}$ are fully tractable.

The approximation at order $n = 4$ is a classic computation [27, 40–42], which yields :

$$
\begin{aligned}
G_4(\mathbf{m}, \mathbf{J}) = & \sum_i G(m_i) - \sum_{(ij)} J_{ij} m_i m_j - \frac{1}{2} \sum_{(ij)} J_{ij}^2 c_i c_j \\
& - \sum_{(ijk)} J_{ij} J_{jk} J_{ki} c_i c_j c_k - \frac{2}{3} \sum_{(ij)} J_{ij}^3 m_i c_i m_j c_j \\
& - \sum_{(ijkl)} J_{ij} J_{jk} J_{kl} J_{li} c_i c_j c_k c_l \\
& - 2 \sum_{(ij),k} J_{ij}^2 J_{jk} J_{ki} m_i c_i m_j c_j c_k \\
& + \frac{1}{12} \sum_{(ij)} J_{ij}^4 c_i c_j (1 + 3m_i^2 + 3m_j^2 - 15 m_i^2 m_j^2).
\end{aligned}
$$
$$(\text{B3})$$

Here, (ij), (ijk), $(ijkl)$ indicate respectively all unordered pairs, triplets and quadruplets of distinct spins. $G(m_i)$ is the free energy of each 1-spin marginal distribution (eq. (B1) with $\theta_i = \tanh^{-1}(m_i)$). Finally, we use the shorthand $c_i = 1 - m_i^2$.

By differentiating this function wrt. $\mathbf{m}$, we obtain a fixed point characterization of its extremum(s) $\mathbf{m}$. Stopping at order 1 in $\mathbf{J}$ yields the "naive" mean field equation. Stopping at order 2 (first line) yields the TAP equation. Stopping at order 3 (two first lines) yields eq. (29) from the main text.

Bethe approximation

In one particular case, $G(\mathbf{m}, \mathbf{J})$ in eq. (B1) is tractable exactly. This is when the underlying couplings J_{ij} define a *tree topology*, that is, they are zero except on a subset of the edges defining a graph without loops. In that case, the test Ising distribution $Q(\mathbf{m}, \mathbf{J})$ can be written as

$$Q(\mathbf{s}|\mathbf{m}, \mathbf{J}) = \prod_{\langle ij \rangle} \frac{Q(s_i, s_j)}{Q(s_i) Q(s_j)} \prod_i Q(s_i), \qquad (\text{B4})$$

where $\langle ij \rangle$ denotes all edges in the tree. This can be proved by repeated applications of Bayes' formula, starting from any leaf of the tree. Besides, each marginal $Q(s_i, s_j)$ is a 2-spin Ising distribution with coupling parameter J_{ij}, as proved directly by integrating out the remaining variables from the original Ising formula.

In consequence, the (exact) free energy writes

$$G_{\mathrm{B}}(\mathbf{m}, \mathbf{J}) = \sum_{i<j} G(m_i, m_j, J_{ij}) - (N-2) \sum_i G(m_i),$$
(B5)

$G(m_i)$ and $G(m_i, m_j, J_{ij})$ being the respective free energies of the marginal distributions $Q(s_i)$ and $Q(s_i, s_j)$, as defined by eq. (B1). Note that the sum can be made over all spin pairs $i < j$, and not just neighboring pairs $\langle ij \rangle$ in the tree. Indeed, unconnected spin pairs yield a zero contribution, as $G(m_i, m_j, 0) = G(m_i) + G(m_j)$.

The *Bethe approximation* consists in using eq. (B5) as an approximation for the free energy, even when the J_{ij} do not have a tree topology.

Imposing that $\partial_{\mathbf{m}} G_{\mathrm{B}} = 0$ and using eq. (B2), leads to the N equations

$$\forall i, \quad (N-2) \tanh^{-1}(m_i) = \Big(\sum_j \theta_i^{(j)} \Big) - h_i, \quad (B6)$$

where the so-called *cavity fields* $(\theta_i^{(j)}, \theta_j^{(i)})$ are the natural field parameters of each 2-spin Ising distribution $Q(s_i, s_j)$, that is, tractable functions of (m_i, m_j, J_{ij}). In fact, it is easily shown that the 2-spin Ising distribution of natural parameters $(\theta_i^{(j)}, \theta_j^{(i)}, J_{ij})$ has moments

