

Loop Calculus and Belief Propagation for q -ary Alphabet: Loop Tower

Vladimir Y. Chernyak

Department of Chemistry, Wayne State University
5101 Cass Ave Detroit, MI 48202
chernyak@chem.wayne.edu

Michael Chertkov

Theoretical Division, T-13 and Center for Nonlinear Studies,
LANL, MS B213, T-13, Los Alamos, NM 87545
chertkov@lanl.gov

Abstract—Loop Calculus introduced in [1], [2] constitutes a new theoretical tool that explicitly expresses the symbol Maximum-A-Posteriori (MAP) solution of a general statistical inference problem via a solution of the Belief Propagation (BP) equations. This finding brought a new significance to the BP concept, which in the past was thought of as just a loop-free approximation. In this paper we continue a discussion of the Loop Calculus. We introduce an invariant formulation which allows to generalize the Loop Calculus approach to a q -ary alphabet.

The manuscript is organized as follows. In Section I we introduce a new formulation of the Loop Calculus in terms of a set of gauge transformations that keep the partition function of the problem invariant. The full expression contains two terms referred to as the “ground state” and “excited states” contributions. The BP equations are interpreted as a special (BP) gauge fixing condition that emerges as a special orthogonality constraint between the ground and the excited states. Stated differently, it selects the generalized loop contributions as the only ones that survive among the excited states. In Section II we demonstrate how the invariant interpretation of the Loop Calculus, introduced in the Section I, allows a natural extension to the case of a general q -ary alphabet. This is achieved via a loop tower sequential construction. The ground level in the tower is exactly equivalent to assigning one color (out of q available) to the “ground state” and considering all “excited” states to be colored in the remaining $(q - 1)$ colors, according to the loop calculus rule. Sequentially, the second level in the tower corresponds to selecting a loop from the previous step, colored in $(q - 1)$ colors, and repeating the same ground vs excited states partitioning procedure into one and the remaining $(q - 2)$ colors, respectively. The construction proceeds until the complete set of $(q - 1)$ levels in the loop tower (including the corresponding contributions to the partition function) is established. In Section III we discuss an ultimate relation between the loop calculus and the Bethe free energy variational approach of [3].

We start with defining a statistical inference problem using the so-called Forney-style graphical model formulation [4], [5]. The basic graph, $C_0 = (\mathcal{V}_0, \mathcal{E}_0)$, is described in terms of vertices, $\mathcal{V}_0 = \{a\}$ and edges, $\mathcal{E}_0 = \{(ab)\}$. Variables, associated with the edges, assume their values in a q -ary alphabet, $\sigma_{ab} = \sigma_{ba} = 0, \dots, (q - 1)$. The probability of a given configuration of variables $\sigma = \{\sigma_{ab} | (ab) \in \mathcal{E}_0\}$ on

the entire graph is described by

$$p(\sigma) = Z_{C_0}^{-1} \prod_a f_a(\sigma_a), \quad Z_{C_0} = \sum_{\sigma} \prod_a f_a(\sigma_a), \quad (1)$$

where Z_{C_0} is the normalization coefficient, also known as the partition function; $f_a(\sigma_a)$ is an arbitrary positive function of the variables, $\sigma_a \equiv \{\sigma_{ab} | b \in a, \mathcal{E}_0\}$, associated with all edges attached to vertex a . $b \in a$ (or conversely $a \in b$) indicates that the vertices b and a share an actual edge of the graph, $(ab) \in \mathcal{E}_0$. The marginal probabilities, e.g. associated with edges and vertices,

$$p_a(\sigma_a) \equiv \sum_{\sigma \setminus \sigma_a} p(\sigma), \quad p_{ab}(\sigma_{ab}) \equiv \sum_{\sigma \setminus \sigma_{ab}} p(\sigma), \quad (2)$$

constitute what one normally needs to evaluate in order to solve a statistical inference problem. The marginal probabilities can be also expressed in terms of derivatives of the so-called equilibrium free energy, $\mathcal{F}_{C_0} = -\ln Z_{C_0}$, with respect to relevant parameters of the factor functions.

I. GAUGE-INVARIANT FORMULATION OF LOOP CALCULUS

Formally, loop calculus suggests an explicit decomposition of the partition function Z_{C_0} in terms of a sum over certain loops on the graph C_0 . Below we re-derive the loop calculus in more general terms compared to [1], [2].

