Message Passing for Graphical Models

Hai-Jun Zhou*
Institute of Theoretical Physics, Chinese Academy of Sciences,
Beijing 100190

June 29, 2017

Abstract

This is the notes of a tutorial lecture given by the author at the Kavli Institute for Theoretical Sciences on the occasion of the "Workshop on Machine Learning and Many-Body Physics" (June 28–30, 2017, Beijing). The concept of probabilistic graphical models and the factor-graph representation are introduced and several well-known examples are mentioned. Message-passing methods for approximately solving a graphical model are introduced, especially the naive mean field method (based on the Weiss approximation) and the belief-propagation method (based on the Bethe-Peierls approximation). Application of message-passing methods in combinatorial optimization and statistical inference is illustrated by two examples. Some important advanced topics not covered in the lecture are also listed.

Contents

*Email: zhouhj@itp.ac.cn

This is the notes of a tutorial lecture given by the author at the "Workshop on Machine Learning and Many-Body Physics" (Kavli Institute for Theoretical Sciences, June 28–30, 2017, Beijing). With the conservative *a priori* assumption that the audience has little background knowledge on probabilistic graphical models and message-passing procedures, the author has tried to make the lecture and the notes as self-contained as possible, and has restricted all the discussions at the elementary level.

1 Probabilistic graphical models

Probabilistic graphical models are widely encountered in statistical physics and in computer science. Here we introduce the factor-graph representation for this type of many-body interaction systems, and discuss the partition function and free energy of a general probabilistic graphical model. We begin with some simple examples.

1.1 Some examples of energy functions

The *(generalized) Ising model* is defined on a graph of vertices and edges and it has the following energy function

$$E(\sigma_1, \sigma_2, \ldots, \sigma_N) = -\sum_{(i,j)\in G} J_{ij}\sigma_i\sigma_j - \sum_{k=1}^{N} h_k^0 \sigma_k \ . \tag{1}$$

Here J_{ij} denotes the spin coupling constant on an edge (i, j) of the graph between two vertices i and j, and h_k^0 is the external (magnetic) field on vertex k. The state σ_i of vertex $i \in \{1, 2, \ldots, N\}$ is assumed to be binary so $\sigma_i \in \pm 1$. If all the couplings J_{ij} are non-negative the model is the ferromagnetic Ising model. If J_{ij} has roughly equal chances of being positive and negative, the model is then a spin glass system.

A representative many-body interaction model is the *exclusive-or (XOR) satisfiability problem* involving N binary spins $\sigma_i \in \pm 1$, with energy function

$$E(\sigma_1, \sigma_2, \ldots, \sigma_N) = \sum_{a=1}^{M} \frac{1}{2}\Big(1 - J_a \prod_{i\in\partial a} \sigma_i\Big) \ , \tag{2}$$

where $J_a \in \pm 1$ is the fixed coupling constant of interaction a, which involves a set (denoted as ∂a) of vertices. Notice the energy of an interaction a is zero if $\prod_{i\in\partial a} \sigma_i$ has the same sign as J_a; otherwise the interaction is unity. This model is very important in low-density parity-check (LDPC) coding theory.

The energy function of the *restricted Boltzmann machine (RBM)* is similar to Eq. (1) but it involves two sets of spins:

$$E(\sigma_1, \ldots, \sigma_N; s_1, \ldots, s_{N'}) = -\sum_{i=1}^{N} h_i^0 \sigma_i - \sum_{i=1}^{N}\sum_{\mu=1}^{N'} J_{i\mu}\sigma_i s_\mu - \sum_{\mu=1}^{N'} w_\mu^0 s_\mu \ . \tag{3}$$

Here $\vec{\sigma} \equiv (\sigma_1, \sigma_2, \ldots, \sigma_N)$ is a configuration of N visible spins $\sigma_i \in \pm 1$; $\vec{s} \equiv (s_1, s_2, \ldots, s_{N'})$ is a configuration of N' hidden spins $s_\mu \in \pm 1$; h_i^0 and w_μ^0 are, respectively, the external field (also called bias) on visible vertex i and hidden vertex μ; and $J_{i\mu}$ is the coupling constant between the visible vertex i and hidden vertex μ. The RBM is widely adopted in machine learning tasks.

The *Amari-Hopfield model* is a fundamental neural network model for associative memory (Amari, 1977; Hopfield, 1982). Its energy function is

$$E(\sigma_1, \sigma_2, \ldots, \sigma_N) = \sum_{a=1}^{M} \Big(\frac{1}{N}\sum_{i=1}^{N} \xi_i^a \sigma_i\Big)^p \ , \tag{4}$$

where $\vec{\xi}^a \equiv (\xi_1^a, \xi_2^a, \ldots, \xi_N^a)$ with $a = 1, 2, \ldots, M$ denotes one of the M memorized patterns. The energy coefficient p is often chosen to be $p = 2$ for theoretical convenience, but a value of $p \geq 3$ is desirable in terms of memory capacity.

1.2 The factor-graph representation

The energy functions of the preceding subsection have the following general form

$$E(\vec{\sigma}) = \sum_{i=1}^{N} E_i(\sigma_i) + \sum_{a=1}^{M} E_a(\vec{\sigma}_{\partial a}) \ . \tag{5}$$

(Here and in all the following discussions we use letters $i, j, k, \ldots$ to denote a generic vertex and use letters $a, b, c, \ldots$ to denote a generic interaction.) The state vector $\vec{\sigma} = (\sigma_1, \ldots, \sigma_N)$ denotes a generic configuration of the system, σ_i could be discrete or be continuous-valued; the function $E_i(\sigma_i)$ is the vertex energy which depends only on the state of a single vertex i; and $E_a(\vec{\sigma}_{\partial a})$ is the energy of interaction a which depends on the states of all the vertices in the set ∂a, with $\vec{\sigma}_{\partial a} \equiv \{\sigma_i : i \in \partial a\}$. The equilibrium Boltzmann distribution of the model (5) is

$$B(\vec{\sigma}) = \frac{1}{Z(\beta)} e^{-\beta E(\vec{\sigma})} \ , \tag{6}$$

where $\beta = 1/T$ is the inverse temperature and T is the temperature. The normalization constant $Z(\beta)$ is referred to as the partition function, and it is a weighted sum over all the microscopic configurations

$$Z(\beta) = \sum_{\vec{\sigma}} \exp\bigl(-\beta E(\vec{\sigma})\bigr) = \sum_{\vec{\sigma}} \prod_{i=1}^{N} \psi_i(\sigma_i) \prod_{a=1}^{M} \psi_a(\vec{\sigma}_{\partial a}) \ , \tag{7}$$

where the vertex and interaction Boltzmann factors are, respectively,

$$\psi_i(\sigma_i) \equiv e^{-\beta E_i(\sigma_i)} \ , \tag{8}$$

$$\psi_a(\vec{\sigma}_{\partial a}) \equiv e^{-\beta E_a(\vec{\sigma}_{\partial a})} \ . \tag{9}$$