$$m_i = \frac{t_i + t_j t_{ij}}{1 + t_i t_j t_{ij}} \quad , \quad c_{ij} = t_{ij} \frac{(1 - t_i^2)(1 - t_j^2)}{(1 + t_{ij} t_i t_j)^2}, \quad (B7)$$

with $t_{ij} = \tanh(J_{ij})$, $t_i = \tanh(\theta_i^{(j)})$, $t_j = \tanh(\theta_j^{(i)})$. Inserting this expression for m_i into eq. (B6), with $\tanh^{-1}(m) = \frac{1}{2}(\ln(1+m) - \ln(1-m))$, and after some linear recombinations, we obtain

$$\forall (i,j), \quad \theta_i^{(j)} = h_i + \sum_{k \neq j} \tanh^{-1} \big(t_{ik} \tanh(\theta_k^{(i)}) \big), \quad (B8)$$

and the optimum is now characterized by $N(N-1)$ equations over the $N(N-1)$ cavity variables. This switching from N principal variables (the m_i) to $N(N-1)$ auxiliary variables (the $\theta_i^{(j)}$) can also be interpreted as a dual Lagrangian optimization procedure [31].

In a tree-like topology, eq. (B8) can be solved iteratively starting from any leaf of the tree, allowing to recover the exact values for all the $\theta_i^{(j)}$, and thus, for magnetizations $m_i^\star$. The resulting algorithm is known as *belief propagation*, or *sum-product*. In a general topology, the fixed point approach to characterize solutions of eq. (B8) is known as *loopy belief propagation*. It is not guaranteed to have a single solution anymore – and it only characterizes an approximation for the magnetizations m_i.

The covariances C_{ij} can be approximated in turn, based on the linear response formula, eq. (22). Applied to each 2-spin distribution of natural parameters $(\theta_i^{(j)}, \theta_j^{(i)}, J_{ij})$, it implies the differential equality

$$\forall (i,j), \quad (\partial m_i) = (1 - m_i^2)(\partial \theta_i^{(j)}) + c_{ij}(\partial \theta_j^{(i)}) \quad (B9)$$

with c_{ij} given by eq. (B7). We can then differentiate eq. (B6) as a function of (∂h_i) and linearly eliminate the cavity fields $(\partial \theta_i^{(j)})$ thanks to eq. (B9). As a result, we express (∂h_i) as a function of the (∂m_j) only, and this provides the linear response prediction :

$$(\mathbf{C}_{\mathrm{B}}^{-1})_{ij} = \left(1 + \sum_k \frac{c_{ik}^2}{(1 - m_i^2)(1 - m_k^2) - c_{ik}^2} \right) \frac{\delta_{ij}}{1 - m_i^2}$$
$$- \frac{c_{ij}}{(1 - m_i^2)(1 - m_j^2) - c_{ij}^2}. \quad (B10)$$

This derivation, which I could not find in the literature, expresses the linear response matrix $\mathbf{C}_{\mathrm{B}}$ in an alternative form than in [36].

To derive the weak coupling (Plefka) expansion of the Bethe approximation, note that each $G(m_i, m_j, J_{ij})$ in eq. (B5) is an exact Ising free energy over two spins i and j. Compared to a generic Ising free energy over N spins, its expansion only contains the two-spin diagrams, summed over the single spin pair involved. It follows, after summing over all spin pairs $i < j$ in eq. (B5), that the Plefka expansion of the Bethe free energy is obtained by keeping only the pairwise diagrams in the expansion for the true Ising free energy.

Appendix C: Cavity method and adaptive TAP equations

Cavity method

The *cavity method* is a classic approach allowing to recover many analytical properties of the Ising model, and other multivariate exponential models [4, 5]. Singling out an arbitrary spin location i, one can rewrite eq. (1) as

$$P(\mathbf{s}_{\setminus i}, s_i) \sim P_{\setminus i}(\mathbf{s}_{\setminus i}) \exp \Big(s_i \big(h_i + \sum_{j \neq i} J_{ij} s_j \big) \Big),$$

where $\mathbf{s}_{\setminus i}$ denotes the remaining $N-1$ spins, and the so-called *cavity distribution* $P_{\setminus i}$ is the Ising distribution obtained by deleting line and column i from $(\mathbf{h}, \mathbf{J})$. The remaining spins interact with i only through the random variable

$$\tilde{r}_i := h_i + \sum_{j \neq i} J_{ij} s_j, \quad (C1)$$

and we can write

$$P(\tilde{r}_i, s_i) \sim P_{\setminus i}(\tilde{r}_i) \exp(s_i \tilde{r}_i), \quad (C2)$$

$P_{\backslash i}(\tilde{r}_i)$ indicating the distribution of variable $\tilde{r}_i$ when the spins $\mathbf{s}_{\backslash i}$ follow the cavity distribution $P_{\backslash i}$.

