PHYSICAL REVIEW E **73**, 065102(R) (2006)

Loop calculus in statistical physics and information science

Michael Chertkov[1] and Vladimir Y. Chernyak[2]
[1]*Theoretical Division and Center for Nonlinear Studies, LANL, Los Alamos, New Mexico 87545, USA*
[2]*Department of Chemistry, Wayne State University, 5101 Cass Avenue, Detroit, Michigan 48202, USA*
(Received 20 January 2006; published 1 June 2006)

Considering a discrete and finite statistical model of a general position we introduce an exact expression for the partition function in terms of a finite series. The leading term in the series is the Bethe-Peierls (belief propagation) (BP) contribution; the rest are expressed as loop contributions on the factor graph and calculated directly using the BP solution. The series unveils a small parameter that often makes the BP approximation so successful. Applications of the loop calculus in statistical physics and information science are discussed.

DOI: 10.1103/PhysRevE.73.065102 PACS number(s): 05.50.+q, 89.70.+c

Discrete statistical models, the Ising model being the most famous example, play a prominent role in theoretical and mathematical physics. They are typically defined on a lattice, and major efforts in the field focused primarily on the case of the infinite lattice size. Similar statistical models emerge in information science. However, the most interesting questions there are related to graphs that are very different from a regular lattice. Moreover, it is often important to consider large but finite graphs. Statistical models on graphs with long loops are of particular interest in the fields of error correction and combinatorial optimization. These graphs are treelike locally.

A theoretical approach pioneered by Bethe [1] and Peierls [2] (see also [3]), who suggested analyzing statistical models on perfect trees, has largely remained a useful efficiently solvable toy. Indeed, these models on trees are effectively one dimensional, and thus exactly solvable in the theoretical sense, while computational effort scales linearly with the generation number. The exact tree results have been extended to higher-dimensional lattices as uncontrolled approximations. In spite of the absence of analytical control the Bethe-Peierls approximation gives remarkably accurate results, often outperforming standard mean-field results. The *ad hoc* approach was also restated in a variational form [4,5]. Except for two recent papers [6,7] that will be discussed later in this Rapid Communication, no systematic attempts to construct a regular theory with a well-defined small parameter and the Bethe-Peierls as its leading approximation have been reported.

A similar tree-based approach in information science has been developed by Gallager [8] in the context of error-correction theory. Gallager introduced so-called low-density parity-check (LDPC) codes, defined on locally treelike Tanner graphs. The problem of ideal decoding, i.e., restoring the most probable preimage out of the exponentially large pool of candidates, is identical to solving a statistical model on the graph [9]. An approximate yet efficient decoding belief-propagation algorithm introduced by Gallager constitutes an iterative solution of the Bethe-Peierls equations derived as if the statistical problem was defined on a tree that locally represents the Tanner graph. We utilize this coincidence to call the Bethe-Peierls and belief-propagation equations by the same acronym BP. Recent resurgence of interest in LDPC codes [10], as well as proliferation of the BP approach to other areas of information and computer science, e.g., artificial intelligence [11] and combinatorial optimization [12], where interesting statistical models on graphs with long loops are also involved, posed the following questions. Why does the BP method perform so well on graphs with loops? What is the hidden small parameter that ensures exceptional performance of the BP approach? How can we systematically correct the BP equations? This Rapid Communication provides systematic answers to all these questions.

The Rapid Communication is organized as follows. We start with introducing notations for a generic statistical model, formulated in terms of interacting Ising variables with the network described via a factor graph. We next state our main result: a decomposition of the partition function of the model in a finite series. The BP expression for the model represents the first term in the series. All other terms correspond to closed, undirected subgraphs of the factor graph, possibly branching yet not terminating at a node, which are referred to as generalized loops. The simplest diagram is a single loop. An individual contribution is the product of local terms along a generalized loop, expressed explicitly in terms of simple correlation functions calculated within the BP approach. We proceed with discussing the meaning of the BP equation as a successful approximation in terms of the loop series, followed by presenting a clear derivation of the loop series. The derivation includes three steps. We first introduce a family of local gauge transformations, two per original Ising variable. The gauge transformation changes individual terms in the expansion with the full expression for the partition function natually remaining unchanged. We then fix the gauge in such a way that only those terms that correspond to generalized loops contribute to the modified series. Finally, we show that the first term in the resulting generalized loop series corresponds exactly to the standard BP approximation. This interprets the BP approach as a special gauge choice. We conclude with clarifying the relation of this work to other recent advances in the subject, and discuss possible applications and generalizations of the approach.

