Journal of Machine Learning Research 7 (2006) 2031-2064 Submitted 4/06; Revised 8/06; Published 10/06

Walk-Sums and Belief Propagation in Gaussian Graphical Models*

Dmitry M. Malioutov DMM@MIT.EDU
Jason K. Johnson JASONJ@MIT.EDU
Alan S. Willsky WILLSKY@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Michael I. Jordan

Abstract

We present a new framework based on walks in a graph for analysis and inference in Gaussian graphical models. The key idea is to decompose the correlation between each pair of variables as a sum over all walks between those variables in the graph. The weight of each walk is given by a product of edgewise partial correlation coefficients. This representation holds for a large class of Gaussian graphical models which we call walk-summable. We give a precise characterization of this class of models, and relate it to other classes including diagonally dominant, attractive, non-frustrated, and pairwise-normalizable. We provide a walk-sum interpretation of Gaussian belief propagation in trees and of the approximate method of loopy belief propagation in graphs with cycles. The walk-sum perspective leads to a better understanding of Gaussian belief propagation and to stronger results for its convergence in loopy graphs.

Keywords: Gaussian graphical models, walk-sum analysis, convergence of loopy belief propagation

1. Introduction

We consider multivariate Gaussian distributions defined on undirected graphs, which are often referred to as Gauss-Markov random fields (GMRFs). The nodes of the graph denote random variables and the edges capture the statistical dependency structure of the model. The family of all Gauss-Markov models defined on a graph is naturally represented in the *information form* of the Gaussian density. The key parameter of the information form is the *information matrix*, which is the inverse of the covariance matrix. The information matrix is sparse, reflecting the structure of the defining graph such that only the diagonal elements and those off-diagonal elements corresponding to edges of the graph are non-zero.

Given such a model, we consider the problem of computing the mean and variance of each variable, thereby determining the marginal densities as well as the mode. In principle, these can be obtained by inverting the information matrix, but the complexity of this computation is cubic in the number of variables. More efficient recursive calculations are possible in graphs with very sparse

*. This paper elaborates upon our earlier brief publication (Johnson, Malioutov, and Willsky, 2006) and presents subsequent developments. This research was supported by the Air Force Office of Scientific Research under Grants FA9550-04-1-0351, FA9550-06-1-0324, and the Army Research Office under Grant W911NF-05-1-0207. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Air Force or Army.

structure—for example, in chains, trees and in graphs with "thin" junction trees. For these models, belief propagation (BP) or its junction tree variants efficiently compute the marginals (Pearl, 1988; Cowell et al., 1999). In large-scale models with more complex graphs, for example, for models arising in oceanography, 3D-tomography, and seismology, even the junction tree approach becomes computationally prohibitive. Iterative methods from numerical linear algebra (Varga, 2000) can be used to compute the marginal means. However, in order to efficiently compute both means and variances, approximate methods such as loopy belief propagation (LBP) are needed (Pearl, 1988; Yedidia, Freeman, and Weiss, 2003; Weiss and Freeman, 2001; Rusmevichientong and Van Roy, 2001). Another important motivation for using LBP, emphasized for example by Moallemi and Van Roy (2006a), is its distributed nature which is important for applications such as sensor networks. While LBP has been shown to often provide good approximate solutions for many problems, it is not guaranteed to do so in general, and may even fail to converge.

In prior work, Rusmevichientong and Van Roy (2001) analyzed Gaussian LBP on the turbo-decoding graph. For this special case they established that variances converge, means follow a linear system upon convergence of the variances, and that if means converge then they are correct. Weiss and Freeman (2001) analyzed LBP from the computation tree perspective to give a sufficient condition (equivalent to diagonal dominance of the information matrix) for convergence, and also showed correctness of the means upon convergence. Wainwright et al. (2003) introduced the tree reparameterization view of belief propagation and, in the Gaussian case, also showed correctness of the means upon convergence. Convergence of other forms of LBP are analyzed by Ihler et al. (2005), and Mooij and Kappen (2005), but unfortunately their sufficient conditions are not directly applicable to the Gaussian case.

We develop a "walk-sum" formulation for computation of means, variances and correlations as sums over certain sets of weighted walks in a graph.[1,2] This walk-sum formulation applies to a wide class of Gauss-Markov models which we call *walk-summable*. We characterize the class of walk-summable models and show that it contains (and extends well beyond) some "easy" classes of models, including models on trees, attractive, non-frustrated, and diagonally dominant models. We also show the equivalence of walk-summability to the fundamental notion of pairwise-normalizability, and that inference in walk-summable models can be reduced to inference in an attractive model based on a certain extended graph.

We use the walk-sum formulation to develop a new interpretation of BP in trees and of LBP in general. Based on this interpretation we are able to extend the previously known sufficient conditions for convergence of LBP to the class of walk-summable models. Our sufficient condition is stronger than that given by Weiss and Freeman (2001) as the class of diagonally dominant models is a strict subset of the class of pairwise-normalizable models. Our results also explain why they did not find any examples where LBP does not converge. The reason is that they presumed pairwise-normalizability. We also give a new explanation, in terms of walk-sums, of why LBP converges to the correct means but not to the correct variances. The reason is that LBP captures all of the walks needed to compute the means but only computes a subset of the walks needed for the variances.

1. After submitting the paper we became aware of a related decomposition for non-Gaussian classical spin systems in statistical physics developed by Brydges et al. (1983). Similarly to our work, the decomposition is connected to the Neumann series expansion of the matrix inverse, but in addition to products of edge weights, their weight of a walk includes a complicated multi-dimensional integral.
2. Another interesting decomposition of the covariance in Gaussian models in terms of *path sums* has been proposed in Jones and West (2005). It is markedly different from our approach (e.g., unlike paths, walks can cross an edge multiple times, and the weight of a path is rather hard to calculate, as opposed to our walk-weights).

In general, walk-summability is not necessary for LBP convergence. Hence, we also provide a tighter (essentially necessary) condition for convergence of LBP *variances* based on a weaker form of walk-summability defined on the LBP computation tree. This provides deeper insight into why LBP can fail to converge—because the LBP computation tree is not always well-posed—which suggests connections to Tatikonda and Jordan (2002).

In related work, concurrent with Johnson et al. (2006), Moallemi and Van Roy (2006a) have shown convergence of their consensus propagation algorithm, which uses a pairwise-normalized model. In this paper, we demonstrate the equivalence of pairwise-normalizability and walk-summability, which suggests a connection between their results and ours. In their more recent work (Moallemi and Van Roy, 2006b), concurrent with this paper, they make use of our walk-sum analysis of LBP, assuming pairwise-normalizability, to consider other initializations of the algorithm.[3] However, the critical condition is still walk-summability, which is presented in this paper.