We start with an observation that the partition function, Z_{C_0} , is invariant with respect to a group of linear gauge transformations of the factor functions

$$f_a(\sigma_a = (\sigma_{ab}, \dots)) \rightarrow \sum_{\sigma'_{ab}} G_{ab}(\sigma_{ab}, \sigma'_{ab}) f_a(\sigma'_{ab}, \dots), \quad (3)$$

described by $\hat{G} = \{G_{ab}(\sigma_{ab}, \sigma'_{ab}); (ab) \in \mathcal{E}_0\}$ provided the pairs of conjugated matrices G_{ab} and G_{ba} are related to each other by the special constraint

$$\sum_{\sigma_{ab}} G_{ab}(\sigma_{ab}, \sigma') G_{ba}(\sigma_{ab}, \sigma'') = \delta(\sigma', \sigma''), \quad (4)$$

where $\delta(x, y)$ is 1 if $x = y$ and 0, otherwise. Except as prescribed by Eq. (4), the gauges are chosen independently at different edges of the graphs. This local freedom in selecting

\hat{G} is the key to our further analysis of the partition function, Z_{C_0} now expressed as

$$Z_{C_0} = \sum_{\sigma} \prod_a \left(\sum_{\sigma'_a} f_a(\sigma'_a) \prod_{b \in a} G_{ab}(\sigma_{ab}, \sigma'_{ab}) \right) \equiv \sum_{\sigma} \bar{p}\{\hat{G}|\sigma\} \equiv \text{Tr} \left(\bar{p}\{\hat{G}|\sigma\} \right), \quad (5)$$

where $\sigma_{ab} = \sigma_{ba}$. We will refer to summation over all allowed configurations of σ in Eq. (5) as computing a graphic trace: a conventional trace can be considered as a special case of the graphic trace for a graph that consists of a single vertex and a single edge. Our next step in evaluation of Eq. (5) is fixing the gauges, which means imposing constraints on \hat{G} in addition to Eq. (4).

It is convenient to distinguish a special term in the sum/trace over σ in Eq. (5) with all $\sigma_{ab} = 0$. We will refer to this term as the ground or, alternatively, uncolored state (term), while all the other terms in the sum, which contain at least one edge with $\sigma_{ab} > 0$, are called excited (colored) states. Obviously for a general gauge choice \hat{G} all kinds of excited states, e.g. with only one edge being excited/colored, provide nonzero contributions to Z . Discussing individual terms in the σ -sum in Eq. (5) we call a vertex colored if at least one edge attached to it is excited/colored.

A BP-gauge corresponds to such a special choice of \hat{G} that makes vanish any contribution in the σ -sum in Eq. (5) that has at least one vertex with only one attached colored edge. Stated differently a BP-gauge prohibits loose excited/colored edges at any vertex. Formally it is expressed as the following set of conditions

$$\sum_{\sigma'_a} f_a(\sigma'_a) G_{ab}^{(bp)}(\sigma_{ab} \neq 0, \sigma'_{ab}) \prod_{c \in a}^{c \neq b} G_{ac}^{(bp)}(0, \sigma'_{ac}) = 0, \quad (6)$$

enforced independently at any vertex of the graph. Combined with the constraints (4), Eq. (6) can be re-stated in the vector form depending only on the ground state part of the gauges:

$$G_{ba}^{(bp)}(0, \sigma'_{ab}) = \rho_a^{-1} \sum_{\sigma'_a \setminus \sigma'_{ab}} f_a(\sigma'_a) \prod_{c \in a}^{c \neq b} G_{ac}^{(bp)}(0, \sigma'_{ac}). \quad (7)$$

with

$$\rho_a = \sum_{\sigma'_a} f_a(\sigma'_a) \prod_{c \in a} G_{ac}^{(bp)}(0, \sigma'_{ac}). \quad (8)$$