Consider a generic discrete statistical model defined for an arbitrary finite undirected graph Γ, with bits $a,b=1,\ldots,m$ with the neighbors connected by edges $(a,b),\ldots$, the neighbor relation expressed as $a\in b$ or $b\in a$. Configurations $\boldsymbol{\sigma}$ are characterized by sets of binary (spin) variables $\sigma_{ab}=\pm 1$, associated with the graph edges: $\boldsymbol{\sigma}=\{\sigma_{ab};(a,b)\in\Gamma\}$. The probability of configuration $\boldsymbol{\sigma}$ is

 ©2006 The American Physical Society

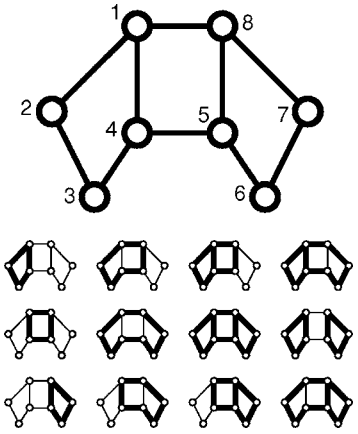

FIG. 1. Example of a factor graph. Twelve possible marked paths (generalized loops) are shown in bold in the bottom part.

$$p(\boldsymbol{\sigma}) = Z^{-1} \prod_{a \in \Gamma} f_a(\boldsymbol{\sigma}_a), \quad Z = \sum_{\boldsymbol{\sigma}} \prod_{a \in \Gamma} f_a(\boldsymbol{\sigma}_a), \tag{1}$$

$f_a(\boldsymbol{\sigma}_a)$ being a non-negative function of $\boldsymbol{\sigma}_a$, a vector built of σ_{ab} with $b \in a$: $\boldsymbol{\sigma}_a = \{\sigma_{ab}; b \in a\}$. The notation assumes $\sigma_{ab} = \sigma_{ba}$. Our vertex model generalizes the celebrated six- and eight-vertex models of Baxter [3]. An example of a factor graph with $m=8$ that corresponds to $p(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = Z^{-1} \Pi_{a=1}^{8} f_a(\boldsymbol{\sigma}_a)$, where $\boldsymbol{\sigma}_1 \equiv (\sigma_2, \sigma_4, \sigma_8)$, $\boldsymbol{\sigma}_2 \equiv (\sigma_1, \sigma_3)$, $\boldsymbol{\sigma}_3 \equiv (\sigma_2, \sigma_4)$, $\boldsymbol{\sigma}_4 \equiv (\sigma_1, \sigma_3, \sigma_5)$, $\boldsymbol{\sigma}_5 \equiv (\sigma_4, \sigma_6, \sigma_8)$, $\boldsymbol{\sigma}_6 \equiv (\sigma_5, \sigma_7)$, $\boldsymbol{\sigma}_7 \equiv (\sigma_6, \sigma_8)$, $\boldsymbol{\sigma}_8 \equiv (\sigma_1, \sigma_5, \sigma_7)$, is shown in Fig. 1.

The main exact result of this Rapid Communication is decomposition of the partition function defined by Eq. (1) in a finite series:

$$Z = Z_0 \left(1 + \sum_{C} \frac{\prod_{a \in C} \mu_a(C)}{\prod_{(a,b) \in C} (1 - m_{ab}(C)^2)} \right), \tag{2}$$

$$m_{ab}(C) = \sum_{\sigma_{ab}} \sigma_{ab} b_{ab}(\sigma_{ab}), \tag{3}$$

$$\mu_a = \sum_{\sigma_a} \prod_{b \in a, C}^{b \neq a} (\sigma_{ab} - m_{ab}) b_a(\boldsymbol{\sigma}_a), \tag{4}$$

where the summation goes over all allowed (marked) paths C, or generalized loops. They consist of bits each with at least two distinct neighbors along the path. Twelve allowed marked paths for our example are shown in the bottom part of Fig. 1. A generalized loop can be disconnected, e.g., the last one in the second row shown in Fig. 1. In Eqs. (2) $b_{ab}(\sigma_{ab})$, $b_a(\boldsymbol{\sigma}_a)$, and Z_0 are beliefs (probabilities) defined on edges, bits, and the partition function, respectively, calculated within the BP approach. A BP solution can be interpreted as an exact solution in an infinite tree built by unwrapping the factor graph. A BP solution can be also interpreted [5] as a set of beliefs that minimize the Bethe free energy