In Section 2 we introduce Gaussian graphical models and describe exact BP for tree-structured graphs as well as approximate BP for loopy graphs, and their connection to Gaussian elimination. Next, in Section 3 we describe our walk-based framework for inference, define walk-summable models, and explore the connections between walk-summable models and other subclasses of Gaussian models. We present the walk-sum interpretation of LBP and our conditions for its convergence in Section 4. We discuss non-walksummable models, and tighter conditions for LBP convergence in Section 5. Finally, conclusions and directions for further work are discussed in Section 6. Detailed proofs omitted from the main body of the paper appear in the appendices.

2. Preliminaries

In this section we give a brief background of Gaussian graphical models (Section 2.1) and of Gaussian elimination and its relation to belief propagation (Section 2.2).

2.1 Gaussian Graphical Models

A Gaussian graphical model is defined by an undirected graph $G = (V,E)$, where V is the set of nodes (or vertices) and E is the set of edges (a set of unordered pairs $\{i, j\} \subset V$), and a collection of jointly Gaussian random variables $x = (x_i, i \in V)$. The probability density is given by

$$p(x) \propto \exp\{-\tfrac{1}{2}x^T Jx + h^T x\} \tag{1}$$

where J is a symmetric, positive definite matrix ($J \succ 0$) that is sparse so as to respect the graph G: if $\{i, j\} \notin E$ then $J_{ij} = 0$. The condition $J \succ 0$ is necessary so that (1) defines a *valid* (i.e., normalizable) probability density. This is the *information form* of the Gaussian density. We call J the *information matrix* and h the *potential vector*. They are related to the standard Gaussian parameterization in terms of the mean $\mu \triangleq \mathbb{E}\{x\}$ and covariance $P \triangleq \mathbb{E}\{(x-\mu)(x-\mu)^T\}$ as follows:

$$\mu = J^{-1}h \quad \text{and} \quad P = J^{-1}.$$

This class of densities is precisely the family of non-degenerate Gaussian distributions which are Markov with respect to the graph G (Speed and Kiiveri, 1986): if a subset of nodes $B \subset V$ separates

3. Here, we choose one particular initialization of LBP. However, fixing this initialization does not restrict the class of models or applications for which our results apply. For instance, the application considered by Moallemi and Van Roy (2006a) can also be handled in our framework by a simple reparameterization.

two other subsets $A \subset V$ and $C \subset V$ in G, then the corresponding subsets of random variables x_A and x_C are conditionally independent given x_B. In particular, define the neighborhood of a node i to be the set of its neighbors: $\mathcal{N}(i) = \{j \mid \{i,j\} \in E\}$. Then, conditioned on $x_{\mathcal{N}(i)}$, the variable x_i is independent of the rest of the variables in the graph.

The *partial correlation coefficient* between variables x_i and x_j measures their conditional correlation given the values of the other variables $x_{V\setminus ij} \triangleq (x_k, k \in V \setminus \{i,j\})$. These are computed by normalizing the off-diagonal entries of the information matrix (Lauritzen, 1996):

$$r_{ij} \triangleq \frac{\mathrm{cov}(x_i; x_j | x_{V\setminus ij})}{\sqrt{\mathrm{var}(x_i|x_{V\setminus ij})\mathrm{var}(x_j|x_{V\setminus ij})}} = -\frac{J_{ij}}{\sqrt{J_{ii}J_{jj}}}. \tag{2}$$

Hence, we observe the relation between the sparsity of J and conditional independence between variables. In agreement with the Hammersley-Clifford theorem (Hammersley and Clifford, 1971), for Gaussian models we may factor the probability distribution

$$p(x) \propto \prod_{i\in V} \psi_i(x_i) \prod_{\{i,j\}\in E} \psi_{ij}(x_i, x_j)$$

in terms of node and edge *potential functions*:[4]

$$\psi_i(x_i) = \exp\{-\tfrac{1}{2}A_i x_i^2 + h_i x_i\} \quad \text{and} \quad \psi_{ij}(x_i,x_j) = \exp\{-\tfrac{1}{2}\left[\begin{smallmatrix} x_i & x_j \end{smallmatrix}\right] B_{ij} \left[\begin{smallmatrix} x_i \\ x_j \end{smallmatrix}\right]\}. \tag{3}$$

Here, A_i and B_{ij} must add up to J such that

$$x^T J x = \sum_i A_i x_i^2 + \sum_{\{i,j\}\in E} \left(\begin{smallmatrix} x_i & x_j \end{smallmatrix}\right) B_{ij} \left(\begin{smallmatrix} x_i \\ x_j \end{smallmatrix}\right).$$

The choice of a decomposition of J into such A_i and B_{ij} is not unique: the diagonal elements J_{ii} can be split in various ways between A_i and B_{ij}, but the off-diagonal elements of J are copied directly into the corresponding B_{ij}. It is *not* always possible to find a decomposition of J such that both $A_i > 0$ and $B_{ij} \succ 0$.[5] We call models where such a decomposition exists *pairwise-normalizable*.

Our analysis is not limited to pairwise-normalizable models. Instead we use the decomposition $A_i = J_{ii}$ and $B_{ij} = \left[\begin{smallmatrix} 0 & J_{ij} \\ J_{ij} & 0 \end{smallmatrix}\right]$, which always exists, and leads to the following node and edge potentials:

$$\psi_i(x_i) = \exp\{-\tfrac{1}{2}J_{ii}x_i^2 + h_i x_i\} \quad \text{and} \quad \psi_{ij}(x_i,x_j) = \exp\{-x_i J_{ij} x_j\}. \tag{4}$$

Note that *any* decomposition in (3) can easily be converted to our decomposition (4) using local operations (the required elements of J can be read off by adding overlapping matrices).

We illustrate this framework with a prototypical estimation problem. Suppose that we wish to estimate an unknown signal x (e.g., an image) based on noisy observations y. A commonly used prior model in image processing is the *thin membrane model* $p(x) \propto \exp(\{-\frac{1}{2}((\alpha\sum_i x_i^2 +$

4. To be precise, it is actually the negative logarithms of ψ_i and ψ_{ij} that are usually referred to as potentials in the statistical mechanics literature. We abuse the terminology slightly for convenience.

5. For example the model with $J = \left[\begin{smallmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{smallmatrix}\right]$ is a valid model with $J \succ 0$, but no decomposition into single and pairwise positive definite factors exists. This can be verified by posing an appropriate semidefinite feasibility problem, or as we discuss later through walk-summability.

$\beta \sum_{\{i,j\}\in E}(x_i - x_j)^2))))\}$ where $\alpha, \beta > 0$ and E specifies nearest neighbors in the image. This model is described by a sparse information matrix with $J_{ii} = \alpha + \beta|\mathcal{N}(i)|$ and $J_{ij} = -\beta$ for $\{i, j\} \in E$.