We can alternatively derive Eq. (7) for BP gauges using a variational approach. To that end we introduce a functional

$$\mathcal{Z}_0(\hat{\epsilon}) \equiv \bar{p}\{G|\mathbf{0}\} = \prod_a \rho_a(\epsilon_a), \quad (9)$$

where $\epsilon_{ab}(\sigma_{ab}) \equiv G_{ab}(0, \sigma_{ab})$, $\epsilon_a \equiv \{\epsilon_{ab}|b \in a, \mathcal{V}_0\}$, $\hat{\epsilon} \equiv \{\epsilon_{ab}|(ab) \in \mathcal{E}_0\}$, and $\rho_a(\epsilon_a)$ is given by Eq. (8) with $G_{ac}^{(bp)}(0)$ replaced by ϵ_{ac} . The conditions for the stationary points of $\mathcal{F}_0 \equiv -\ln \mathcal{Z}_0$ with respect to $\hat{\epsilon}$ under the corresponding (0-related) part of the orthogonality constraints (4) recovers Eqs. (7). Note that the functional $\mathcal{F}_0(\hat{\epsilon})$ as well as the

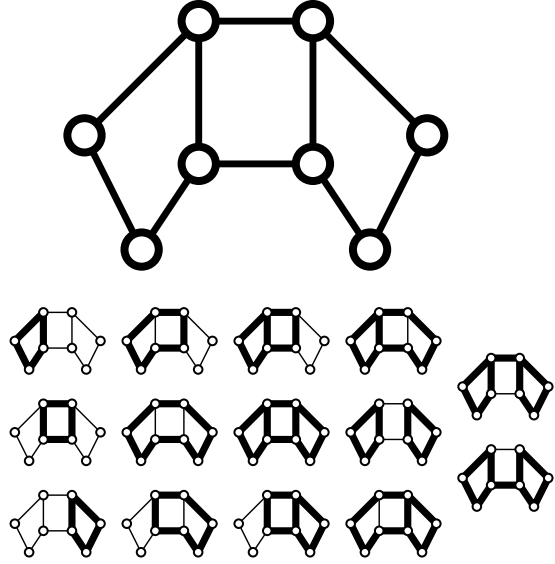


Fig. 1. Example of a factor graph, C_0 with fourteen possible generalized loops, $\Omega(C_0) = \{C_1\}$, shown in bold on the bottom.

BP equations (7,8) possess some remaining irrelevant gauge freedom with respect to a set of transformations $\epsilon_{ab} \rightarrow \kappa_{ab} \epsilon_{ab}$ with $\kappa_{ab} \kappa_{ba} = 1$. Stated differently, the BP equations fix only the relevant part of the gauge freedom. A connection between the functional $\mathcal{F}_0(\hat{\epsilon})$ and the variational Bethe free energy will be established in Section III.

The conventional form of the BP equations, in terms of the “messages” $\eta_{ab}(\sigma_{ab})$,

$$\frac{\exp \left(\eta_{ab}^{(bp)}(\sigma_{ab}) \right)}{\sum_{\sigma_{ab}} \exp \left(\eta_{ab}^{(bp)}(\sigma_{ab}) + \eta_{ba}^{(bp)}(\sigma_{ab}) \right)} = \frac{\sum_{\sigma_a \setminus \sigma_{ab}} f_a(\sigma_a) \exp \left(\sum_{b \in a} \eta_{ab}^{(bp)}(\sigma_{ab}) \right)}{\sum_{\sigma_a} f_a(\sigma_a) \exp \left(\sum_{b \in a} \eta_{ab}^{(bp)}(\sigma_{ab}) \right)}, \quad (10)$$

is recovered using the following parametrization

$$\epsilon_{ab} = G_{ab}(0, \sigma) = \frac{\exp(\eta_{ab}(\sigma))}{\sum_{\sigma} \exp(\eta_{ab}(\sigma) + \eta_{ba}(\sigma))}. \quad (11)$$