$$\mathcal{F} = \sum_{a} \sum_{\boldsymbol{\sigma}_a} b_a(\boldsymbol{\sigma}_a) \ln \frac{b_a(\boldsymbol{\sigma}_a)}{f_a(\boldsymbol{\sigma}_a)} - \sum_{(a,b)} \sum_{\sigma_{ab}} b_{ab}(\sigma_{ab}) \ln b_{ab}(\sigma_{ab}),$$

under the set of realizability $0 \leqslant b_a(\boldsymbol{\sigma}_a), b_{ab}(\sigma_{ab}) \leqslant 1$, normalization $\Sigma_{\sigma_a} b_a(\boldsymbol{\sigma}_a) = \Sigma_{\sigma_{ab}} b_{ab}(\sigma_{ab}) = 1$, and consistency $\Sigma_{\boldsymbol{\sigma}_a \setminus \sigma_{ab}} b_a(\boldsymbol{\sigma}_a) = b_{ab}(\sigma_{ab})$ constraints. The term associated with a marked path is the ratio of the products of irreducable correlation functions (4) and the quadratic magnetization at-edge functions (3) calculated along the marked path C within the BP approximation.

As usual in statistical mechanics exact expressions for the spin correlation functions can be obtained by differentiating Eq. (2) with respect to the proper factor functions. In the tree (no loops) case only the unity term in the right-hand side (RHS) of Eq. (2) survives. In the general case Eq. (2) provides a clear criterion for the BP approximation validity: The sum over the loops in the RHS of Eq. (2) should be small compared to 1. The number of terms in the series increases exponentially with the number of bits. Therefore, Eq. (2) becomes useful for selecting a smaller than exponential number of leading contributions. In a large system the leading contribution comes from the paths with the number of degree-2 connectivity nodes substantially exceeding the number of branching nodes, i.e., the ones with higher connectivity degree. According to Eq. (2) the contribution of a long path is given by the ratio of the along-the-path product of the irreducible nearest-neighbor spin correlation functions associated with a bit μ_a to the along-the-path product of the edge contributions $1/(1-m_{ab}^2)$. All are calculated within the BP approximation. Therefore, the small parameter in the perturbation theory is $\varepsilon = \Pi_{a \in C} \mu_a(C) / \Pi_{(a,b) \in C} (1 - m_{ab}^2)$. If ε is much smaller than 1 for all marked paths the BP approximation is valid. We anticipate the loop formula (2) to be extremely useful for analysis and possible differentiation between the loop contributions. Whether the series is dominated by a single-loop contribution or some number of comparable loop corrections will depend on the problem specifics (form of the factor graph and functions). In the former case the leading correction to the BP result is given by the marked path with the largest ε.

We now turn to derivation of the loop formula. Let us relax the condition $\sigma_{ab} = \sigma_{ba}$ in Eq. (1) and treat σ_{ab} and σ_{ba} as independent variables. This allows us to represent the partition function in the form

$$Z = \sum_{\boldsymbol{\sigma}'} \prod_{a} f_a(\boldsymbol{\sigma}_a) \prod_{(b,c)} \frac{1 + \sigma_{bc} \sigma_{cb}}{2}, \tag{5}$$

where there are twice more components since any pair of variables σ_{ab} and σ_{ba} enters $\boldsymbol{\sigma}'$ independently. It is also assumed in Eq. (5) that each edge contributes to the product over (b,c) only once. The representation (5) is advantageous over the original one (1) since $\boldsymbol{\sigma}_a$ at different bits become independent. We further introduce a parameter vector $\boldsymbol{\eta}$ with independent components η_{ab} (i.e., $\eta_{ab} \neq \eta_{ba}$). Making use of the key identity

$$\frac{\cosh(\eta_{bc}+\eta_{cb})(1+\sigma_{bc}\sigma_{cb})}{(\cosh\eta_{bc}+ma_{bc}\sinh\eta_{bc})(\cosh\eta_{cb}+\sigma_{cb}\sinh\eta_{cb})}=V_{bc},$$

$$V_{bc}(\sigma_{bc},\sigma_{cb})=1+[\sinh(\eta_{bc}+\eta_{cb})-\sigma_{bc}\cosh(\eta_{bc}+\eta_{cb})]\times[\sinh(\eta_{bc}+\eta_{cb})-\sigma_{cb}\cosh(\eta_{bc}+\eta_{cb})], \tag{6}$$

we transform the product over edges on the RHS of Eq. (5) to arrive at

$$Z=\left(\prod_{(b,c)}2\cosh(\eta_{bc}+\eta_{cb})\right)^{-1}\sum_{\boldsymbol{\sigma}'}\prod_{a}P_a\prod_{bc}V_{bc}, \tag{7}$$

$$P_a(\boldsymbol{\sigma}_a)=f_a(\boldsymbol{\sigma}_a)\prod_{b\in a}(\cosh\eta_{ab}+\sigma_{ba}\sinh\eta_{ab}). \tag{8}$$