We can represent the generic model (5) by a bipartite graph of N circles (each of which corresponding to a vertex i of the system), and M squares (each of which corresponding to an interaction a), see the illustration in Fig. 1. If a vertex j participates in an interaction b (so $j \in \partial b$), a link (j, b) is then set up between the corresponding circle and square. This bipartite graph is referred to as a factor graph (Kschischang et al., 2001)), and it is denoted as G in the following discussions.

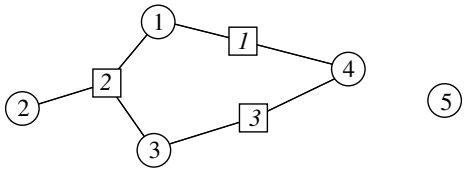


Figure 1: Factor-graph representation for a simple probabilistic graphical model, which has $N = 5$ circles (vertices) and $M = 3$ squares (interactions). The vertex with index $i = 5$ does not participate in any interactions so it is isolated. The vertex $i = 1$ participates in two interactions (whose indices are $a = 1$ and $a = 2$) so it has two attached links. Figure copied from Zhou (2015).

The energy function (5), represented as a factor graph G, together with the partition function (7) define a probabilistic graphical model. The biggest challenge for probabilistic graphical models is to compute the partition function $Z(\beta)$ or equivalently the free energy $F(\beta)$ defined by

$$F(\beta) \equiv -\frac{1}{\beta} \ln Z(\beta) \ . \tag{10}$$

For the special case of the factor graph being a tree (or a forest) which is free of any loops, the free energy can be computed with arbitrary precision by recursively integrating the states of the leaf vertices. But the factor graphs of most non-trivial many-body systems are extremely rich in loops. The existence of a huge number of loops cause complicated correlations among the states of different vertices. It is then very hard to compute the free energy with very high precision. We review in the following three sections some intuitive message-passing methods to treat such loop-rich statistical physics systems approximately but efficiently.

2 The cavity picture

The cavity picture was first developed in studying the Sherrington-Krikpatrick spin glass model (Mézard et al., 1987) and was later employed in deriving the first-step replica-symmetric breaking mean field theory of finite-connectivity spin glasses (Mézard and Parisi, 2001). In this section we employ the cavity picture to reach an intuitive formula for the free energy $F(\beta)$.

2.1 Single-vertex marginal probability

Let us first focus on the behavior of a single vertex, say vertex i. This vertex might be involved in some interactions and this set of interactions is denoted as ∂i (e.g., $\partial i = \{a, b, c\}$ in Fig. 2a). The marginal distribution of vertex i, $q_i(\sigma_i)$, is expressed as

$$\begin{aligned} q_i(\sigma_i) &= \frac{1}{Z(\beta)} \sum_{\vec{\sigma}\backslash\sigma_i} e^{-\beta E(\vec{\sigma})} \\ &\propto \psi_i(\sigma_i) \sum_{\vec{\sigma}\backslash\sigma_i} \prod_{a\in\partial i} \psi_a(\vec{\sigma}_{\partial a}) \Big\{ \prod_{j\neq i} \psi_j(\sigma_j) \prod_{b\notin\partial i} \psi_b(\vec{\sigma}_{\partial b}) \Big\} . \end{aligned} \tag{11}$$

Notice the terms inside the curly brakets of Eq. (11) do not depend on the state σ_i of vertex i, they are contributed by the interactions and site energies of the "cavity" factor graph, $G_{\backslash i}$, formed by the $(N-1)$ vertices except i (Fig. 2b). Vertex i is only directly affected by the vertices at the "inner boundary" of this cavity graph, and we denote this set of vertices as $n(i)$, i.e., $n(i) \equiv \{j : j \in \partial a\backslash i; a \in \partial i\}$. For the particular example of Fig. 2 the vertex set $n(i) = \{j, k, l, m, n, o, p\}$.

Let us denote a generic configuration of the cavity system $G_{\backslash i}$ as $\vec{\sigma}_{\backslash i}$ ($\equiv \{\sigma_j : j \in \{1, 2, \ldots, N\}\backslash i\}$). The partition function of this cavity graph is

$$Z_{\backslash i}(\beta) = \sum_{\vec{\sigma}_{\backslash i}} \prod_{j\neq i} \psi_j(\sigma_j) \prod_{b\notin\partial i} \psi_b(\vec{\sigma}_{\partial b}) . \tag{12}$$

Then the Boltzmann distribution $B_{\backslash i}(\vec{\sigma}_{\backslash i})$ for the cavity system is

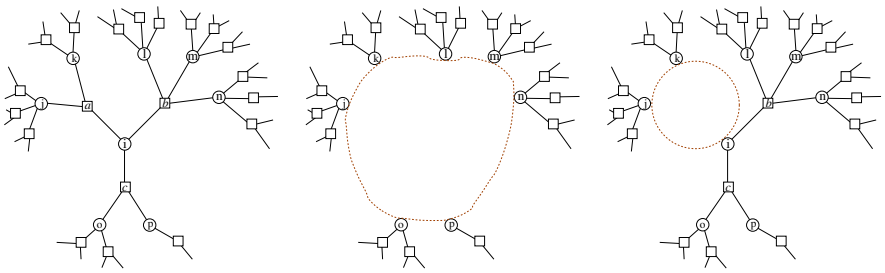
$$B_{\backslash i}(\vec{\sigma}_{\backslash i}) = \frac{1}{Z_{\backslash i}(\beta)} \prod_{j\neq i} \psi_j(\sigma_j) \prod_{b\notin\partial i} \psi_b(\vec{\sigma}_{\partial b}) . \tag{13}$$

The boundary vertices (those in the set $n(i)$) of the cavity graph $G_{\backslash i}$ then has the following joint distribution

$$P^{cavity}_{n(i)}(\vec{\sigma}_{n(i)}) = \sum_{\vec{\sigma}_{\backslash i}\backslash\vec{\sigma}_{n(i)}} B_{\backslash i}(\vec{\sigma}_{\backslash i}) . \tag{14}$$