Our discussion has been applied so far to the case of a general q -ary alphabet. We now turn to the simplest binary case $q = 2$, where the ground state parametrization (11) unambiguously fixes the excited states: $G_{ab}(1, \sigma) = (1 - 2\sigma)G_{ba}(0, (\sigma - 1)^2)$. Substituting the latter expression and Eqs. (6,7) into Eq. (5) we arrive at the main formula of the loop calculus for the binary alphabet

$$Z_{C_0} = Z_{0;C_0} \left(1 + \sum_{C_1} r(C_1) \right), \quad r(C_1) \equiv Z_{0;C_0}^{-1} \bar{p}\{G|\sigma_{C_1}\},$$

$$Z_{0;C_0} \equiv \bar{p}\{G|\sigma_0\}, \quad \sigma_0 \equiv \{\sigma_{ab} = 0 | (ab) \in C_0\},$$

$$\sigma_{C_1} \equiv \left\{ \begin{array}{ll} \sigma_{ab} = 1 & (ab) \in C_1 \\ \sigma_{ab} = 0 & (ab) \in C_0 \setminus C_1 \end{array} \right\}.$$

where $\{C_1\} = \Omega(C_0)$ is the set of generalized loops on the graph, defined as subgraphs of C_0 without loose ends, i.e.

with degree of connectivity at any vertex (within the subgraph) being two or larger.

Beliefs are defined here as substitutes for the exact marginal probabilities (2) truncated at the first, ground state, term

$$b_{ab}^{(bp)}(\sigma_{ab}) = G_{ab}^{(bp)}(0, \sigma_{ab}), \quad (12)$$

$$b_a^{(bp)}(\sigma_a) = \frac{f_a(\sigma_a) \prod_{b \in a} G_{ab}^{(bp)}(0, \sigma_{ab})}{\sum_{\sigma_a} f_a(\sigma_a) \prod_{b \in a} G_{ab}^{(bp)}(0, \sigma_{ab})}. \quad (13)$$

Then a single generalized loop contribution, r_{C_1} , is expressed in terms of the ground state beliefs in the following simple way

$$r(C_1) = \frac{\prod_{a \in C_1} \mu_a}{\prod_{(ab) \in C_1} (1 - m_{ab}^2)}, \quad m_{ab} \equiv \sum_{\sigma_{ab}} \sigma_{ab} b_{ab}^{(bp)}(\sigma_{ab}),$$

$$\mu_a \equiv \sum_{\sigma_a} \left(\prod_{b \in a, C_1} (\sigma_{ab} - m_{ab}) \right) b_a^{(bp)}(\sigma_a).$$

The loop calculus construction for a simple example is illustrated schematically in Fig. 1.

II. LOOP TOWER FOR q -ARY ALPHABET

Turning to the general q -ary alphabet case we first notice that all considerations and formulas of the introduction and the first part of Section I, all the way up to Eq. (7), actually apply to the general q -ary case. Partitioning the sum/trace over σ in Eq. (5) into the ground-state term, with σ_0 , and the remaining excited-state terms $\{\sigma \setminus \sigma_0\}$, and emergence of the self-consistent set of equations for the ground state gauges (11) are important general features of the gauge fixing construction. Of course, all the preceding formulas should be understood in terms of the edge variables that assume values from $\{0, \dots, q-1\}$. Generalization of Eq. (14) to a general q -ary alphabet reads

$$Z_{C_0} = Z_{0;C_0} + \sum_{C_1 \in \Omega(C_0)} Z_{C_1}, \quad Z_{C_1} = \sum_{\sigma_{C_1}} \bar{p}(G^{(bp)} | \sigma_{C_1}). \quad (14)$$

The additional summation over the colored/excited σ_{C_1} in Eq. (14) is a consequence of the fact that for $q > 2$, σ_{C_1} , is not fixed unambiguously, but rather represents summation over the reduced $(q-1)$ -colors rich set, $1, \dots, q-1$. The BP-gauges for the original graphical model are described by Eqs. (4,6). The set of excited states gets larger in the q -ary case and, consequently, there is a big freedom in selecting the orthogonal basis set of excited gauges. Selecting one such solution of Eqs. (4,6), $\{G_{ab;C_0}^{(bp)}(\sigma_{ab}, \sigma'_{ab}); (ab) \in C_0\}$, and substituting it in Eq. (14) we find that Z_{C_1} becomes the partition function of a reduced graphical model, defined on a subset $C_1 \subset C_0$ of C_0 ,

$$Z_{C_1} = \sum_{\sigma_{C_1}} \prod_{a \in C_1} f_{1;a}(\sigma_{a;C_1}), \quad f_{1;a}(\sigma_{a;C_1}) = \quad (15)$$

$$= \sum_{\sigma'_a} f_a(\sigma'_a) \prod_{b \in a, C_0} G_{ab;C_0}^{(bp)}(\sigma_{ab}, \sigma'_{ab}) \prod_{b \in a, C_0}^{b \notin C_1} \delta(\sigma_{ab}, 0).$$