In general, distribution $P_{\backslash i}(\tilde{r}_i)$ is not tractable exactly [3]. But in many circumstances, since $\tilde{r}_i$ is the sum of variables with many degrees of freedom, it can be assumed to have a normal distribution :

$$P_{\backslash i}(\tilde{r}_i) = \mathcal{N}(\tilde{r}_i | \theta_i, V_i),$$

and eq. (C2) becomes, approximately :

$$P(\tilde{r}_i, s_i) \sim \exp\left(-\frac{1}{2V_i}(\tilde{r}_i - \theta_i)^2 + s_i \tilde{r}_i\right). \qquad \text{(C3)}$$

This equation is the starting point of the classic cavity method. Note the formal similarity with our definition for the joint probability of spins and latent fields, eq. (4), so we can simply recycle our results. From eq. (7), $P(s_i)$ is the Bernoulli distribution of parameter θ_i, so $m_i = \tanh(\theta_i)$. From eq. (14), we have $\mathrm{E}(\tilde{r}_i) = \theta_i + V_i m_i$. Taken together, this yields the generalized TAP equation :

$$m_i = \tanh\left(h_i + \sum_j J_{ij} m_j - m_i V_i\right). \qquad \text{(C4)}$$

For the SK model in the limit $N \to \infty$, it can be shown that the different summands to $\tilde{r}_i$ in the cavity distribution (eq. (C1)) become linearly independent [4, 6], so the variance writes $V_i \simeq \sum_j J_{ij}^2(1 - m_j^2)$ and we recover the classic TAP equation.

Adaptive TAP approximation

Instead, in the *adaptive TAP* method [33, 34], V_i is left as a free variable which can adapt to any statistical structure of the couplings $\mathbf{J}$. First, differentiating eq. (C4) wrt. $\mathbf{h}$ (but neglecting the dependency of V_i itself) leads to a linear response prediction for $\mathbf{C}$:

$$(\mathbf{C}_{\mathrm{A}}^{-1})_{ij} = \left(1 + (1 - m_i^2)V_i\right)\frac{\delta_{ij}}{1 - m_i^2} - J_{ij}. \qquad \text{(C5)}$$

Second, given magnetization m_i, the individual variance of spin i should be $1 - m_i^2$. Self-coherence of the variance prediction imposes that

$$1 - m_i^2 = (\mathbf{C}_{\mathrm{A}})_{ii}. \qquad \text{(C6)}$$

Taken together, eq. (C4)-(C6) constitute a system on variables $\{m_i, V_i\}$, which can be solved by classic iterative methods [34].

I now turn to the weak coupling (Plefka) expansion of eq. (C4)-(C6), when magnetizations m_i are fixed, and the

---

[3] Except when the couplings J_{ij} have a tree-like topology. In this case, $P_{\backslash i}(\tilde{r}_i)$ is a factorized product over the neighboring spins of i, and this is another way of deriving the Bethe equation (B8) [32].

coupling matrix writes $\alpha \mathbf{J}$. Given the form of eq. (C4), this only requires to obtain the expansion for $m_i V_i$ or, after a convenient rescaling, for variable

$$x_i := V_i(1 - m_i^2),$$

for which we want to establish the Taylor development

$$x_i = \alpha x_i^{[1]} + \alpha^2 x_i^{[2]} + \alpha^3 x_i^{[3]} + \dots$$

(note that $x_i = V_i = 0$ when $\alpha = 0$).

From eq. (C5), it is clear that the solution depends on the diagonal of $\mathbf{J}$ only through the simple offset $V_i \to V_i + J_{ii}$, so we may assume $J_{ii} = 0$ without loss of generality.