Then we see from Eq. (11) that

$$q_i(\sigma_i) = \frac{1}{z_i} \psi_i(\sigma_i) \sum_{\vec{\sigma}_{n(i)}} P^{cavity}_{n(i)}(\vec{\sigma}_{n(i)}) \prod_{a\in\partial i} \psi_a(\vec{\sigma}_a) , \tag{15}$$



(a) (b) (c)

Figure 2: Creating cavities in the factor graph G. (a) The local environment of a vertex i which participates in three interactions ($\partial i = \{a, b, c\}$). (b) After vertex i and the attached interactions are deleted, the remaining part of the factor graph is a cavity graph, $G_{\backslash i}$. The states of vertices j and k in $G_{\backslash i}$ might still be correlated due to the interactions not shown here; however due to the absence of interaction a, these two vertices should be less correlated in $G_{\backslash i}$ than in the original system G. (c) The cavity graph $G_{\backslash a}$ obtained by deleting a single interaction a from the original factor graph.

where z_i is the normalization constant defined by

$$z_i = \sum_{\sigma_i} \psi_i(\sigma_i) \sum_{\vec{\sigma}_{n(i)}} P^{cavity}_{n(i)}(\vec{\sigma}_{n(i)}) \prod_{a \in \partial i} \psi_a(\vec{\sigma}_{\partial a}) \; . \tag{16}$$

Equation (15) has a clear physical meaning. The marginal probability of σ_i depends on the probability of the states $\vec{\sigma}_{n(i)}$ of boundary vertices in the cavity system $G_{\backslash i}$ (*before* vertex i is added into the system), and also on the extra Boltzmann weight $\prod_{a \in \partial i} \psi_a(\vec{\sigma}_{\partial a})$ due to the interactions between i and the cavity boundary, and finally on the additional Boltzmann weight $\psi_i(\sigma_i)$ coming from the site energy of vertex i.

It is helpful to emphasize again that, to exactly compute $q_i(\sigma_i)$ we need to know the joint probability $P^{cavity}_{n(i)}(\vec{\sigma}_{n(i)})$ in the cavity system $G_{\backslash i}$ instead of the full system G.

2.2 Contribution of a vertex and its attached interactions

There is an important relation between the partition functions of the full system G and the cavity system $G_{\backslash i}$:

$$Z(\beta) = z_i(\beta) Z_{\backslash i}(\beta) \; . \tag{17}$$

This is easy to check since $Z(\beta)$ can be expressed as

$$\frac{\sum_{\sigma_i} \sum_{\vec{\sigma}_{\backslash i}} \psi_i(\sigma_i) \prod_{a \in \partial i} \psi_a(\vec{\sigma}_{\partial a}) \prod_{j \neq i} \psi_j(\sigma_j) \prod_{b \notin \partial i} \psi_b(\vec{\sigma}_{\partial b})}{\sum_{\vec{\sigma}_{\backslash i}} \prod_{j \neq i} \psi_j(\sigma_j) \prod_{b \notin \partial i} \psi_b(\vec{\sigma}_{\partial b})} \sum_{\vec{\sigma}_{\backslash i}} \prod_{j \neq i} \psi_j(\sigma_j) \prod_{b \notin \partial i} \psi_b(\vec{\sigma}_{\partial b}) \; . \tag{18}$$

In terms of free energy, we see from Eq. (17) that

$$F(\beta) = f_i(\beta) + F_{\backslash i}(\beta) \; , \tag{19}$$

where $F_{\backslash i}(\beta) \equiv -(1/\beta) \ln Z_{\backslash i}(\beta)$ and

$$f_i(\beta) \equiv -\frac{1}{\beta} \ln z_i(\beta) \tag{20}$$

$$= -\frac{1}{\beta} \ln \Big[\sum_{\sigma_i} \psi_i(\sigma_i) \sum_{\vec{\sigma}_{n(i)}} P^{cavity}_{n(i)}(\vec{\sigma}_{n(i)}) \prod_{a \in \partial i} \psi_a(\vec{\sigma}_{\partial a}) \Big] \; . \tag{21}$$

The total free energy $F(\beta)$ of the full system G is therefore decomposed into two parts, the free energy $F_{\backslash i}(\beta)$ of the cavity system $G_{\backslash i}$ and the free energy contribution $f_i(\beta)$ from vertex i and all its attached interactions. Notice that $F_{\backslash i}(\beta)$ is completely independent of vertex i and all the interactions in the set ∂i.

2.3 Contribution of a single interaction

We can also consider the free energy contribution of a single interaction a which directly affect a set ∂a of vertices. This contribution can be obtained by comparing the original factor graph G and the cavity graph (denoted as $G_{\backslash a}$) obtained by deleting the interaction a (Fig. 2c). Similar to Eq. (17), we obtain

$$Z(\beta) = z_a(\beta) Z_{\backslash a}(\beta) \ , \tag{22}$$

where $Z_{\backslash a}$ is the partition function of the cavity sytem $G_{\backslash a}$. The factor $z_a(\beta)$ is expressed as

$$z_a(\beta) = \sum_{\vec{\sigma}_{\partial a}} \psi_a(\vec{\sigma}_{\partial a}) P_a^{cavity}(\vec{\sigma}_{\partial a}) \ , \tag{23}$$

with $P_a^{cavity}(\vec{\sigma}_{\partial a})$ being the joint state distribution of all vertices in the set ∂a (the inner boundary vertices, see Fig. 2c) in the cavity system $G_{\backslash a}$.

In terms of free energy we therefore have

$$F(\beta) = f_a(\beta) + F_{\backslash a}(\beta) \ . \tag{24}$$

This relation is very similar to Eq. (19). It means that the total free energy of the full system G is equal to the free energy $F_{\backslash a}$ of the cavity system $G_{\backslash a}$ plus the additional free energy contribution $f_a(\beta)$ from the interaction a. The explicit expression for $f_a(\beta)$ is

$$f_a(\beta) \equiv -\frac{1}{\beta} \ln z_a(\beta) = -\frac{1}{\beta} \ln \Big[\sum_{\vec{\sigma}_{\partial a}} \psi_a(\vec{\sigma}_{\partial a}) P_a^{cavity}(\vec{\sigma}_{\partial a}) \Big] \ . \tag{25}$$

Notice again that $F_{\backslash a}(\beta)$ is completely independent of the interaction a, simply because this interaction is absent in the cavity system $G_{\backslash a}$ (Fig. 2c).