Now, consider local observations y, such that $p(y|x) = \prod_i p(y_i|x_i)$. The distribution of interest is then $p(x|y) \propto p(y|x)p(x)$, which is Markov with respect to the same graph as $p(x)$, but with modified information parameters. For instance, let $y = x + v$ where v is Gaussian distributed measurement noise with zero mean and covariance $\sigma^2 I$. Then $p(x|y) \propto \exp\{-\frac{1}{2}x^T \hat{J} x + h^T x\}$, where $\hat{J} = J + \frac{1}{\sigma^2} I$ and $h = \frac{1}{\sigma^2} y$. Hence, introducing local observations only changes the potential vector h and the diagonal of the information matrix J. Without loss of generality, in subsequent discussion we assume that any observations have already been absorbed into J and h.

2.2 Belief Propagation and Gaussian Elimination

An important inference problem for a graphical model is computing the marginals $p_i(x_i)$, obtained by integrating $p(x)$ over all variables except x_i, for each node i.[6] This problem can be solved very efficiently in graphs that are trees by a form of variable elimination, known as belief propagation, which also provides an approximate method for general graphs.

Belief Propagation in Trees In principle, the marginal of a given node can be computed by recursively eliminating variables one by one until just the desired node remains. Belief propagation in trees can be interpreted as an efficient form of variable elimination. Rather than computing the marginal for each variable independently, we instead compute these together by sharing the results of intermediate computations. Ultimately each node j must receive information from each of its neighbors, where the *message*, $m_{i\to j}(x_j)$, from neighbor i to j represents the result of eliminating all of the variables in the subtree rooted at node i and including all of its neighbors other than j (see Figure 1). Since each of these messages is itself made up of variable elimination steps corresponding to the subtrees rooted at the other neighbors of node i, there is a set of fixed-point equations that relate messages throughout the tree:

$$m_{i\to j}(x_j) = \int \psi_{ij}(x_i, x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i)\ dx_i. \tag{5}$$

Given these fixed-point messages, the marginals are obtained by combining messages at each node,

$$p_i(x_i) \propto \psi_i(x_i) \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i),$$

and normalizing the result.

The equations (5) can be solved in a finite number of steps using a variety of *message schedules*, including one schedule that corresponds roughly to sequential variable elimination and back-substitution (a first pass from leaf nodes toward a common, overall "root" node followed by a reverse pass back to the leaf nodes) and a fully parallel schedule in which each node begins by sending non-informative messages (all $m_{i\to j}$ initially set to 1), followed by iterative computation of (5) throughout the tree. For trees, either message schedule will terminate with the correct values after a finite number of steps (equal to the diameter of the tree in the case of the fully parallel iteration).

6. Another important problem is computation of max-marginals $\hat{p}_i(x_i)$, obtaining by *maximizing* with respect to the other variables, which is useful to determine the mode $\hat{x} = \arg\max p(x)$. In Gaussian models, these are equivalent inference problems because marginals are proportional to max-marginals and the mean is equal to the mode.

As we have discussed, there are a variety of ways in which the information matrix in GMRFs can be decomposed into edge and node potential functions, and each such decomposition leads to BP iterations that are different in detail.[7] In our development we will use the simple decomposition in (4), directly in terms of the elements of J.

For Gaussian models expressed in information form, variable elimination/marginalization corresponds to *Gaussian elimination*.[8] For example, if we wish to eliminate a single variable i to obtain the marginal over $U = V \backslash i$, the formulas yielding the information parameterization for the marginal on U are:

$$\hat{J}_U = J_{U,U} - J_{U,i} J_{ii}^{-1} J_{i,U} \quad \text{and} \quad \hat{h}_U = h_U - J_{U,i} J_{ii}^{-1} h_i.$$

Here $\hat{J}_U$ and $\hat{h}_U$ specify the marginal density on x_U, whereas $J_{U,U}$ and h_U are a submatrix and a subvector of the information parameters on the full graph. The messages in Gaussian models can be parameterized in information form

$$m_{i\to j}(x_j) \triangleq \exp\{-\tfrac{1}{2}\Delta J_{i\to j} x_j^2 + \Delta h_{i\to j} x_j\}, \tag{6}$$

so that the fixed-point equations (5) can be stated in terms of these information parameters. We do this in two steps. The first step corresponds to preparing the message to be sent from node i to node j by collecting information from all of the other neighbors of i:

$$\hat{J}_{i\backslash j} = J_{ii} + \sum_{k\in\mathcal{N}(i)\backslash j} \Delta J_{k\to i} \quad \text{and} \quad \hat{h}_{i\backslash j} = h_i + \sum_{k\in\mathcal{N}(i)\backslash j} \Delta h_{k\to i}. \tag{7}$$

The second step produces the information quantities to be propagated to node j:

$$\Delta J_{i\to j} = -J_{ji} \hat{J}_{i\backslash j}^{-1} J_{ji} \quad \text{and} \quad \Delta h_{i\to j} = -J_{ji} \hat{J}_{i\backslash j}^{-1} \hat{h}_{i\backslash j}. \tag{8}$$

As before, these equations can be solved by various message schedules, ranging from leaf-root-leaf Gaussian elimination and back-substitution to fully parallel iteration starting from the non-informative messages in which all $\Delta J_{i\to j}$ and $\Delta h_{i\to j}$ are set to zero. When the fixed point solution is obtained, the computation of the marginal at each node is obtained by combining messages and local information:

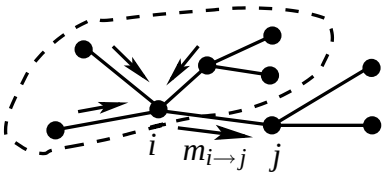
$$\hat{J}_i = J_{ii} + \sum_{k\in\mathcal{N}(i)} \Delta J_{k\to i} \quad \text{and} \quad \hat{h}_i = h_i + \sum_{k\in\mathcal{N}(i)} \Delta h_{k\to i}, \tag{9}$$

which can be easily inverted to recover the marginal mean and variance:

$$\mu_i = \hat{J}_i^{-1} \hat{h}_i \quad \text{and} \quad P_{ii} = \hat{J}_i^{-1}.$$

In general, performing Gaussian elimination corresponds, upto a permutation, to computing an LDL^T factorization of the information matrix—that is, $PJP^T = LDL^T$ where L is lower-triangular, D is diagonal and P is a permutation matrix corresponding to a particular choice of elimination order. This factorization exists if J is non-singular. In trees, the elimination order can be chosen such that at each step of the procedure, the next node eliminated is a leaf node of the remaining subtree. Each node elimination step then corresponds to a message in the "upward" pass of the leaf-root-leaf form

7. One common decomposition for pairwise-normalizable models selects $A_i > 0$ and $B_{ij} \succ 0$ in (3) (Plarre and Kumar, 2004; Weiss and Freeman, 2001; Moallemi and Van Roy, 2006a).
8. The connection between Gaussian elimination and belief propagation has been noted before by Plarre and Kumar (2004), although they do not use the information form.



A message $m_{i\to j}$ passed from node i to node $j \in \mathcal{N}(i)$ captures the effect of eliminating the subtree rooted at i.