Introducing the variables

$$d_i := \frac{1 - m_i^2}{1 + (1 - m_i^2)V_i} = (1 - m_i^2)\left[1 - x_i + x_i^2 - x_i^3 + \dots\right],$$

and matrix $\mathbf{D} = \mathrm{diag}(d_i)$, we rewrite eq. (C5) as

$$\mathbf{C}^{-1} = \mathbf{D}^{-1} - (\alpha \mathbf{J})$$

and thus, after a classic switching from $\mathbf{C}^{-1}$ to $\mathbf{C}$:

$$\mathbf{C} = \mathbf{D} + \alpha \mathbf{DJD} + \alpha^2 \mathbf{D}(\mathbf{JD})^2 + \alpha^3 \mathbf{D}(\mathbf{JD})^3 + \dots,$$

allowing to easily express the development of $\mathbf{C}$ from that of d_i.

Then, noting $c_i := 1 - m_i^2$ for concision, the fixed point equation (C6) imposes, at order 4 :

$$\begin{aligned}
1 = {} & \left(1 - x_i + x_i^2 - x_i^3 + x_i^4\right) \\
& + \alpha^2 \sum_j c_i J_{ij}^2 c_j \left(1 - 2x_i + 3x_i^2 - x_j + 2x_i x_j + x_j^2\right) \\
& + \alpha^3 \sum_{j,k} c_i J_{ij} c_j J_{jk} c_k J_{ki}\left(1 - 2x_i - x_j - x_k\right) \\
& + \alpha^4 \sum_{j,k,l} c_i J_{ij} c_j J_{jk} c_k J_{kl} c_l J_{li} \quad + o(\alpha^4).
\end{aligned}$$

We can then replace x_i by its expansion, and regroup the powers of α. For the equation to be verified at order 1, this imposes that

$$x_i^{[1]} = 0.$$

Using this newly found value, the fixed point equation at order 2 imposes

$$\frac{m_i}{c_i} x_i^{[2]} = m_i \sum_{j \neq i} J_{ij}^2 c_j.$$

Then, the fixed point equation at order 3 imposes

$$\frac{m_i}{c_i} x_i^{[3]} = 2m_i \sum_{(jk|i)} J_{ij} J_{jk} J_{ki} c_j c_k.$$

where $(jk|i)$ denotes all unordered triplets of the form $\{i, j, k\}$ with j and k distinct, and distinct from i. By

inserting these values into eq. (C4), we recover the expansion from the main text, eq. (30).

Finally, the fixed point equation at order 4 yields

$$\frac{m_i}{c_i} x_i^{[4]} = 2m_i \sum_{(jkl|i)} J_{ij} J_{jk} J_{kl} J_{li} c_j c_k c_l - m_i c_i \sum_{j \neq i} J_{ij}^4 c_j^2.$$
(C7)

Note that the sum over $(jkl|i)$ – which involves the most terms and is generally dominant – is identical to that for the true Ising expansion : see eq. (B3).

Appendix D: Weak coupling expansion for the Cox approximation

We consider the 'variational' Cox approximation, solution to the equations

$$m_i = \int_{x \in \mathbb{R}} \tanh\left(\mu_i + x\sqrt{\Sigma_{ii}}\right) \phi(x)\mathrm{d}x,$$
(D1)

$$d_i = \int_{x \in \mathbb{R}} \left(1 - \tanh^2\left(\mu_i + x\sqrt{\Sigma_{ii}}\right)\right) \phi(x)\mathrm{d}x,$$
(D2)

$$(\mathbf{C}^{-1})_{ij} = d_i^{-1}\delta_{ij} - \alpha J_{ij}.$$
(D3)

$$\boldsymbol{\mu} = \mathbf{h} + \alpha \mathbf{Jm},$$
(D4)

$$\boldsymbol{\Sigma} = \alpha \mathbf{J} + \alpha^2 \mathbf{JCJ},$$
(D5)

when magnetizations m_i are fixed, and the coupling matrix writes $\alpha \mathbf{J}$, α being the small parameter of the expansion.

Here, I detail the computation up to order 3, and also provide the result at order 4. The overall structure of the computation is largely similar to that for the adaptive TAP approximation, in the previous paragraph.