2.4 An intuitive expression for the total free energy

Looking at Eq. (19), it is tempting for us to write the total free energy as $F(\beta) = \sum_{i=1}^{N} f_i(\beta)$. But the free energy $f_i(\beta)$ contains not only the contribution of vertex i but the contributions from all the attached interactions in the set ∂i. Therefore there is over-counting of free energy contributions from the interactions in $\sum_i f_i$, which must be properly eliminated. For example, the interaction a in Fig. 2a involves three vertices i, j, and k; its free energy contribution $f_a(\beta)$ is considered three times, respectively, in the free energies f_i, f_j, and f_k; it is intuitively reasonable to subtract a value equal to $2f_a$ from the sum $\sum_i f_i$.

With this analysis, we arrive at the following explicit formula for $F(\beta)$:

$$F(\beta) \approx \sum_{i=1}^{N} f_i(\beta) - \sum_{a=1}^{M} \big(|\partial a| - 1 \big) f_a(\beta) \ , \tag{26}$$

where $|\partial a|$ means the size of the vertex set ∂a, i.e., the number of vertices participating in interaction a. If the factor graph G is a tree or is a forest (a collection of trees), then there is no loop in G. It can be proven that Eq. (26) is exact for such a loop-free system (Zhou, 2015). Usually the factor graph G contains a huge number of loops, and then the expression (26) is only an approximation to the true free energy.

To actually compute the free energy using Eq. (26), we need first to compute the cavity joint probabilities $P_{n(i)}^{cavity}(\vec{\sigma}_{n(i)})$ and $P_a^{cavity}(\vec{\sigma}_{\partial a})$. These tasks are not easy and we have to rely on approximate methods. We now describe two of the simplest approximate (mean-field) solution protocols.

3 The Weiss mean field theory

Let us make two simplifying approximations: (1) the inner boundary vertices in the cavity graphs $G_{\backslash a}$ (see Fig. 2c) are all independent; and (2) each of these boundary vertices $i \in \partial a$ has the same marginal distribution $q_i(\sigma_i)$ in the cavity system $G_{\backslash a}$ as in the full system G. Then the joint distribution $P_a^{cavity}(\vec{\sigma}_{\partial a})$ will become factorized, namely

$$P_a^{cavity}(\vec{\sigma}_{\partial a}) \approx \prod_{j \in \partial a} q_j(\sigma_j) \; . \tag{27}$$

Similarly, we assume that all the inner boundary vertices in the cavity graph $G_{\backslash i}$ are independent and their marginal distributions are the same as in the full graph G. Then

$$P_{n(i)}^{cavity}(\vec{\sigma}_{n(i)}) \approx \prod_{a \in \partial i} \prod_{j \in \partial a \backslash i} q_j(\sigma_j) \; . \tag{28}$$

Inserting the approximate expression (28) into Eq. (15), we get the following self-consistent equation for the single-vertex marginal probability $q_i(\sigma)$:

$$q_i(\sigma_i) \propto \psi_i(\sigma_i) \prod_{a \in \partial i} \Big[\sum_{\vec{\sigma}_{\partial a \backslash \sigma_i}} \psi_a(\vec{\sigma}_{\partial a}) \prod_{j \in \partial a \backslash i} q_j(\sigma_j) \Big] \; . \tag{29}$$

For example, in the case of an interaction a involving only two vertices i and j, with energy $E_a = -J_{ij}\sigma_i\sigma_j$, the corresponding term within the square brackets of this equation is equal to $\sum_{\sigma_j} q_j(\sigma_j) e^{\beta J_{ij}\sigma_i\sigma_j}$; in the case of a three-body interaction b involving vertices i, j, and k with $E_b = -J_{ijk}\sigma_i\sigma_j\sigma_k$, the corresponding term within the square brackets is equal to $\sum_{\sigma_j} \sum_{\sigma_k} q_j(\sigma_j) q_k(\sigma_k) e^{\beta J_{ijk}\sigma_i\sigma_j\sigma_k}$.

Equation (29) is the simplest message-passing equation for solving a probabilistic graphical model. In the statistical physics literature it is widely referred to as the Weiss mean field equation. One can try to iterate this equation on all the vertices of the graph to reach a fixed-point solution. At a fixed point of Eq. (29) is reached, the free energy contributions from all the vertices and all the interactions can be evaluated as:

$$f_i(\beta) = -\frac{1}{\beta} \ln \Big[\sum_{\sigma_i} \psi_i(\sigma_i) \prod_{a \in \partial i} \Big[\sum_{\vec{\sigma}_{\partial a \backslash \sigma_i}} \psi_a(\vec{\sigma}_{\partial a}) \prod_{j \in \partial a \backslash i} q_j(\sigma_j) \Big] \Big] \; , \tag{30}$$

$$f_a(\beta) = -\frac{1}{\beta} \ln \Big[\sum_{\vec{\sigma}_a} \psi_a(\vec{\sigma}_a) \prod_{j \in \partial a} q_j(\sigma_j) \Big] \; . \tag{31}$$

Then the total free energy $F(\beta)$ can be evaluated using Eq. (26).

4 The Bethe-Peierls mean field theory

An improvement over the Weiss mean field theory is to distinguish the cavity graphs (e.g., $G_{\backslash i}$ and $G_{\backslash a}$) and the full factor graph G. Notice that in the cavity graph $G_{\backslash a}$ of Fig. 2c the vertex j does not feel the interaction a. Let us denote its marginal probability in $G_{\backslash a}$ as $q_{j \to a}(\sigma_a)$ to distinguish it with the marginal probability $q_j(\sigma_j)$ in G. Under the factorization assumption Eq. (27) will then be modified to be

$$P_a^{cavity}(\vec{\sigma}_{\partial a}) \approx \prod_{j \in \partial a} q_{j \to a}(\sigma_j) \; . \tag{32}$$

Similarly, Eq. (28) for the cavity graph $G_{\backslash i}$ is changed to be

$$P^{cavity}_{n(i)}(\vec{\sigma}_{n(i)}) \approx \prod_{a\in\partial i}\prod_{j\in\partial a\backslash i} q_{j\to a}(\sigma_j)\ . \tag{33}$$