Figure 1: An illustration of BP message-passing on trees.

of Gaussian BP. In particular, $D_{ii} = \hat{J}_{i\backslash j}$ at all nodes i except the last (here, j is the parent of node i when i is eliminated) and $D_{ii} = \hat{J}_i$ for that last variable corresponding to the root of the tree. It is clear that $D_{ii} > 0$ for all i if and only if J is positive definite. We conclude that for models on trees, J being positive definite is equivalent to all of the quantities $\hat{J}_{i\backslash j}$ and $\hat{J}_i$ in (7),(9) being positive, a condition we indicate by saying that BP on this tree is *well-posed*. Thus, performing Gaussian BP on trees serves as a simple test for validity of the model. The importance of this notion will become apparent shortly.

Loopy Belief Propagation The message passing formulas derived for tree models can also be applied to models defined on graphs with cycles, even though this no longer corresponds precisely to variable elimination in the graph. This approximation method, called *loopy belief propagation* (LBP), was first proposed by Pearl (1988). Of course in this case, since there are cycles in the graph, only iterative message-scheduling forms can be defined. To be precise, a message schedule $\{\mathcal{M}^{(n)}\}$ specifies which messages $m_{i\to j}^{(n)}$, corresponding to directed edges $(i,j) \in \mathcal{M}^{(n)}$,[9] are updated at step n. The messages in $\mathcal{M}^{(n)}$ are updated using

$$m_{i\to j}^{(n)}(x_j) = \int \psi_{ij}(x_i,x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\backslash j} m_{k\to i}^{(n-1)}(x_i)\ dx_i \tag{10}$$

and $m_{i\to j}^{(n)} = m_{i\to j}^{(n-1)}$ for the other messages. For example, in the fully parallel case *all* messages are updated at each iteration whereas, in serial versions, only one message is updated at each iteration. For GMRFs, application of (10), with messages $m_{i\to j}^{(n)}$ parameterized as in (6), reduces to iterative application of equations (7),(8). We denote the information parameters at step n by $\Delta J_{i\to j}^{(n)}$ and $\Delta h_{i\to j}^{(n)}$. We initialize LBP with non-informative zero values for all of the information parameters in these messages. It is well known that LBP may or may not converge. If it does converge, it will not, in general, yield the correct values for the marginal distributions. In the Gaussian case, however, it is known (Weiss and Freeman, 2001; Rusmevichientong and Van Roy, 2001) that if LBP converges, it yields the correct mean values but, in general, incorrect values for the variances. While there has been considerable work on analyzing the convergence of LBP in general and for GMRFs in particular, the story has been far from complete. One major contribution of this paper is analysis that both provides new insights into LBP for Gaussian models and also brings that story several steps closer to completion.

A key component of our analysis is the insightful interpretation of LBP in terms of the so-called *computation tree* (Yedidia et al., 2003; Weiss and Freeman, 2001; Tatikonda and Jordan, 2002), which captures the structure of LBP computations. The basic idea here is that to each message $m_{i\to j}^{(n)}$

9. For each undirected edge $\{i,j\} \in E$ there are two messages: $m_{i\to j}$ for direction (i,j), and $m_{j\to i}$ for (j,i).

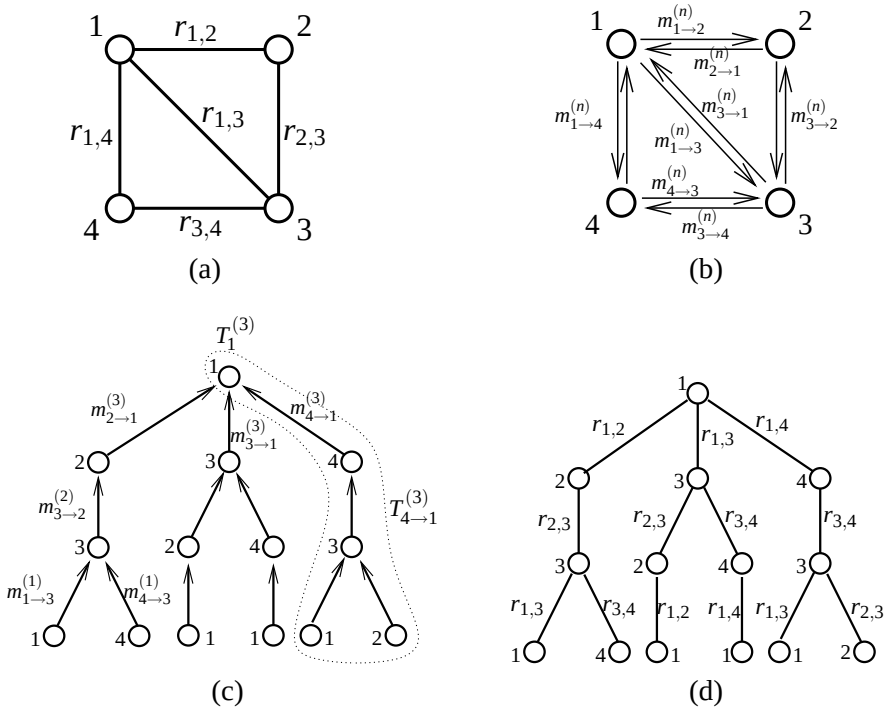


Figure 2: (a) Graph of a Gauss-Markov model with nodes $\{1,2,3,4\}$ and with edge weights (partial correlations) as shown. (b) The parallel LBP message passing scheme. In (c), we show how, after 3 iterations, messages link up to form the computation tree $T_1^{(3)}$ of node 1 (the subtree $T_{4\to 1}^{(3)}$, associated with message $m_{4\to 1}^{(3)}$, is also indicated within the dotted outline). In (d), we illustrate an equivalent Gauss-Markov tree model, with edge weights copied from (a), which has the same marginal at the root node as computed by LBP after 3 iterations.

and marginal estimate $p_i^{(n)}$ there are associated computation trees $T_{i\to j}^{(n)}$ and $T_i^{(n)}$ that summarize their pedigree. Initially, these trees are just single nodes. When message $m_{i\to j}^{(n)}$ is computed, its computation tree $T_{i\to j}^{(n)}$ is constructed by joining the trees $T_{k\to i}^{(n-1)}$, for all neighbors k of i except j, at their common root node i and then adding an additional edge (i,j) to form $T_{i\to j}^{(n)}$ rooted at j. When marginal estimate $p_i^{(n)}$ is computed, its computation tree $T_i^{(n)}$ is formed by joining the trees $T_{k\to i}^{(n-1)}$, for all neighbors k of i, at their common root. Each node and edge of the original graph may be replicated many times in the computation tree, but in a manner which preserves the local neighborhood structure. Potential functions are assigned to the nodes and edges of $T_i^{(n)}$ by copying these from the corresponding nodes and edges of the original loopy graphical model. In this manner, we obtain a Markov tree model in which the marginal at the root node is precisely $p_i^{(n)}$ as computed by LBP. In the case of the fully parallel form of LBP, this leads to a collection of "balanced" computation trees $T_i^{(n)}$ (assuming there are no leaf nodes in G) having uniform depth n, as illustrated in Figure 2. The same construction applies for other message schedules with the only difference being that the resulting computation trees may grow in a non-uniform manner. Our walk-

sum analysis of LBP in Section 6, which relies on computation trees, applies for general message passing schedules.