When $\alpha = 0$, the solution is obvious : couplings $\alpha \mathbf{J}$ vanish, and so does the covariance matrix $\boldsymbol{\Sigma}$. The Cox distribution $Q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is simply a Bernoulli distribution $\mathcal{B}(\boldsymbol{\mu})$ with $\boldsymbol{\mu} = \mathbf{h}$, and the fixed point equations impose that

$$\mu_i = \tanh^{-1}(m_i),$$
$$d_i = 1 - m_i^2.$$

We now seek a Taylor expansion for the solution of eq. (D1)-(D5) when α is small but nonzero, and magne-

tizations m_i are fixed. More precisely, noting

$$\Delta_i := \mu_i - \tanh^{-1}(m_i),$$

our purpose is to find the parameters in the following Taylor expansions :

$$\Delta_i = \alpha \Delta_i^{[1]} + \alpha^2 \Delta_i^{[2]} + \alpha^3 \Delta_i^{[3]} + \dots$$
$$\Sigma_{ij} = \alpha \Sigma_{ij}^{[1]} + \alpha^2 \Sigma_{ij}^{[2]} + \alpha^3 \Sigma_{ij}^{[3]} + \dots$$

When inserted into eq. (D4), the development of Δ_i will exactly provide the desired Plefka expansion.

Development for equation (D1)

Equation (D1) writes

$$m_i = \int_{x \in \mathbb{R}} \tanh\left(\tanh^{-1}(m_i) + \Delta_i + x\sqrt{\Sigma_{ii}}\right) \phi(x)\mathrm{d}x,$$

where Δ_i is of leading order α, and $\sqrt{\Sigma_{ii}}$ is of leading order $\alpha^{1/2}$. Applying the Taylor development of tanh :

$$\tanh\left(\tanh^{-1}(m) + X\right) = m + (1 - m^2)\left[X - mX^2 + \dots\right]$$

up to order 6 (because $\sqrt{\Sigma_{ii}}$ is of leading order $\alpha^{1/2}$), and using the classic integration formulas :

$$\int_x x^n \phi(x)\mathrm{d}x = \begin{cases} (n-1)(n-3)\dots & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases}$$
(D6)

we obtain

$$\begin{aligned} m_i = & m_i + (1 - m_i^2)\Big[\Delta_i - m_i\big(\Delta_i^2 + \Sigma_{ii}\big) \\ & + \big(m_i^2 - \tfrac{1}{3}\big)\big(\Delta_i^3 + 3\Delta_i\Sigma_{ii}\big) \\ & + \big(-m_i^3 + \tfrac{2}{3}m_i\big)\big(\Delta_i^4 + 6\Delta_i^2\Sigma_{ii} + 3\Sigma_{ii}^2\big) \\ & + \big(m_i^4 - m_i^2 + \tfrac{2}{15}\big)\big(10\Delta_i^3\Sigma_{ii} + 15\Delta_i\Sigma_{ii}^2\big) \\ & + \big(-m_i^5 + \tfrac{4}{3}m_i^3 - \tfrac{17}{45}m_i\big)\big(45\Delta_i^2\Sigma_{ii}^2 + 15\Sigma_{ii}^3\big)\Big] + o(\alpha^3). \end{aligned}$$

Notice that, after integration by the Gaussian kernel, only integer powers of Σ_{ii} remain.

For this equation to be verified, the term inside square brackets must be equal to zero up to order α^3. Expanding Δ_i and Σ_{ii} with the shorthand $X_k = \Delta_i^{[k]}$, $Y_k = \Sigma_{ii}^{[k]}$, and regrouping the powers of α, we obtain :

$$0 = \alpha\Big[X_1 - m_i Y_1\Big] \tag{D7}$$

$$+ \alpha^2\Big[X_2 - m_i(X_1^2 + Y_2) + (3m_i^2 - 1)X_1Y_1 + (-3m_i^3 + 2m_i)Y_1^2\Big] \tag{D8}$$

$$+ \alpha^3\Big[X_3 - m_i(2X_2X_1 + Y_3) + (m_i^2 - \tfrac{1}{3})(X_1^3 + 3X_2Y_1 + 3X_1Y_2) + (-6m_i^2 + 4m_i)(X_1^2Y_1 + Y_2Y_1)$$

$$+ (15m_i^4 - 15m_i^2 + 2)X_1Y_1^2 + (-15m_i^5 + 20m_i^3 - \tfrac{17}{3}m_i)Y_1^3\Big] + o(\alpha^3). \tag{D9}$$

Solution at order 2

At this point, we can readily find the two first orders of the solution. Indeed, we have $\boldsymbol{\Sigma} = \alpha\mathbf{J} + \alpha^2\mathbf{JCJ}$ and $C_{ij} = (1 - m_i^2)\delta_{ij} + o(1)$, and so

$$\Sigma_{ij}^{[1]} = J_{ij},$$
$$\Sigma_{ij}^{[2]} = \sum_k J_{ik}J_{jk}(1 - m_k^2).$$