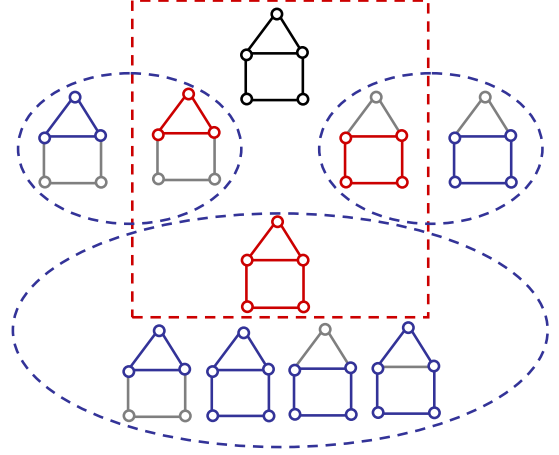


Fig. 2. Example of a loop tower construction for three colors, $q = 0, 1, 2$, shown in the Figure in black, red and blue respectively. First layer of the tower is bounded by the red dashed box, with the original graph, C_0 , shown in black and three generalized loops, $\{C_1\} = \Omega(C_0)$, shown in red. On the second layer of the tower each graph from $\Omega(C_0)$ generates its own set of generalized loops. The next layer of generalized loops, shown in blue, are bounded by three dashed blue boxes with red graph in a box showing respective element of $\{C_1\}$.

Here $\sigma_{a;C_1}$ is the vector constructed of σ_{ab} with $b \in C_1$, with the components labeled by $\{1, \dots, q-1\}$. Z_{C_1} may be understood as a partition function of a reduced graphical model, defined on the graph C_1 in terms of a reduced (one element shorter) alphabet and with the factor functions $f_{1;a}$.

This reformulation of the partition function of the original problem in terms of a sum of partition functions over reduced graphical problems can be repeated sequentially: $Z_{C_0} \rightarrow Z_{C_1} \rightarrow Z_{C_2} \rightarrow \dots \rightarrow Z_{C_{q-1}}$ where $C_0 \supset C_1 \supset \dots \supset C_{q-2}$ is the tower of loops and Z_{C_j} is the partition function of the graphical model defined in terms of a $(q-j)$ -ary variables on the graph C_j , which is a generalized loop of C_{j-1} . All together one arrives at Eq. (14) supplemented by the sequence

$$j=1, \dots, q-2: \quad Z_{C_j} = Z_{0;C_j} + \sum_{C_{j+1} \in \Omega(C_j)} Z_{C_{j+1}}. \quad (16)$$

Generalization of Eq. (15) becomes

$$Z_{C_j} = \sum_{\sigma_{C_j}} \prod_{a \in C_j} f_{j;a}(\sigma_{a;C_j}), \quad (17)$$

$$f_{j;a}(\sigma_{a;C_j}) = \sum_{\sigma'_{a;C_{j-1}}} f_{j-1;a}(\sigma'_{a;C_{j-1}})$$

$$\times \prod_{b \in a, C_{j-1}} G_{ab;C_{j-1}}^{(bp)}(\sigma_{ab}, \sigma'_{ab}) \prod_{b \in a, C_{j-1}}^{b \notin C_j} \delta(\sigma_{ab}, j-1),$$

where σ_{C_j} is a vector constructed out of the variables defined on all edges of the graph C_j with the components labeled by $\{j, \dots, q-1\}$. The BP gauges in Eqs. (17), $G_{ab;C_j}^{(bp)}$, are solutions of Eqs. (4,6) with the original factor functions, $f = f_0$ replaced by f_j .