As we have mentioned, BP on trees, which corresponds to performing Gaussian elimination, is well-posed if and only if J is positive definite. LBP on Gaussian models corresponds to Gaussian elimination in the computation tree, which has its own information matrix corresponding to the unfolding illustrated in Figure 2 and involving replication of information parameters of the original loopy graphical model. Consequently, LBP is well-posed, yielding non-negative variances at each stage of the iteration, if and only if the model on the computation tree is *valid*, that is, if and only if the information matrix for the computation tree is positive definite. Very importantly, this is *not* always the case (even though the matrix J on the original graph is positive definite). The analysis in this paper, among other things, makes this point clear through analysis of the situations in which LBP converges and when it fails to converge.

3. Walk-Summable Gaussian Models

Now we describe our walk-sum framework for Gaussian inference. It is convenient to assume that we have normalized our model (by rescaling variables) so that $J_{ii} = 1$ for all i. Then, $J = I - R$ where R has zero diagonal and the off-diagonal elements are equal to the partial correlation coefficients r_{ij} in (2). We label each edge $\{i, j\}$ of the graph G with partial correlations r_{ij} as edge weights (e.g., see Figures 3 and 5).

3.1 Walk-Summability

A *walk* of length $l \geq 0$ in a graph G is a sequence $w = (w_0, w_1, \ldots, w_l)$ of nodes $w_k \in V$ such that each step of the walk (w_k, w_{k+1}) corresponds to an edge of the graph $\{w_k, w_{k+1}\} \in E$. Walks may visit nodes and cross edges multiple times. We let $l(w)$ denote the length of walk w. We define the *weight* of a walk to be the product of edge weights along the walk:

$$\phi(w) = \prod_{k=1}^{l(w)} r_{w_{k-1},w_k} \ .$$

We also allow zero-length "self" walks $w = (v)$ at each node v for which we define $\phi(w) = 1$. To make a connection between these walks and Gaussian inference, we decompose the covariance matrix using the Neumann power series for the matrix inverse:[10]

$$P = J^{-1} = (I - R)^{-1} = \sum_{k=0}^{\infty} R^k, \quad \text{for} \quad \rho(R) < 1.$$

Here $\rho(R)$ is the spectral radius of R, the maximum absolute value of eigenvalues of R. The power series converges if $\rho(R) < 1$.[11] The (i, j)-th element of R^l can be expressed as a sum of weights of walks w that go from i to j and have length l (denoted $w : i \xrightarrow{l} j$):

$$(R^l)_{ij} = \sum_{w_1,\ldots,w_{l-1}} r_{i,w_1} r_{w_1,w_2} \ldots r_{w_{l-1},j} = \sum_{w : i \xrightarrow{l} j} \phi(w).$$

10. The Neumann series holds for the unnormalized case as well: $J = D - K$, where D is the diagonal part of J. With the weight of a walk defined as $\phi(w) = \prod_{k=1}^{l(w)} K_{w_{k-1},w_k} / \prod_{k=0}^{l(w)} D_{w_k,w_k}$, all our analysis extends to the unnormalized case.
11. Note that $\rho(R)$ can be greater than 1 while $I - R \succ 0$. This occurs if R has an eigenvalue less than -1. Such models are not walk-summable, so the analysis in Section 5 (rather than Section 4.2) applies.

The last equality holds because only the terms that correspond to walks in the graph have non-zero contributions: for all other terms at least one of the partial correlation coefficients $r_{w_k,w_{k+1}}$ is zero. The set of walks from i to j of length l is finite, and the sum of weights of these walks (the walk-sum) is well-defined. We would like to also define walk-sums over arbitrary countable sets of walks. However, care must be taken, as walk-sums over countably many walks may or may not converge, and convergence may depend on the order of summation. This motivates the following definition:

We say that a Gaussian distribution is *walk-summable* (WS) if for all $i, j \in V$ the unordered sum over all walks w from i to j (denoted $w : i \to j$)

$$\sum_{w:i\to j} \phi(w)$$

is well-defined (i.e., converges to the same value for every possible summation order). Appealing to basic results of analysis (Rudin, 1976; Godement, 2004), the unordered sum is well-defined if and only if it *converges absolutely*, that is, if $\sum_{w:i\to j} |\phi(w)|$ converges.

Before we take a closer look at walk-summability, we introduce additional notation. For a matrix A, let $\bar{A}$ be the element-wise absolute value of A, that is, $\bar{A}_{ij} = |A_{ij}|$. We use the notation $A \geq B$ for element-wise comparisons, and $A \succeq B$ for comparisons in positive definite ordering. The following version of the Perron-Frobenius theorem (Horn and Johnson, 1985; Varga, 2000) for non-negative matrices (here $\bar{R} \geq 0$) is used on several occasions in the paper:

Perron-Frobenius theorem There exists a non-negative eigenvector $x \geq 0$ of $\bar{R}$ with eigenvalue $\rho(\bar{R})$. If the graph G is connected (where $r_{ij} \neq 0$ for all edges of G) then $\rho(\bar{R})$ and x are strictly positive and, apart from γx with $\gamma > 0$, there are no other non-negative eigenvectors of $\bar{R}$.

In addition, we often use the following monotonicity properties of the spectral radius:

$$\text{(i)}\ \ \rho(R) \leq \rho(\bar{R}) \quad \text{(ii) If } \bar{R}_1 \leq \bar{R}_2 \ \text{ then } \ \rho(\bar{R}_1) \leq \rho(\bar{R}_2). \tag{11}$$

We now present several equivalent conditions for walk-summability:

Proposition 1 (Walk-Summability) *Each of the following conditions are equivalent to walk-summability:*

(i) $\sum_{w:i\to j} |\phi(w)|$ *converges for all* $i, j \in V$.

(ii) $\sum_l \bar{R}^l$ *converges.*

(iii) $\rho(\bar{R}) < 1$.

(iv) $I - \bar{R} \succ 0$.