Then, at order 1, the fixed point equation above (line (D7)) imposes that

$$\Delta_i^{[1]} = m_i\Sigma_{ii}^{[1]} \quad = m_i J_{ii}.$$

Using this new value, the fixed point equation at order 2 (line (D8)) imposes that

$$\Delta_i^{[2]} = m_i\big((\Delta_i^{[1]})^2 + \Sigma_{ii}^{[2]}\big) - (3m_i^2 - 1)\Delta_i^{[1]}\Sigma_{ii}^{[1]} + \dots$$
$$= m_i \sum_{j \neq i} J_{ij}^2(1 - m_j^2),$$

which is identical to the order 2 coefficient in the exact Ising model – see eq. (29). Note that the diagonal terms J_{ii} are nonzero and an active part of the derivation, but cancel out in the final result, so they play no role in the expansion up to order 2.

Development for C

In general, to establish the development of $\boldsymbol{\Sigma}$ at any given order n, we need the expansion of $\mathbf{C}$ up to order $n-2$, because $\boldsymbol{\Sigma} = \alpha\mathbf{J} + \alpha^2\mathbf{JCJ}$. Thus, to expand $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ at order 3, we must first establish the development for $\mathbf{C}$ at order 1, based on the development of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ at order 1 established above.

Regarding eq. (D2), a very similar computation (development of tanh at order 2, and simplification by the integral formulas of eq. (D6)) yields :

$$d_i = (1 - m_i^2) - \alpha(1 - m_i^2)^2 J_{ii} + o(\alpha).$$

Introducing matrix $\mathbf{D} = \mathrm{diag}(d_i)$, eq. (D3) writes

$$\mathbf{C}^{-1} = \mathbf{D}^{-1} - (\alpha\mathbf{J}),$$

and thus, after a classic switching from $\mathbf{C}^{-1}$ to $\mathbf{C}$:

$$\mathbf{C} = \mathbf{D} + \alpha\mathbf{DJD} + o(\alpha).$$

So finally :

$$C_{ij} = \delta_{ij}(1 - m_i^2)$$
$$+ \alpha(1 - \delta_{ij})(1 - m_i^2)(1 - m_j^2)J_{ij} + o(\alpha). \quad \text{(D10)}$$

Solution at orders 3 and 4

The coefficient $C_{ij}^{[1]}$ found in eq. (D10) is pasted into $\boldsymbol{\Sigma} = \alpha\mathbf{J} + \alpha^2\mathbf{JCJ}$, to obtain :

$$\Sigma_{ij}^{[3]} = \sum_{k \neq l} J_{ik}J_{kl}J_{lj}(1 - m_k^2)(1 - m_l^2). \quad \text{(D11)}$$

Then, line (D9) allows to find the order 3 coefficient of $\boldsymbol{\mu}$. After computation, this gives :

$$\Delta_i^{[3]} = 2m_i \sum_{(jk|i)} J_{ij}J_{jk}J_{ki}(1 - m_j^2)(1 - m_k^2)$$
$$- 2(m_i^2 - \tfrac{1}{3})J_{ii}^3 m_i(1 - m_i^2),$$

where $(jk|i)$ denotes all unordered triplets of the form $\{i, j, k\}$ with j and k distinct, and distinct from i.

Using the values found for $\Delta_i^{[1]}$, $\Delta_i^{[2]}$ and $\Delta_i^{[3]}$, and the fact that $\mu_i = h_i + \alpha \sum_j J_{ij}m_j$, yields eq. (31) from the main text.

Pushing all computations one order further, with the help of the computer algebra system MAXIMA, yields :

$$\Delta_i^{[4]} = 2m_i \sum_{(jkl|i)} J_{ij}J_{jk}J_{kl}J_{li}c_jc_kc_l - m_ic_i \sum_{j \neq i} J_{ij}^4 c_j^2$$
$$- 2m_ic_i(3m_i^2 - 1)J_{ii}^2 \sum_{j \neq i} J_{ij}^2 c_j$$
$$- 2m_ic_i(7m_i^4 - 8m_i^2 + \tfrac{5}{3})J_{ii}^4$$
$$- 2m_i \sum_{j \neq i} J_{ij}^2 J_{jj}^2 m_j^2 c_j^2, \quad \text{(D12)}$$

with the shorthand $c_i = 1 - m_i^2$. The two first terms are identical to the adaptive TAP expansion, eq. (C7). The remaining terms involve the diagonal weights J_{ii}, and would be absent if the coupling matrix was such that $\mathrm{diag}(\mathbf{J}) = \mathbf{0}$.