Equations like (32) and (33) are commonly referred to as the Bethe-Peierls approximation in the statistical physics literature. Under the Bethe-Peierls approximation the marginal distribution of vertex i is then

$$q_i(\sigma_i) \propto \psi_i(\sigma_i)\prod_{a\in\partial i}\Big[\sum_{\vec{\sigma}_{\partial a}\backslash\sigma_i}\psi_a(\vec{\sigma}_{\partial a})\prod_{j\in\partial a\backslash i} q_{j\to a}(\sigma_j)\Big]\ , \tag{34}$$

and the expressions for the free energy contributions are

$$f_i(\beta) = -\frac{1}{\beta}\ln\Big[\sum_{\sigma_i}\psi_i(\sigma_i)\prod_{a\in\partial i}\Big[\sum_{\vec{\sigma}_{\partial a}\backslash\sigma_i}\psi_a(\vec{\sigma}_{\partial a})\prod_{j\in\partial a\backslash i} q_{j\to a}(\sigma_j)\Big]\Big]\ , \tag{35}$$

$$f_a(\beta) = -\frac{1}{\beta}\ln\Big[\sum_{\vec{\sigma}_a}\psi_a(\vec{\sigma}_a)\prod_{j\in\partial a} q_{j\to a}(\sigma_j)\Big]\ . \tag{36}$$

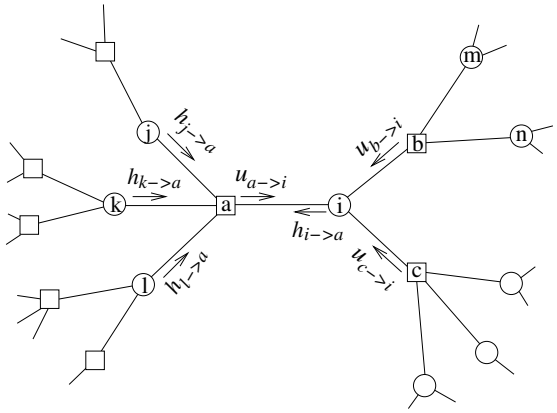


Figure 3: Message-passing by belief propagation on a factor graph G. In this figure the message $h_{i\to a}$ from vertex i to interaction a denotes the set of parameters for the cavity distribution $q_{i\to a}(\sigma_i)$; similarly the message $u_{b\to i}$ from interaction b to vertex i denotes the set of parameters for the function $p_{b\to i}(\sigma_i)$ defined in Eq. (38). Vertex i collects the incoming messages $u_{b\to i}$ and $u_{c\to i}$ and produces an outgoing message $h_{i\to a}$ to interaction a; while interaction a collects the incoming messages from vertices j, k, and l and produces an outgoing message $u_{a\to i}$ to vertex i. Figure copied from Zhou (2015).

Similar to Eq. (34), we can write down an equation for the cavity probability distribution $q_{j\to a}(\sigma_j)$ as

$$q_{j\to a}(\sigma_j) = \frac{1}{z_{j\to a}}\psi_j(\sigma_j)\prod_{b\in\partial j\backslash a} p_{b\to j}(\sigma_j)\ , \tag{37}$$

where $z_{j\to a}$ is a normalization constant and the function $p_{b\to j}(\sigma_j)$ is defined by the expression

$$p_{b\to j}(\sigma_j) \equiv \sum_{\vec{\sigma}_{\partial b}\backslash\sigma_j}\psi_b(\vec{\sigma}_{\partial b})\prod_{k\in\partial b\backslash j} q_{k\to b}(\sigma_k)\ . \tag{38}$$

Equations (37) and (38), taking together, are referred to as the belief-propagation (BP) equation for the probabilistic graphical model. The functions $q_{j\to a}(\sigma_j)$ and $p_{a\to j}(\sigma_j)$ can be understood as a pair of message functions on each link (j,a) of the factor graph G between a vertex j and an interaction a, see Fig. 3.

The BP equations (37) and (38) can be iterated on the factor graph G as a message-passing process. After a fixed-point solution is reached, the free energy $F(\beta)$ is then obtained through Eq. (26) and the marginal probabilities of all the vertices are obtained through Eq. (34). In practical applications the Bethe-Peierls mean field theory often offers rather good predictions.

5 Application in optimization and inference

Here we describe two relatively simple applications of the belief-propagation method, one on combinatorial optimization and the other on statistical inference.

5.1 The minimum vertex cover problem

The minimum vertex cover problem is a basic NP-hard combinatorial optimization problem in computer science, it also has wide practical relevance (Weigt and Hartmann, 2000). This problem is defined on a conventional graph of vertices $i, j, \ldots$ and edges $(i, j), (k, l), \ldots$ between these vertices, see Fig. 4. Each vertex i has two states $s_i = 0$ (empty) or $s_i = 1$ (occupied). The optimization goal is to minimize the total number of occupied vertices. But for each edge (i, j) between two vertices i and j, at least one of the incident vertices should be occupied, namely

$$s_i + s_j \geq 1 \qquad \forall (i, j) \; . \tag{39}$$

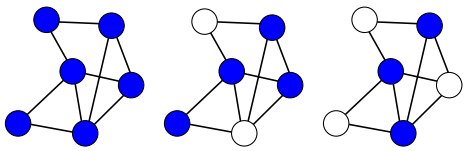


Figure 4: The minimum vertex cover problem defined on a graph with $N = 6$ vertices and $M = 9$ edges. Each edge in the graph requires that at least one of the incident vertices should be occupied (shown as filled circles). The left panel is a trivial fully occupied solution with maximum energy $E = 6$; the middle panel is a non-optimal solution with energy $E = 4$; and the right panel is an optimal solution with global minimum energy $E = 3$. Figure copied from Zhao and Zhou (2014).

Notice that each edge constraint of Eq. (39) can be regarded as an interaction with energy being infinity if $s_i = s_j = 0$ and energy being zero if $s_i = 1$ or $s_j = 1$ or both. For such a two-body interaction system, since each edge already represents an interaction, we do not need to work with a bipartite factor graph but can work directly on the conventional graph of vertices and edges.