The proof appears in Appendix A. It uses absolute convergence to rearrange walks in order of increasing length, and the Perron-Frobenius theorem for part (iv). The condition $\rho(\bar{R}) < 1$ is stronger than $\rho(R) < 1$. The latter is sufficient for the convergence of the walks ordered by increasing length, whereas walk-summability enables convergence to the same answer in arbitrary order of summation. Note that (iv) implies that the model is walk-summable if and only if we can replace all negative partial correlation coefficients by their absolute values and still have a well-defined model (i.e., with information matrix $I - \bar{R} \succ 0$). We also note that condition (iv) relates walk-summability to the

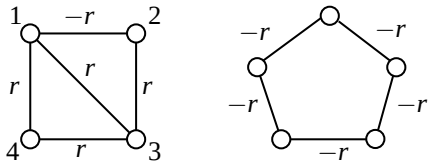


Figure 3: Example graphs: (a) 4-cycle with a chord. (b) 5-cycle.

so-called H-matrices in linear algebra (Horn and Johnson, 1991; Varga, 2000).[12] As an immediate corollary, we identify the following important subclass of walk-summable models:

Corollary 2 (Attractive Models) *Let $J = I - R$ be a valid model ($J \succ 0$) with non-negative partial correlations $R \geq 0$. Then, $J = I - R$ is walk-summable.*

A superclass of attractive models is the set of *non-frustrated models*. A model is non-frustrated if it does not contain any frustrated cycles, that is, cycles with an odd number of negative edge weights. We show in Appendix A (in the proof of Corollary 3) that if the model is non-frustrated, then one can negate some of the variables to make the model attractive[13]. Hence, we have another subclass of walk-summable models (the inclusion is strict as some frustrated models are walk-summable, see Example 1):

Corollary 3 (Non-frustrated models) *Let $J = I - R$ be valid. If R is non-frustrated then J is walk-summable.*

Example 1. In Figure 3 we illustrate two small Gaussian graphical models, which we use throughout the paper. In both models the information matrix J is normalized to have unit diagonal and to have partial correlations as indicated in the figure. Consider the 4-cycle with a chord in Figure 3(a). The model is frustrated (due to the opposing sign of one of the partial correlations), and increasing r worsens the frustration. For $0 \leq r \leq 0.39039$, the model is valid and walk-summable: for example, for $r = 0.39$, $\lambda_{\min}(J) = 0.22 > 0$, and $\rho(\bar{R}) \approx 0.9990 < 1$. In the interval $0.39039 \leq r \leq 0.5$ the model is valid, but not walk-summable: for example, for $r = 0.4$, $\lambda_{\min} = 0.2 > 0$, and $\rho(\bar{R}) \approx 1.0246 > 1$. Also, note that for R (as opposed to $\bar{R}$), $\rho(R) \leq 1$ for $r \leq 0.5$ and $\rho(R) > 1$ for $r > 0.5$. Finally, the model stops being diagonally dominant above $r = \frac{1}{3}$, but walk-summability is a strictly larger set and extends until $r \approx 0.39039$. We summarize various critical points for this model and for the model in Figure 3(b) in the diagram in Figure 4.

Here are additional useful implications of walk-summability, with proof in Appendix A:

Proposition 4 (WS Necessary Conditions) *All of the following are implied by walk-summability:*

12. A (possibly non-symmetric) matrix A is an *H-matrix* if all eigenvalues of the matrix $M(A)$, where $M_{ii} = |A_{ii}|$, and $M_{ij} = -|A_{ij}|$ for $i \neq j$, have positive real parts. For symmetric matrices this is equivalent to M being positive definite. In (iv) J is an H-matrix since $M(J) = I - \bar{R} \succ 0$.
13. This result is referred to in Kirkland et al. (1996). However, in addition to proving that there exists such a sign similarity, our proof also gives an algorithm which checks whether or not the model is frustrated, and determines which subset of variables to negate if the model is non-frustrated.

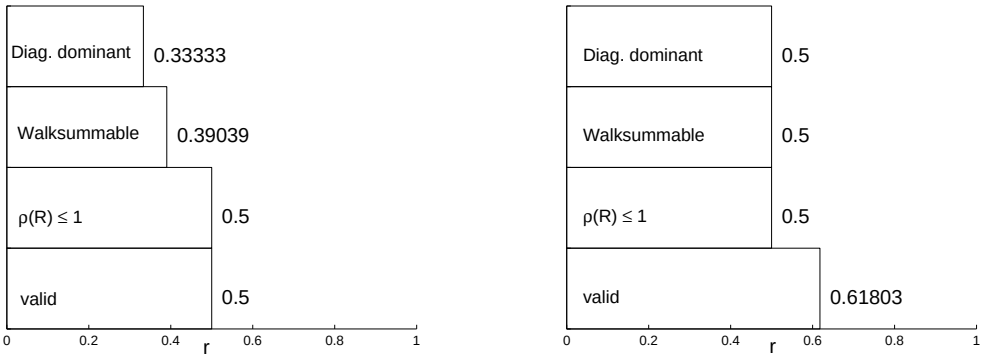


Figure 4: Critical regions for example models from Figure 3. (a) 4-cycle with a chord. (b) 5-cycle.

(i) $\rho(R) < 1$.

(ii) $J = I - R \succ 0$.

(iii) $\sum_k R^k = (I-R)^{-1}$.

Implication (ii) shows that walk-summability is a sufficient condition for validity of the model. Also, (iii) shows the relevance of walk-sums for inference since $P = J^{-1} = (I-R)^{-1} = \sum_k R^k$ and $\mu = J^{-1}h = \sum_k R^k h$.

3.2 Walk-Sums for Inference

Next we show that, in walk-summable models, means and variances correspond to walk-sums over certain sets of walks.

Proposition 5 (WS Inference) *If $J = I - R$ is walk-summable, then the covariance $P = J^{-1}$ is given by the walk-sums:*

$$P_{ij} = \sum_{w: i \to j} \phi(w).$$

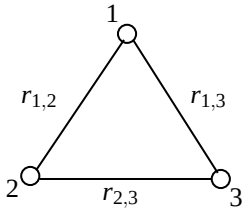
Also, the means are walk-sums reweighted by the value of h at the start of each walk:

$$\mu_i = \sum_{w: * \to i} h_* \phi(w)$$

*where the sum is over **all** walks which end at node i (with arbitrary starting node), and where $*$ denotes the starting node of the walk w.*

Proof. We use the fact that $(R^l)_{ij} = \sum_{w: i \overset{l}{\to} j} \phi(w)$. Then,

$$P_{ij} = \sum_l (R^l)_{ij} = \sum_l \sum_{w: i \overset{l}{\to} j} \phi(w) = \sum_{w: i \to j} \phi(w)$$



Single walk: $w = (1,2,3)$. Weight: $\phi(w) = r_{1,2}r_{2,3}$,
$\phi_h(w) = h_1 r_{1,2}r_{2,3}$.