III. RELATION TO THE BETHE FREE ENERGY APPROACH

It is known that the exact (equilibrium) free energy of any classical statistical model can be obtained from a variational principle based on an exact non-equilibrium variational functional of the full belief, $b(\sigma)$,

$$\mathcal{F}_{exact}\{b(\sigma)\} = \sum_{\sigma} b(\sigma) \ln \frac{b(\sigma)}{\prod_a f_a(\sigma_a)}. \quad (18)$$

The only stationary point of the functional under the normalization condition

$$\sum_{\sigma} b(\sigma) = 1, \quad (19)$$

reproduces the probability distribution $b(\sigma) = p(\sigma)$, where $p(\sigma)$ is defined by Eq. (1). This stationary point is actually a minimum. The value of the exact variational functional at its minimum is equal to the exact free energy

$$\mathcal{F}_{exact}\{p(\sigma)\} = F = -\ln Z, \quad (20)$$

where $Z = Z_{C_0}$, the latter defined above by Eq. (1). Hereafter we skip the graph, C_0 , index to simplify notations.

Introducing an approximate variational ansatz

$$b(\sigma) \approx \frac{\prod_a b_a(\sigma_a)}{\prod_{(ab)} b_{ab}(\sigma_{ab})}, \quad (21)$$

where b_a and b_{ab} are approximations for the corresponding (exact) marginal probabilities, we substitute it (in the spirit of [3]) into Eq. (18). We further invoke another approximation (both approximations are actually exact in the case of a tree, i.e. a graph with no loops)

$$b_a(\sigma_a) \approx \sum_{\sigma \setminus \sigma_a} b(\sigma), \quad b_{ab}(\sigma_{ab}) \approx \sum_{\sigma \setminus \sigma_{ab}} b(\sigma). \quad (22)$$

This results in the so-called Bethe (approximate) free energy functional of beliefs $b_a(\sigma_a), b_{ac}(\sigma_{ac})$:

$$\begin{aligned} \Phi_{Bethe} = & \sum_a \sum_{\sigma_a} b_a(\sigma_a) \ln \left(\frac{b_a(\sigma_a)}{f_a(\sigma_a)} \right) \\ & - \sum_{(ab)} \sum_{\sigma_{ab}} b_{ab}(\sigma_{ab}) \ln b_{ab}(\sigma_{ab}). \end{aligned} \quad (23)$$

We require the beliefs to obey the positivity, normalizability and compatibility constraints, the features borrowed from the corresponding exact probabilities given by Eqs. (2). Thus, we have $\forall a, c; c \in a$ (and inversely $a \in c$):

$$0 \leq b_a(\sigma_a), b_{ac}(\sigma_{ac}) \leq 1, \quad (24)$$

$$\sum_{\sigma_a} b_a(\sigma_a) = 1, \quad \sum_{\sigma_{ab}} b_{ab}(\sigma_{ab}) = 1, \quad (25)$$

$$b_{ac}(\sigma_{ac}) = \sum_{\sigma_a \setminus \sigma_{ac}} b_a(\sigma_a), \quad b_{ac}(\sigma_{ca}) = \sum_{\sigma_c \setminus \sigma_{ca}} b_c(\sigma_c). \quad (26)$$

To establish a connection between the Bethe free energy and the functional \mathcal{F}_0 we introduce the effective Lagrangian

$$\begin{aligned} \mathcal{L}_{Bethe} = & \sum_a \sum_{\sigma_a} b_a(\sigma_a) \ln \left(\frac{b_a(\sigma_a)}{f_a(\sigma_a)} \right) \\ & - \sum_{(ab)} \sum_{\sigma_{ab}} b_{ab}(\sigma_{ab}) \ln b_{ab}(\sigma_{ab}) \\ & + \sum_{(ab)} \left(\sum_{\sigma_{ab}} \ln(\varepsilon_{ab}(\sigma_{ab})) \right) \left(b_{ab}(\sigma_{ab}) - \sum_{\sigma_a \setminus \sigma_{ab}} b_a(\sigma_a) \right) \\ & + \sum_{\sigma_{ba}} \ln(\varepsilon_{ba}(\sigma_{ba})) \left(b_{ba}(\sigma_{ba}) - \sum_{\sigma_b \setminus \sigma_{ba}} b_b(\sigma_b) \right), \end{aligned} \quad (27)$$

that depends on all beliefs that satisfy the normalization constraints (25) with no constraints on $\varepsilon_{ab}(\sigma_{ab})$. Requiring vanishing of the variation with respect to ε_{ab} obviously leads to the constraints given by Eq. (26), and once all constraints are fulfilled the functional does not depend on ε_{ab} (which should be considered as gauge symmetry) and coincides with Φ_{Bethe} as a function of the beliefs. This implies a one-to-one correspondence between the extrema of \mathcal{L}_{Bethe} and Bethe free energy Φ_{Bethe} .