The desired decomposition Eq. (2) is obtained by choosing some special values for the η variables (fixing the gauge) and expanding the V terms in Eq. (7) in a series followed by a local computation (summations over σ variables at the edges). Individual contributions to the series are naturally identified with subgraphs of the original graph defined by a simple rule: Edge (a,b) belongs to the subgraph if the corresponding "vertex" V_{ab} on the RHS of Eq. (7) contributes using its second (nonunity) term, naturally defined according to Eq. (6). We next utilize the freedom in the choice of $\boldsymbol{\eta}$. The contributions that originate from subgraphs with loose ends vanish provided the following system of equations is satisfied:

$$\sum_{\boldsymbol{\sigma}_a}[\tanh(\eta_{ab}+\eta_{ba})-\sigma_{ba}]P_a(\boldsymbol{\sigma}_a)=0. \tag{9}$$

The number of equations is exactly equal to the number of η variables. Moreover, Eqs. (9) are nothing but BP equations: simple algebraic manipulations (see [13] for details) allow one to recast Eq. (9) in a more traditional BP form

$$\tanh\eta_{ba}=\frac{\sum_{\boldsymbol{\sigma}_a}\sigma_{ab}f_a(\boldsymbol{\sigma}_a)\prod_{c\in a}^{c\neq b}(\cosh\eta_{ac}+\sigma_{ac}\sinh\eta_{ac})}{\sum_{\boldsymbol{\sigma}_a}f_a(\boldsymbol{\sigma}_a)\prod_{c\in a}^{c\neq b}(\cosh\eta_{ac}+\sigma_{ac}\sinh\eta_{ac})},$$

with the relation between the beliefs that minimize the Bethe free energy $\mathcal{F}$ and the η fields according to

$$b_a(\boldsymbol{\sigma}_a)=\frac{P_a(\boldsymbol{\sigma}_a)}{\sum_{\boldsymbol{\sigma}_a}P_a(\boldsymbol{\sigma}_a)}.$$

The final expression Eq. (2) emerges as a result of direct expansion of the V term in Eq. (5), performing summations over local σ variables, making use of Eqs. (3) and (4), and also identifying the BP expression for the partition function as

$$Z_0=\frac{\prod_a P_a(\boldsymbol{\sigma}_a)}{\prod_{(b,c)}2\cosh(\eta_{bc}+\eta_{cb})}.$$

To summarize, Eq. (2) represents a finite series where all individual contributions are related to the corresponding generalized loops. This fine feature is achieved via a special selection of the BP gauge (9). The condition enforces the "no loose ends" rule, thus prohibiting anything but generalized loop contributions to Eq. (2). Any individual contribution is expressed explicitly in terms of the BP solution.

We expect that BP equations may have multiple solutions for the model with loops. This expectation naturally follows from the notion of the infinite covering graph, as different BP solutions correspond to different ways to spontaneously break symmetry on the infinite structure. These different BP solutions will generate loop series (2) that are different term by term but give the same result for the sum. Finding the "optimal" BP solution with the smallest ε, characterizing loop corrections to the BP solution, is important for applications. A solution related to the absolute minimum of the Bethe free energy would be a natural candidate. However, one cannot guarantee that the absolute minimum, as opposed to other local minima of $\mathcal{F}$, is always optimal for arbitrary f_α.

We further briefly discuss other models related to the general one discussed in the paper. The vertex model can be considered on a graph of the special oriented or biparitite type. A bipartite graph contains two families of nodes, referred to as bits and checks, so that the neighbor relations occur only between nodes from opposite families. A bipartite factor-graph model with an additional property that any factor associated with a bit is nonzero only if all Ising variables at the neighboring edges are the same leads to the factor-graph model considered in [5]. Actually, this factorization condition means reassignment of the Ising variables, defined at the edges of the original vertex model, to the corresponding bits of the bipartite factor-graph model. Furthermore, if only checks of degree 2 (each connected to only two bits) are considered, the bipartite factor graph model is reduced to the standard binary-interaction Ising model. The loop series derived in this Rapid Communication is obviously valid for all less general aforementioned models. Also note that the bipartite factor-graph model was chosen in [13] to introduce an alternative derivation of the loop series via an integral representation, where the BP approximation corresponds to the saddle-point approximation for the resulting integral.