Self-return walks, $\mathcal{W}(1 \to 1)$: $\{(1),(1,2,1),(1,3,1),$
$(1,2,3,1),(1,3,2,1),(1,2,1,2,1),...\}$
$P_{1,1} = \phi(1 \to 1) = 1 + r_{1,2}r_{2,1} + r_{1,3}r_{3,1} + r_{1,2}r_{2,3}r_{3,1} + ...$

Set of walks $\mathcal{W}(* \to 1)$: $\{(1),(2,1),(3,1),(2,3,1),$
$(3,2,1),(1,2,1)(1,3,1),...\}$
$\mu_1 = \phi_h(* \to 1) = h_1 + h_2 r_{2,1} + h_3 r_{3,1} + h_2 r_{2,3}r_{3,1} + ...$

Figure 5: Illustration of walk-sums for means and variances.

and

$$\mu_i = \sum_j h_j P_{ji} = \sum_j \sum_{w: j \xrightarrow{l} i} h_j \phi(w) = \sum_{w: * \to i} h_* \phi(w) \ \square$$

Walk-Sum Notation We now provide a more compact notation for walk-sets and walk-sums. In general, given a set of walks $\mathcal{W}$ we define the walk-sum:

$$\phi(\mathcal{W}) = \sum_{w \in \mathcal{W}} \phi(w)$$

and the reweighted walk-sum:

$$\phi_h(\mathcal{W}) = \sum_{w \in \mathcal{W}} h_{w_0} \phi(w)$$

where w_0 denotes the initial node in the walk w. Also, we adopt the convention that $\mathcal{W}(\dots)$ denotes the set of all walks having some property $\dots$ and denote the associated walk-sums simply as $\phi(\dots)$ or $\phi_h(\dots)$. For instance, $\mathcal{W}(i \to j)$ denotes the set of all walks from i to j and $\phi(i \to j)$ is the corresponding walk-sum. Also, $\mathcal{W}(* \to i)$ denotes the set all walks that end at node i and $\phi_h(* \to i)$ is the corresponding reweighted walk-sum. In this notation, $P_{ij} = \phi(i \to j)$ and $\mu_i = \phi_h(* \to i)$. An illustration of walk-sums and their connection to inference appears in Figure 5 where we list some walks and walk-sums for a 3-cycle graph.

Walk-Sum Algebra We now show that the walk-sums required for inference in walk-summable models can be significantly simplified by exploiting the recursive structure of walks. To do so, we make use of some simple algebraic properties of walk-sums. The following lemmas all assume that the model is walk-summable.

Lemma 6 *Let $\mathcal{W} = \cup_{k=1}^{\infty} \mathcal{W}_k$ where the subsets $\mathcal{W}_k$ are disjoint. Then, $\phi(\mathcal{W}) = \sum_{k=1}^{\infty} \phi(\mathcal{W}_k)$.*

Proof. By the sum-partition theorem for absolutely convergent series (Godement, 2004): $\sum_{w \in \mathcal{W}} \phi(w) = \sum_k \sum_{w \in \mathcal{W}_k} \phi(w)$. $\square$

Lemma 7 *Let $\mathcal{W} = \cup_{k=1}^{\infty} \mathcal{W}_k$ where $\mathcal{W}_k \subset \mathcal{W}_{k+1}$ for all k. Then, $\phi(\mathcal{W}) = \lim_{k \to \infty} \phi(\mathcal{W}_k)$.*

Proof. Let $\mathcal{W}_0$ be the empty set. Then, $\mathcal{W} = \cup_{k=1}^{\infty}(\mathcal{W}_k \setminus \mathcal{W}_{k-1})$. By Lemma 6,

$$\phi(\mathcal{W}) = \sum_{k=1}^{\infty} \phi(\mathcal{W}_k \setminus \mathcal{W}_{k-1}) = \lim_{N\to\infty} \sum_{k=1}^{N} (\phi(\mathcal{W}_k) - \phi(\mathcal{W}_{k-1})) = \lim_{N\to\infty} (\phi(\mathcal{W}_N) - \phi(\mathcal{W}_0))$$

where we use $\phi(\mathcal{W}_0) = 0$ in the last step to obtain the result. □

Given two walks $u = (u_0, \ldots, u_n)$ and $v = (v_0, \ldots, v_m)$ with $u_n = v_0$ (walk v begins where walk u ends) we define the product of walks $uv = (u_0, \ldots, u_n, v_1, \ldots, v_m)$. Let $\mathcal{U}$ and $\mathcal{V}$ be two countable sets of walks such that every walk in $\mathcal{U}$ ends at a given node i and every walk in $\mathcal{V}$ begin at this node. Then we define the product set $\mathcal{UV} = \{uv \mid u \in \mathcal{U}, v \in \mathcal{V}\}$. We say that $(\mathcal{U}, \mathcal{V})$ is a *valid decomposition* if for every $w \in \mathcal{UV}$ there is a unique pair $(u, v) \in \mathcal{U} \times \mathcal{V}$ such that $uv = w$.

Lemma 8 *Let $(\mathcal{U}, \mathcal{V})$ be a valid decomposition. Then, $\phi(\mathcal{UV}) = \phi(\mathcal{U})\phi(\mathcal{V})$.*

Proof. For individual walks it is evident that $\phi(uv) = \phi(u)\phi(v)$. Note that $\mathcal{UV} = \cup_{u\in\mathcal{U}} u\mathcal{V}$, where the sets $u\mathcal{V} \triangleq \{uv | v \in \mathcal{V}\}$ are mutually disjoint. By Lemma 6,

$$\phi(\mathcal{UV}) = \sum_{u\in\mathcal{U}} \phi(u\mathcal{V}) = \sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} \phi(uv) = \sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} \phi(u)\phi(v) = \left(\sum_{u\in\mathcal{U}} \phi(u)\right)\left(\sum_{v\in\mathcal{V}} \phi(v)\right)$$

where we have used $\phi(u\mathcal{V}) = \sum_{v\in\mathcal{V}} \phi(uv)$ because $u\mathcal{V}$ is one-to-one with $\mathcal{V}$. □

Note that $\mathcal{W}(i \to i)$ is the set of *self-return walks* at node i, that is, walks which begin and end at node i. These self-return walks include walks which return to i many times. Let $\mathcal{W}(i \xrightarrow{\setminus i} i)$ be the set of all walks with non-zero length that begin and end at i but do not visit i at any other point in between. We call these the *single-revisit* self-return walks at node i. The set of self-return walks that return exactly k times is generated by taking the product of k copies of $\mathcal{W}(i \xrightarrow{\setminus i} i)$ denoted by $\mathcal{W}^k(i \xrightarrow{\setminus i} i)$. Thus, we obtain all self-return walks as

$$\mathcal{W}(i \to i) = \cup_{k\geq 0} \mathcal{W}^k(i \xrightarrow{\setminus i} i) \tag{12}$$

where $\mathcal{W}^0(i \xrightarrow{\setminus i} i) \triangleq \{(i)\}$.

Similarly, recall that $\mathcal{W}(* \to i)$ denotes the set of all walks which end at node i. Let $\mathcal{W}(* \xrightarrow{\setminus i} i)$ denote the set of walks with non-zero length which end at node i and do not visit i previously (we call them *single-visit walks*). Thus, all walks which end at i are obtained as:

$$\mathcal{W}(* \to i) = \left(\{(i)\} \cup \mathcal{W}(* \xrightarrow{\setminus i} i)\right) \mathcal{W}(i \to i), \tag{13}$$

which is a valid decomposition.

Now we can decompose means and variances in terms of single-visit and single-revisit walk-sums, which we will use in section 4.1 to analyze BP.