Under the Bethe-Peierls approximation of conditional independence among all the neighboring vertices of a focal vertex i, we can easily write down the following expression for q_i^0, the probability of vertex i being empty:

$$q_i^0 = \frac{\prod\limits_{j \in n(i)} \left(1 - q_{j \to i}^0\right)}{e^{-\beta} + \prod\limits_{j \in n(i)} \left(1 - q_{j \to i}^0\right)} \; , \tag{40}$$

where $n(i)$ denotes the set formed by all the nearest neighboring vertices of vertex i, and $q_{j \to i}^0$ is the probability of vertex j being empty in the absence of edge (i, j). To intuitively understand this expression, we notice that the term $\prod_{j \in n(i)} (1 - q_{j \to i}^0)$ is the probability of all the vertices in $n(i)$ being occupied in the absence of vertex i and all its attached edges (when i is added to the graph

it can then be empty), while the term $e^{-\beta}$ is the penalty weight for vertex i being occupied. The cavity probability $q^0_{j\to i}$ can be written down similarly:

$$q^0_{j\to i} = \frac{\prod\limits_{k\in n(j)\backslash i}\left(1-q^0_{k\to j}\right)}{e^{-\beta}+\prod\limits_{k\in n(j)\backslash i}\left(1-q^0_{k\to j}\right)} \,. \tag{41}$$

The above equation is the BP equation for the minimum vertex cover problem. The iteration of this equation on a given graph instance is simple to implement.

We can turn the BP iteration process into a heuristic algorithm to construct close-to-minimum vertex cover solutions. In the belief propagation-guided decimation (BPD) procedure, for instance, we can fix a tiny fraction of the vertices i with highest estimated empty probabilities ($q^0_i \approx 1$) to the empty state and then simplify the graph; the BP iteration is then carried out on the reduced graph for a number of rounds and the occupation probabilities for all the remaining vertices are evaluated again using Eq. (40) to prepare for the next round of decimation (more details were reviewed in Zhao and Zhou (2014)). We illustrate by Fig. 5 the typical performance of such a BPD algorithm. Empirical results on random graph instances and real-world complex networks both suggested that the BPD algorithm is highly competent both in terms of solution quality and in terms of search time.

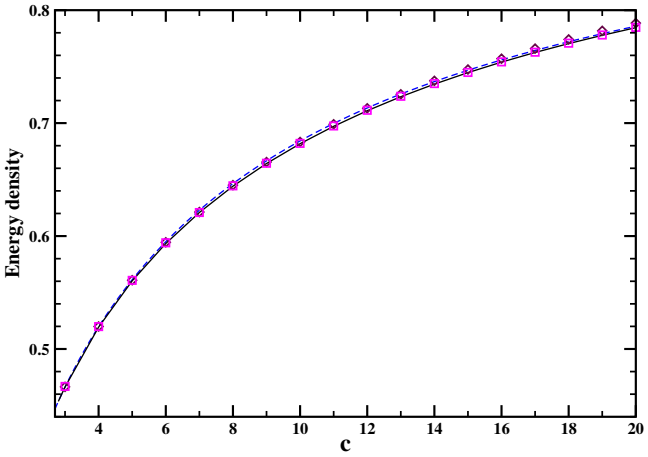


Figure 5: Comparing theoretical and algorithmic results for the random minimum vertex cover problem. Square symbols and diamond symbols are, respectively, the energy density (fraction of occupied vertices) of solutions obtained by the BPD algorithm (with $\beta = 10$) and another more advanced message-passing algorithm, on a single random graph instance of size $N = 10^5$ and mean vertex degree c. The solid and dashed lines are the theoretical predictions obtained by two mean field theories. Figure copied from Zhao and Zhou (2014).

Besides the vertex cover problem, the BPD algorithm and other message-passing algorithms have been successfully tested in many other combinatorial optimization problems and constraint satisfaction problems. They are applicable not only to problems with local constraints such as those of Eq. (39) but also to problems with global constraints (such as the cycle-constrained minimum feedback vertex set problem (Zhou, 2013)). Research efforts in applying message-passing methods to hard machine-learning tasks are now active and fruitful.

5.2 The inverse Ising problem

Consider generating many equilibrium spin configurations from the Ising model (1) with energy function

$$E(\sigma_1, \sigma_2, \ldots, \sigma_N) = -\sum_{(i,j)} J_{ij}\sigma_i\sigma_j - \sum_{k=1}^{N} h_k^0 \sigma_k \ . \tag{42}$$

Let us denote the set of configurations as $\{\vec{\sigma}^{(a)} \ : \ a = 1, 2, \ldots, M\}$. These configurations are sampled *independently* from the equilibrium Boltzmann distribution of model (42) at a fixed inverse temperature β, see the illustration in Fig. 6. The joint probability of getting these M configurations is

$$\Pr\Big[\{\vec{\sigma}^{(a)}\}\Big] = \prod_{a=1}^{M} B(\vec{\sigma}^{(a)}) \ . \tag{43}$$

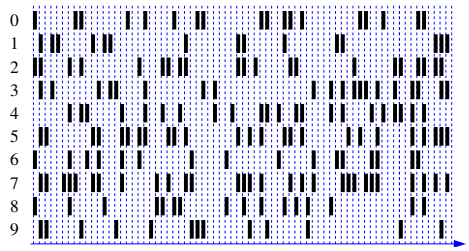


Figure 6: The inverse Ising problem. In the example shown here for the coarse-grained activity levels of neurons, the states of $N = 10$ vertices (neurons or clusters of neurons) are recorded at fixed time intervals. Each row is for a single vertex, and the horizon arrow shows the direction of time increase, while the configuration of the N vertices at a single time point corresponds to one column of the data matrix. At each time point, the state of vertex i is either active ($\sigma_i = +1$, represented by a black bar) or inactive ($\sigma_i = -1$, no bar). The task is to infer the couplings J_{ij} and external fields h_k^0 of model (42) from the empirical data matrix. Figure copied from Zhou (2015).

If all the model parameters J_{ij} and h_k^0 are not revealed to the observer, could they be reconstructed from the observed data only? The solution to this problem is based on the belief or hypothesis that the M configurations have the largest probability to be observed. The log-likelihood of observing these configurations is expressed as

$$\ln \Pr\big[\{\vec{\sigma}^{(a)}\}\big] \equiv \sum_{a=1}^{M} \ln B(\vec{\sigma}^{(a)}) = -\sum_{a=1}^{M} \beta E(\vec{\sigma}^{(a)}) - M \ln\Big[\sum_{\vec{\sigma}} e^{-\beta E(\vec{\sigma})}\Big] \ . \tag{44}$$

We are looking for the parameters J_{ij} and $h_k^0\}$ which maximize this log-likelihood or equivalently, minimize the lost function

$$\mathcal{L}\big(\{J_{ij}\}, \{h_k^0\}\big) \equiv -\frac{1}{M} \ln \Pr\big[\{\vec{\sigma}^{(a)}\}\big] = \frac{1}{M}\sum_{a=1}^{M} \beta E(\vec{\sigma}^{(a)}) + \ln\Big[\sum_{\vec{\sigma}} e^{-\beta E(\underline{\sigma})}\Big] \ . \tag{45}$$