Appendix E: Approximate formulas for the Cox distribution

The formulas inherent to the Cox distribution, eq. (13), (16), (17) and (21) from the main text, are easily estimated by numerical integration (for example, Simpson quadrature). But the overall computation time quickly becomes forbidding, as these estimations must be done for each pair of spins, and on many iterations to target the fixed point.

Hence, I found it more convenient to use approximate formulas. Let us note L the logistic function at scale $1/2$, that is :

$$L(r) := \frac{1}{1 + e^{-2r}}.$$

Function L is pivotal in the Bernoulli distribution, since $\tanh(r) = 2L(r) - 1$, $\log 2\cosh(r) = 2\int_0^r L(u)\mathrm{d}u - r$, and $1 - \tanh^2(r) = 2L'(r)$.

I suggest to approximate L by the following combination of Gaussian functions :

$$L^{\text{app}}(r) = \Phi\left(\frac{r}{\kappa}\right) + \nu\phi'\left(\frac{r}{\lambda}\right), \qquad \text{(E1)}$$

with $\Phi(x) := \int_{-\infty}^{x} \phi(u)\mathrm{d}u$ the standard normal cumulative distribution, and $\phi'(r) = -r\phi(r)$.

Taking parameters $(\kappa, \nu, \lambda) \simeq (0.7072, 0.1648, 0.9712)$, one has $\|L - L^{\text{app}}\|_{\infty} < 0.001$ on the whole real line. The approximation also applies to the primitive, with $\|\int(L - L^{\text{app}})\|_{\infty} < 0.001$, and to the first derivative, with $\|L' - (L^{\text{app}})'\|_{\infty} < 0.0033$.

As the convolution product of two Gaussian functions remains Gaussian, this replacement allows to compute analytically all the formulas. Here, I only provide the results, and refer to Supplementary Material for the derivation.

Given spin index i, let us introduce the following reduced quantities :

$$x_i := \frac{\mu_i}{\sqrt{\Sigma_{ii}}},$$

$$k_i := \frac{\sqrt{\Sigma_{ii}}}{\sqrt{\kappa^2 + \Sigma_{ii}}},$$

$$l_i := \frac{\sqrt{\Sigma_{ii}}}{\sqrt{\lambda^2 + \Sigma_{ii}}}.$$

Then, the first moment of the Cox distribution, eq. (16) (or equivalently eq. (20)), can be approximated as

$$m_i^{\text{app}} = 2\left[\Phi(x_i k_i) + \nu(1 - l_i^2)\phi'(x_i l_i)\right] - 1 \qquad \text{(E2)}$$

with a guaranteed maximum error $|m_i - m_i^{\text{app}}| < 0.002$.

The variance term d_i in eq. (17) is approximated as

$$d_i^{\text{app}} = 2\Sigma_{ii}^{-1/2}\left[k_i\phi(x_i k_i) + \nu(1 - l_i^2)l_i\phi''(x_i l_i)\right]$$

with a guaranteed maximum error $|d_i - d_i^{\text{app}}| < 0.007$.

To concretely estimate the free energy associated to eq. (13), it is necessary to compute $F_i := \int_x \log 2\cosh(\mu_i + x\sqrt{\Sigma_{ii}})\phi(x)\mathrm{d}x$. It is approximated as

$$F_i^{\text{app}} = \mu_i\left[2\Phi(x_i k_i) - 1\right] + 2\sqrt{\Sigma_{ii}}\left[\frac{\phi(x_i k_i)}{k_i} - \nu\frac{\phi''(x_i l_i)}{l_i}\right]$$

with a guaranteed maximum error inferior to 0.002.

The approximate formula for the covariance of the Cox distribution, eq. (21), is quite bulky and provided in Supplementary Material. It has guaranteed maximum error inferior to 0.008.

In my numerical tests, using these approximate formulas instead of lengthier Simpson quadrature yielded no noticeable difference in the final solution of the fixed point equations. At the same time, computation times were cut by (up to) two orders of magnitude.