Finding extrema of \mathcal{L}_{Bethe} with respect to the beliefs (this can be technically achieved by introducing Lagrange multipliers for the set of constraints (25)) leads to beliefs that depend explicitly on $\hat{\varepsilon} \equiv \{\varepsilon_{ab} | (ab) \in \mathcal{E}_0\}$:

$$b_a^{(*)}(\sigma_a) = (\varrho_a(\varepsilon_a))^{-1} f_a(\sigma_a) \prod_{b \in a} \varepsilon_{ab}(\sigma_{ab}) \quad (28)$$

$$b_{ab}^{(*)}(\sigma_{ab}) = \varrho_{ab}^{-1}(\varepsilon_{ab}, \varepsilon_{ba}) \varepsilon_{ab}(\sigma_{ab}) \varepsilon_{ba}(\sigma_{ab}), \quad (29)$$

$$\varrho_a(\varepsilon_a) \equiv \sum_{\sigma_a} f_a(\sigma_a) \prod_{c \in a} \varepsilon_{ac}(\sigma_{ac}), \quad (30)$$

$$\varrho_{ab}(\varepsilon_{ab}, \varepsilon_{ba}) \equiv \sum_{\sigma_{ab}} \varepsilon_{ab}(\sigma_{ab}) \varepsilon_{ba}(\sigma_{ab}), \quad (31)$$

where $\varepsilon_a = \{\varepsilon_{ab}; a \in b, \mathcal{V}_0\}$. Substituting the values of beliefs given by Eqs. (28,29) into Eq. (27) results in a functional that depends on the $\hat{\varepsilon}$ variables only

$$\mathcal{F}_B(\hat{\varepsilon}) = - \sum_a \ln \varrho_a(\varepsilon_a) + \sum_{(ab)} \ln (\varrho_{ab}(\varepsilon_{ab}, \varepsilon_{ba})). \quad (32)$$

The functional \mathcal{F}_B possesses strong gauge symmetry: it is invariant under a set of transformations $\varepsilon_{ab} \rightarrow \kappa_{ab} \varepsilon_{ab}$. The gauge can be partially fixed by implementing a gauge (normalization) condition

$$\sum_{\sigma_{ab}} \varepsilon_{ab}(\sigma_{ab}) \varepsilon_{ba}(\sigma_{ab}) = 1. \quad (33)$$

Implementing this constraint, the second term in Eq. (32) vanishes. This means that switching from the notations of Section I to our current notations, $\epsilon \rightarrow \varepsilon$ and $\rho \rightarrow \varrho$, we arrive at $\mathcal{F}_0 = \mathcal{F}_B$. Stated more formally, \mathcal{F}_0 introduced earlier represents the gauge-invariant functional \mathcal{F}_B in a particular gauge determined by Eq. (33). This implies a one-to-one correspondence of the extrema of \mathcal{F}_0 to the extrema of \mathcal{L}_{Bethe} , and therefore to the extrema of the Bethe free energy Φ_{Bethe} .

IV. DISCUSSIONS AND CONCLUSIONS

We first summarize the results presented in the manuscript. We have introduced a group of gauge transformations that keep the partition function of the graphical model invariant, and naturally split the gauges into the “ground” and “excited” parts. The partition function is decomposed into the principal ground and many excited terms. Each excited contribution is interpreted in terms of an excited subgraph constructed from excited edges. Requiring that only excited subgraphs with no loose ends contribute to the partition function sets the BP equations for the ground gauges. We show that the BP equations can be derived using a variational principle for the partition function as a function of the ground gauges. Further consideration differs for the binary and q -ary alphabets. In the binary case the excited gauges are fixed unambiguously, generating the binary loop series over generalized loops for the partition function [1], [2], [6]. In the q -ary case we pick one (of many possible) excited gauges and presenting the full partition function as a sum over generalized loops. Each contribution labeled by a generalized loop can be viewed as a new graphical model defined on this loop with a new set of factor functions. The loop decomposition procedure is applied again, introducing new ground and excited gauges, fixing the gauges, etc. The procedure repeated for $(q - 1)$ layers builds a q -store loop tower. Finally, we showed that the BP-gauges can be determined using a variational principle and related the corresponding functional \mathcal{F}_0 to the Bethe free energy functional constructed in the spirit of [3].