The first derivative of $\mathcal{L}$ with respect to J_{ij} is

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} = -\beta \frac{1}{M}\sum_{a=1}^{M} \sigma_i^{(a)}\sigma_j^{(a)} + \beta \frac{\sum_{\vec{\sigma}} \sigma_i\sigma_j e^{-\beta E(\vec{\sigma})}}{\sum_{\vec{\sigma}} e^{-\beta E(\vec{\sigma})}} \tag{46}$$

$$= -\beta\Big[\big\langle \sigma_i\sigma_j \big\rangle_{data} - \big\langle \sigma_i\sigma_j \big\rangle_{model}\Big] \ . \tag{47}$$

Therefore if the correlation $\langle \sigma_i \sigma_j \rangle_{data}$ between vertices i and j in the empirical data is larger than the correlation $\langle \sigma_i \sigma_j \rangle_{model}$ predicted by the model, then the coupling J_{ij} should be increased; otherwise J_{ij} should be decreased. As a simple learning rule, we can make a small change $\delta J_{ij}(t)$ to the coupling J_{ij} at time t as

$$\delta J_{ij}(t) = \lambda \Big[\langle \sigma_i \sigma_j \rangle_{data} - \langle \sigma_i \sigma_j \rangle_{model} \Big] , \tag{48}$$

where λ is a small positive-valued learning parameter. Similarly we can derive a learning rule for the external field h_i^0 as

$$\delta h_i^0(t) = \lambda \Big[\langle \sigma_i \rangle_{data} - \langle \sigma_i \rangle_{model} \Big] , \tag{49}$$

where $\langle \sigma_i \rangle_{data}$ is the mean value of σ_i among the M observed configurations, and $\langle \sigma_i \rangle_{model}$ is the predicted value of σ_i from the model (42). The values of $\langle \sigma_i \sigma_j \rangle$ and $\langle \sigma_i \rangle$ are straightforward to compute from the empirical data. The real challenge is to compute the predicted mean values of $\sigma_i \sigma_j$ and σ_i accurately and efficiently. There are of course many different ways for doing this. Here we describe the message-passing inference method.

For binary spin $\sigma_i \in \pm 1$, the marginal distribution of vertex i can be parameterized as $q_i(\sigma_i) = e^{\beta h_i \sigma_i}/[2\cosh(\beta h_i)]$, where h_i is simply the total (magnetic) field on this vertex. Similarly, $q_{i \to j}(\sigma_i) = e^{\beta h_{i \to j} \sigma_i}/[2\cosh(\beta h_{i \to j})]$ with $h_{i \to j}$ being the cavity field on vertex i. Based on the Bethe-Peierls approximation, we have

$$h_i = h_i^0 + \sum_{j \neq i} u_{j \to i} , \tag{50}$$

$$h_{i \to j} = h_i^0 + \sum_{k \neq i,j} u_{k \to i} = h_i - u_{j \to i} . \tag{51}$$

Here $u_{j \to i}$ is the field contribution from vertex j:

$$u_{j \to i} = \frac{1}{\beta} \mathrm{atanh} \Big[\tanh(\beta h_{j \to i}) \tanh(\beta J_{ij}) \Big] . \tag{52}$$

Because each vertex interacts with every other vertex in the system, it may be safe to expect that the cavity field $h_{j \to i}$ differs only very slightly from h_j. Under this approximation we obtain that

$$\tanh(\beta h_{j \to i}) \approx \tanh(\beta h_j) - \big(1 - \tanh^2(\beta h_j)\big) \mathrm{atanh} \big[\tanh(\beta h_i) \tanh(\beta J_{ij}) \big] , \tag{53}$$

and then

$$\begin{aligned} \beta u_{j \to i} \approx\ & \mathrm{atanh} \Big[\tanh(\beta h_j) \tanh(\beta J_{ij}) \Big] \\ & - \frac{\tanh(\beta J_{ij}) \big[1 - \tanh^2(\beta h_j) \big]}{1 - \tanh^2(\beta h_j) \tanh^2(\beta J_{ij})} \mathrm{atanh} \big[\tanh(\beta h_i) \tanh(\beta J_{ij}) \big] . \end{aligned} \tag{54}$$

Therefore we get from Eq. (50) that

$$h_i \approx h_i^0 + \sum_{j \neq i} \frac{1}{\beta} \mathrm{atanh} \big[\tanh(\beta h_j) \tanh(\beta J_{ij}) \big] \tag{55}$$

$$- \sum_{j \neq i} \frac{\tanh(\beta J_{ij}) \big[1 - \tanh^2(\beta h_j) \big]}{\beta \big[1 - \tanh^2(\beta h_j) \tanh^2(\beta J_{ij}) \big]} \mathrm{atanh} \big[\tanh(\beta h_i) \tanh(\beta J_{ij}) \big] . \tag{56}$$

which does not involves the cavity messages and is more convenient for numerical computations. The third term of Eq. (56) is called an Onsager retraction term. If we further assume that $\beta J_{ij} \approx 0$, the above expression can be further simplified as

$$h_i \approx h_i^0 + \sum_{j \neq i} J_{ij} \Big[\tanh(\beta h_j) - \beta J_{ij} \big(1 - \tanh^2(\beta h_j) \big) \tanh(\beta h_i) \Big] . \tag{57}$$

Equation (57) was first derived by Thouless et al. (1977) and is commonly referred to as the TAP equation. After a fixed-point solution of all the N fields h_i is reached by iterating Eq. (56) or Eq. (57), the predicted mean value of σ_i is then simply

$$\langle \sigma_i \rangle_{model} = \tanh\big(\beta h_i\big) \ . \tag{58}$$

To compute the mean value of $\sigma_i \sigma_j$, we notice that the joint distribution of σ_i and σ_j under the Bethe-Peierls approximation is