Let us now comment on two relevant papers [6,7]. The Ising model on a graph with loops has been considered by Montanari and Rizzo [6], where a set of exact equations has been derived that relates the correlation functions to each other. This system of equations is underdefined; however, if irreducible correlations are neglected, the BP result is restored. This feature has been used [6] to generate a perturbative expansion for corrections to the BP equations in terms of irreducible correlations. A complementary approach for the Ising model on a lattice has been taken by Parisi and Slanina [7], who utilized an integral representation developed by Efetov [14]. The saddle point for the integral representation used in [7] turns out to be exactly the BP solution. Calculating perturbative corrections to magnetization, the authors of [7] encountered divergences in their representation for the partition function; however, the divergences canceled out from the leading order correction to the magnetization revealing a sensible loop correction to the BP approximation. These papers, [6,7], became important initial steps toward calculating and understanding loop corrections to the BP approximation. However, both approaches are very far from

being complete and problem-free. Thus, [6] lacks an invariant representation in terms of the partition function, and requires operating with correlation functions instead. Besides, the complexity of the equations related to the higher-order corrections rapidly grows with the order. The complementary approach of [7] contains dangerous (since lacking analytical control) divergences (zero modes), which constitutes a very problematic symptom for any field theory. Both [6,7] focus on the Ising pairwise interaction model. The extensions of the proposed methods to the multibit interaction cases that are most interesting from the information theory viewpoint do not look straightforward. Finally, the approaches of [6,7], if extended to higher-order corrections, will result in infinite series. Resumming the corrections in all orders, so that the result is presented in terms of a finite series, does not look feasible within the proposed techniques.

We conclude with a discussion of possible applications and generalizations. We see a major utility for Eq. (2) in its direct application to models without short loops. In this case Eq. (2) constitutes an efficient tool for improving the BP approximation through accounting for the shortest loop corrections first and then moving gradually (up to the point when complexity is still feasible) to account for longer and longer loops. Another application of Eq. (2) is direct use of ε as a test parameter for the BP approximation validity: If the shortest loop corrections to the BP equations are not small one should either look for another BP solution (hoping that the loop correction will be small within the corresponding loop series) or conclude that no feasible BP solution, resulting in a small ε, can be used as a valid approximation. There is also a strong generalization potential here. If a problem is multiscale with both short and long loops present in the factor graph, a development of a synthetic approach combining the generalized belief propagation approach of [5] (which is efficient in accounting for local correlations) and a corresponding version of Eq. (2) can be beneficial. Finally, our approach can also be useful for analysis of standard (for statistical physics and field theory) lattice problems. A particularly interesting direction will be to use Eq. (2) for introducing a new form of resummation of different scales. This can be applied for analysis of the lattice models at the critical point where correlations are long range.

We are thankful to M. Stepanov for many fruitful discussions. The work at LANL was supported by the LDRD program, and through startup funds at WSU.

[1] H. A. Bethe, Proc. R. Soc. London, Ser. A **150**, 552 (1935).
[2] R. Peierls, Proc. Cambridge Philos. Soc. **32**, 477 (1936).
[3] R. J. Baxter, *Exactly Solvable Models in Statistical Mechanics* (Academic, New York, 1982).
[4] R. Kikuchi, Phys. Rev. **81**, 988 (1951).
[5] J. S. Yedidia, W. T. Freeman, and Y. Weiss, IEEE Trans. Inf. Theory **51**, 2282 (2005).
[6] A. Montanari and T. Rizzo, J. Stat. Mech.: Theory Exp. 2005, P10011.
[7] G. Parisi and F. Slanina, J. Stat. Mech.: Theory Exp. 2006, L02003.
[8] R. G. Gallager, *Low Density Parity Check Codes* (MIT Press, Cambridge, MA, 1963).
[9] N. Sourlas, Nature (London) **339**, 693 (1989).
[10] D. J. C. MacKay, IEEE Trans. Inf. Theory **45**, 399 (1999).
[11] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference* (Kaufmann, San Francisco, 1988).
[12] M. Mezard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).
[13] M. Chertkov and V. Chernyak, e-print cond-mat/0603189.
[14] K. B. Efetov, Physica A **167**, 119 (1990).