These results open new venues for further development, and also raise a set of important and challenging questions listed below. (1) Already the lowest level BP equations in the loop tower, the ground BP-gauge may have multiple solutions. Our construction applied to different solutions will generate different loop decompositions for the partition function. The question is, whether a preferred solution is in a way better than the others? Naive intuition suggests that BP gauge with the highest value of Z_{C_0} would serve better. (2) Furthermore, in the case of a q -ary alphabet with $q > 2$ positivity of the factor functions at higher tower levels is not guaranteed. The positivity would be desirable for interpreting the auxiliary graphical problems as some actual statistical inference problems, with the factor functions related to probabilities. On the other hand, there is a big freedom in selecting the excited gauges, and a question surfaces: could one select the excited gauges in a way to guarantee positivity of the higher-level factor functions? (3) The BP ground state contribution to the partition function, Z_{C_0} is positive by construction, however the signs of the excited terms can alternate. This raises a couple of important questions. How do the signs of the loop terms depend on the factor functions and the graphical model itself? Based on our previous experiments [6], we know that emergence of an excited loop contribution comparable to the ground state alerts for a possible failure of BP as an approximation to exact inference. How exactly does the sign alternation and relative value of the tower loop contributions affect success

or failure of BP as an approximation? (4) The equilibrium Bethe free energy estimates the value of the partition function, however the variational derivation sketched in Section III does not guarantee that the resulting \mathcal{F}_0 is actually larger than the exact \mathcal{F} . Indeed in the transition from Eqs. (18,21,22) to Eqs. (23) we further discuss the latter formulation completely ignoring the fact that the conditions (22) can be violated for the resulting BP solutions. How does this violation affect the relation, $\mathcal{F} \geq \mathcal{F}_0$, and what are the consequences of this inequality for the loop series?

We conclude with mentioning some future research directions. As demonstrated in [6], the loop calculus is suggestive of an efficient truncation of the full series that can potentially improve the BP approximation. This idea can also be extended to the q -ary alphabet case, with the tower truncated at some relatively low level. This approach can obviously find interesting application in decoding of non-binary codes and also in problems, such as computer vision, that require a multi-valued data reconstruction. The loop tower approach can also be extended to the analogous case of continuous alphabet. In this case the ground state gauges satisfy a set of integral equations, while the ground and excited states that define the gauges become elements of functional infinite-dimensional Hilbert spaces, which makes the tower heights unlimited and the tower loop decomposition turns into an infinite series. Finally, we note that the gauge conditions may be chosen in some other non-BP way. BP-gauge is of a special importance for dilute locally tree-like graphs simply because in the loop-free case the whole loop hierarchy (the entire loop tower) disappears. One could conjecture that for some other classes of graphical models, e.g. those naturally defined on regular lattices, similar cancelations can take place for some alternative specially selected gauges.

The work at Los Alamos was carried out under the auspices of the National Nuclear Security Administration of the U.S. Department of Energy at Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396. VYC also acknowledges the support through the start-up grant from Wayne State University.

REFERENCES

- [1] M. Chertkov, V. Chernyak, Loop Calculus in Statistical Physics and Information Science, Phys. Rev. E **73**, 065102(R) (2006); cond-mat/0601487.
- [2] M. Chertkov, V. Chernyak, *Loop series for discrete statistical models on graphs*, J. Stat. Mech. (2006) P06009, cond-mat/0603189.
- [3] J.S. Yedidia, W.T. Freeman, Y. Weiss, *Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms*, IEEE IT **51**, 2282 (2005).
- [4] G. D. Forney, *Codes on Graphs: Normal Realizations*, IEEE IT **47**, 520-548 (2001).
- [5] H.-A. Loeliger, *An Introduction to Factor Graphs*, IEEE Signal Processing Magazine, Jan 2001, p. 28-41.
- [6] M. Chertkov, V. Chernyak, *Loop Calculus Helps to Improve Belief Propagation and Linear Programming Decodings of Low-Density-Parity-Check Codes*, invited talk at 44th Allerton Conference (September 27-29, 2006, Allerton, IL), arXiv:cs.IT/0609154.