---

[1] D. R. Cox, Applied statistics , 113 (1972).
[2] L. P. Zhao and R. L. Prentice, Biometrika **77**, 642 (1990).
[3] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, in *Readings in Computer Vision* (Elsevier, 1987) pp. 522–533.
[4] M. Mézard, G. Parisi, and M. Virasoro, *Spin glass theory and beyond*, Vol. 9 (World Scientific Publishing Company, 1987).
[5] M. Opper and D. Saad, eds., *Advanced mean field methods : Theory and practice* (MIT Press, 2001).
[6] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction*, Vol. 111 (Clarendon Press, 2001).
[7] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proceedings of the National Academy of Sciences **106**, 67 (2009).
[8] E. Schneidman, M. J. Berry II, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).
[9] I. E. Ohiorhenuan, F. Mechler, K. P. Purpura, A. M. Schmid, Q. Hu, and J. D. Victor, Nature **466**, 617 (2010).
[10] K. Hukushima and K. Nemoto, Journal of the Physical Society of Japan **65**, 1604 (1996).
[11] F. Wang and D. P. Landau, Physical review letters **86**, 2050 (2001).
[12] H. J. Kappen and F. d. B. Rodríguez, Neural Computation **10**, 1137 (1998).
[13] V. Sessak and R. Monasson, Journal of Physics A: Mathematical and Theoretical **42**, 055001 (2009).
[14] Y. Roudi, J. Tyrcha, and J. Hertz, Physical Review E **79**, 051915 (2009).
[15] S. Cocco and R. Monasson, Physical review letters **106**, 090601 (2011).
[16] K. Pearson, Biometrika **7**, 96 (1909).
[17] D. R. Cox and N. Wermuth, Biometrika **89**, 462 (2002).
[18] S.-i. Amari, H. Nakahara, S. Wu, and Y. Sakai, Neural computation **15**, 127 (2003).
[19] J. H. Macke, M. Opper, and M. Bethge, Physical Review Letters **106**, 208102 (2011).
[20] D. R. Cox, Journal of the Royal Statistical Society. Series B (Methodological) , 215 (1958).
[21] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes, volume I : Elementary theory and methods* (Springer Science & Business Media, 2003).
[22] D. R. Cox, Journal of the Royal Statistical Society. Series B (Methodological) , 129 (1955).
[23] J. Møller, A. R. Syversveen, and R. P. Waagepetersen, Scandinavian journal of statistics **25**, 451 (1998).
[24] P. J. Diggle, P. Moraga, B. Rowlingson, and B. M. Taylor, Statistical Science , 542 (2013).
[25] M. Krumin and S. Shoham, Neural computation **21**, 1642 (2009).
[26] R. Brette, Neural computation **21**, 188 (2009).
[27] A. N. Vasil'ev and R. Radzhabov, Theoretical and Mathematical Physics **21**, 963 (1974).
[28] C. M. Bishop, *Pattern recognition and machine learning* (Springer Verlag, New York, USA, 2006).
[29] S. Boyd and L. Vandenberghe, *Convex optimization*

(Cambridge university press, 2004).

[30] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Philosophical Magazine **35**, 593 (1977).

[31] J. S. Yedidia, W. T. Freeman, and Y. Weiss, in *Advances in neural information processing systems* (2001) pp. 689–695.

[32] M. Mézard and G. Parisi, The European Physical Journal B-Condensed Matter and Complex Systems **20**, 217 (2001).

[33] M. Opper and O. Winther, Physical Review Letters **86**, 3695 (2001).

[34] M. Opper and O. Winther, Physical Review E **64**, 056131 (2001).

[35] T. Plefka, Journal of Physics A: Mathematical and general **15**, 1971 (1982).

[36] F. Ricci-Tersenghi, Journal of Statistical Mechanics: Theory and Experiment **2012**, P08015 (2012).

[37] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett **35**, 1792 (1975).

[38] J. J. Hopfield, Proceedings of the national academy of sciences **79**, 2554 (1982).

[39] A. Decelle and F. Ricci-Tersenghi, Physical Review E **94**, 012112 (2016).

[40] A. Georges and J. S. Yedidia, Journal of Physics A: Mathematical and General **24**, 2173 (1991).

[41] K. Nakanishi and H. Takayama, Journal of Physics A: Mathematical and General **30**, 8085 (1997).

[42] T. Tanaka, Physical Review E **58**, 2302 (1998).