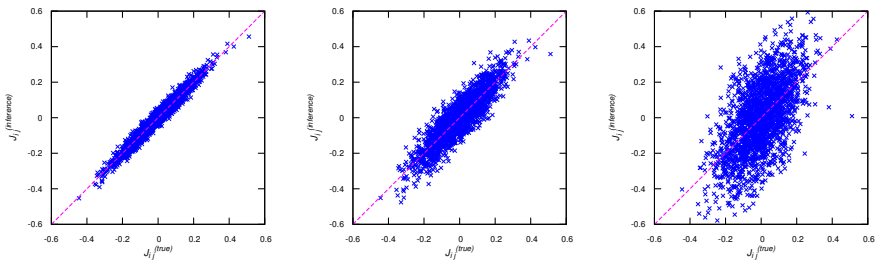
$$\Pr(\sigma_i, \sigma_j) \propto \exp\big(\beta h_{i\to j}\sigma_i + \beta J_{ij}\sigma_i\sigma_j + \beta h_{j\to i}\sigma_j\big) \ . \tag{59}$$

It is then almost straightforward to derive the following mean-field equation for any $i \neq j$

$$\langle \sigma_i \sigma_j \rangle_{model} = \frac{\tanh(\beta J_{ij}) + \tanh(\beta h_i)\tanh(\beta h_j)}{1 + \tanh(\beta J_{ij})\tanh(\beta h_i)\tanh(\beta h_j)} \times \Big\{1 + o(1)\Big\} \ . \tag{60}$$

(The correction term $o(1)$ in this expression comes from approximating $h_{i\to j}$ by h_i and $h_{j\to i}$ by h_j.)

Figure 7 shows the performance of this simple inference algorithm on a problem instance of $N = 64$ vertices and $M = 10^6$ samples. We find the inference accuracy is satisfactory for configurations sampled at low values of the inverse temperature β, but the performance deteriorates with β. The basic reason behind this phenomenon is the build up of correlations within each sampled configurations. As the temperature is lowered (and β increases), the correlations among the states of the vertices become more and more stronger; but these correlations are largely ignored in the simple BP inference method.



(a) (b) (c)

Figure 7: Reconstructing the Sherrington-Kirkpatrick model. A total number of $M = 10^6$ equilibrium configurations were generated by Markov-Chain Monte-Carlo at fixed inverse temperature β for a single instance of the SK model with $N = 64$ vertices. The external fields are set to $h_k^0 = 0$ in the model, and each coupling constant $J_{ij}^{(true)}$ (with $1 \leq i < j \leq N$) of the model is independently sampled from the Gaussian distribution with mean zero and variance $1/N$. The reconstructed values $J_{ij}^{(inference)}$ of the coupling constants are compared with the corresponding true values. (a) Data sampled at $\beta = 0.2$; (b) $\beta = 0.5$; (c) $\beta = 1.0$. Simulation data courtesy of Mr. Chen-Yi Gao (2017).

There are various ways to improve the inference performance, see recent reviews by Cocco et al. (2017) and Nguyen et al. (2017). The inverse Ising problem has important applications in neuron firing data analysis (Schneidman et al., 2006; Roudi et al., 2009) and in protein structure prediction (Weigt et al., 2009). This problem is also referred to as the direct contact analysis in the protein folding research community (Cocco et al., 2017; Nguyen et al., 2017).

Machine learning algorithms are to a great extent (advanced) inference algorithms. Learning rules similar to Eqs. (48) and (49) are commonly encountered in variously machine learning tasks (e.g., the restricted Boltzmann machine). Message-passing methods are indeed very helpful for inference problem (Zdeborová and Krzakala, 2016).

6 Beyond the elementary steps

We introduced some of the basic concepts and methods for solving probabilistic graphical models. Many important issues were not addressed here. For readers interested in exploring more, we list here some of these more advanced topics on probabilistic graphical models.

1. *Ergodicity breaking.* At low temperature (and high inverse temperature β) the relevant configuration space of the graphical model (5) might break into many widely separated sub-spaces. More advanced mean field theories have been developed to tackle this difficult situation (Mézard and Parisi, 2001).

2. *Loop-expansion framework.* Mean field theories and message-passing equations can also be derived by expanding the partition function (7) into a loop series and keeping on the leading term (Zhou, 2015).

3. *Beyond the Bethe-Peierls approximation.* We can adopt the Kikuchi cluster variational framework to include more local correlations among the vertices of a given model and to develop generalized belief-propagation equations (Yedidia et al., 2005; Pelizzola, 2005).

Acknowledgement

The author acknowledges financial support from the Chinese Academy of Sciences and from the National Natural Science Foundation of China (grant number 11647601).

References

S.-I. Amari. Neural theory of association and concept-formation. *Biol. Cybernetics*, 26:175–185, 1977.

S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt. Inverse statistical physics of protein sequences: A key issues review. *e-print*, page arXiv:1703.01222, 2017.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79:2554–2558, 1982.

F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory*, 47:498–519, 2001.

M. Mézard and G. Parisi. The bethe lattice spin glass revisited. *Eur. Phys. J. B*, 20:217–233, 2001.

M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.

H. C. Nguyen, R. Zecchina, and J. Berg. Inverse statistical problems: from the inverse ising problem to data science. *preprint*, page arXiv:1702.01522, 2017.

A. Pelizzola. Cluster variation method in statistical physics and probabilistic graphical models. *J. Phys. A: Meth. Gen.*, 38:R309–R339, 2005.

Y. Roudi, E. Aurell, and J. A. Hertz. Statistical physics of pairwise probability models. *Front. Comput. Neurosci.*, 3:22, 2009. doi: 10.3389/neuro.10.022.2009.

E. Schneidman, M. J. Berry II, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007–1012, 2006.

D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin glass'. *Phil. Mag.*, 35:593–601, 1977.

M. Weigt and A. K. Hartmann. Number of guards needed by a museum: A phase transition in vertex covering of random graphs. *Phys. Rev. Lett.*, 84: 6118–6121, 2000.

M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message-passing. *Proc. Natl. Acad. Sci. USA*, 106:67–72, 2009.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief-propagation algorithms. *IEEE Trans. Inf. Theory*, 51:2282–2312, 2005.

L. Zdeborová and F. Krzakala. Statistical physics of inference: thresholds and algorithms. *Adv. Phys.*, 65:453–552, 2016.

J.-H. Zhao and H.-J. Zhou. Statistical physics of hard combinatorial optimization: Vertex cover problem. *Chin. Phys. B*, 23:078901, 2014. doi: 10.1088/1674-1056/23/7/078901.

H.-J. Zhou. Spin glass approach to the feedback vertex set problem. *Eur. Phys. J. B*, 86:455, 2013.

H.-J. Zhou. *Spin Glass and Message Passing.* Science Press, Beijing, China, 2015.