Proposition 9 *Let $\alpha_i = \phi(i \xrightarrow{\setminus i} i)$ and $\beta_i = \phi_h(* \xrightarrow{\setminus i} i)$. Then,*

$$P_{ii} = \frac{1}{1-\alpha_i} \quad \textit{and} \quad \mu_i = \frac{h_i + \beta_i}{1-\alpha_i}.$$

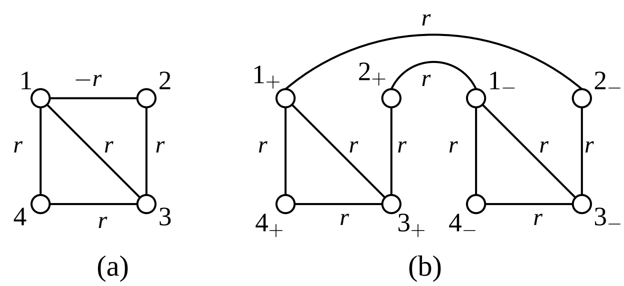


Figure 6: (a) A frustrated model defined on G with one negative edge ($r > 0$). (b) The corresponding attractive model defined on $\hat{G}$.

Proof. First note that the decomposition of $\mathcal{W}^k(i \xrightarrow{\backslash i} i)$ into products of k single-revisit self-return walks is a valid decomposition. By Lemma 8, $\phi(\mathcal{W}^k(i \xrightarrow{\backslash i} i)) = \phi^k(i \xrightarrow{\backslash i} i) = \alpha_i^k$. Then, by (12) and Lemma 6:

$$P_{ii} = \phi(i \to i) = \sum_k \alpha_i^k = \frac{1}{1-\alpha_i}.$$

Walk-summability of the model implies convergence of the geometric series (i.e., $|\alpha_i| < 1$). Lastly, the decomposition in (13) implies

$$\mu_i = \phi_h(* \to i) = (h_i + \phi_h(* \xrightarrow{\backslash i} i))\phi(i \to i) = \frac{h_i + \beta_i}{1-\alpha_i} \quad \square$$

3.3 Correspondence to Attractive Models

We have already shown that attractive models are walk-summable. Interestingly, it turns out that inference in any walk-summable model can be reduced to inference in a corresponding attractive model defined on a graph with twice as many nodes. The basic idea here is to separate out the walks with positive and negative weights.

Specifically, let $\hat{G} = (\hat{V}, \hat{E})$ be defined as follows. For each node $i \in V$ we define two corresponding nodes $i_+ \in V_+$ and $i_- \in V_-$, and set $\hat{V} = V_+ \cup V_-$. For each edge $\{i, j\} \in E$ with $r_{ij} > 0$ we define two edges $\{i_+, j_+\}, \{i_-, j_-\} \in \hat{E}$, and set the partial correlations on these edges to be equal to r_{ij}. For each edge $\{i, j\} \in E$ with $r_{ij} < 0$ we define two edges $\{i_+, j_-\}, \{i_-, j_+\} \in \hat{G}$, and set the partial correlations to be $-r_{ij}$. See Figure 6 for an illustration.

Let $(R_+)_{ij} = \max\{R_{ij}, 0\}$ and $(R_-)_{ij} = \max\{-R_{ij}, 0\}$. Then R can be expressed as the difference of these non-negative matrices: $R = R_+ - R_-$. Based on our construction, we have that $\hat{R} = \begin{pmatrix} R_+ & R_- \\ R_- & R_+ \end{pmatrix}$ and $\hat{J} = I - \hat{R}$. This defines a unit-diagonal information matrix $\hat{J}$ on $\hat{G}$. Note that if $\hat{J} \succ 0$ then this defines a valid attractive model.

Proposition 10 *$\hat{J} = I - \hat{R} \succ 0$ if and only if $J = I - R$ is walk-summable.*

The proof relies on the Perron-Frobenius theorem and is given in Appendix A. Now, let $h = h_+ - h_-$ with $(h_+)_i = \max\{h_i, 0\}$ and $(h_-)_i = \max\{-h_i, 0\}$. Define $\hat{h} = \begin{pmatrix} h_+ \\ h_- \end{pmatrix}$. Now we have the information form model $(\hat{h}, \hat{J})$ which is a valid, attractive model and also has non-negative node

potentials. Performing inference with respect to this augmented model, we obtain the mean vector $\hat{\mu} = \begin{pmatrix} \hat{\mu}_+ \\ \hat{\mu}_- \end{pmatrix} \triangleq \hat{J}^{-1}\hat{h}$ and covariance matrix $\hat{P} = \begin{pmatrix} \hat{P}_{++} & \hat{P}_{+-} \\ \hat{P}_{-+} & \hat{P}_{--} \end{pmatrix} \triangleq \hat{J}^{-1}$. From these calculations, we can obtain the moments (μ, P) of the original walk-summable model (h, J):

Proposition 11 *$P = \hat{P}_{++} - \hat{P}_{+-}$ and $\mu = \hat{\mu}_+ - \hat{\mu}_-$.*

The proof appears in Appendix A. This proposition shows that estimation of walk-summable models may be reduced to inference in an attractive model in which all walk-sums are sums of positive weights. In essence, this is accomplished by summing walks with positive and negative weights separately and then taking the difference, which is only possible for walk-summable models.

3.4 Pairwise-Normalizability

To simplify presentation we assume that the graph does not contain any isolated nodes (a node without any incident edges). Then, we say that the information matrix J is *pairwise-normalizable* (PN) if we can represent J in the form

$$J = \sum_{e \in E} [J_e]$$

where each J_e is a 2×2 symmetric, positive definite matrix.[14] The notation $[J_e]$ means that J_e is zero-padded to a $|V| \times |V|$ matrix with its principal submatrix for $\{i, j\}$ being J_e (with $e = \{i, j\}$). Thus, $x^T[J_e]x = x_e^T J_e x_e$. Pairwise-normalizability implies that $J \succ 0$ because each node is covered by at least one positive definite submatrix J_e. Let $\mathcal{J}_{PN}$ denote the set of $n \times n$ pairwise-normalizable information matrices J (not requiring unit-diagonal normalization). This set has nice convexity properties. Recall that a set $\mathcal{X}$ is *convex* if $x, y \in \mathcal{X}$ implies $\lambda x + (1-\lambda)y \in \mathcal{X}$ for all $0 \leq \lambda \leq 1$ and is a *cone* if $x \in \mathcal{X}$ implies $\alpha x \in \mathcal{X}$ for all $\alpha > 0$. A cone $\mathcal{X}$ is *pointed* if $\mathcal{X} \cap -\mathcal{X} = \{0\}$.

Proposition 12 (Convexity of PN models) *The set $\mathcal{J}_{PN}$ is a convex pointed cone.*

The proof is in Appendix A. We now establish the following fundamental result:

Proposition 13 (WS $\Leftrightarrow$ PN) *$J = I - R$ is walk-summable if and only if it is pairwise-normalizable.*

Our proof appears in in Appendix A. An equivalent result has been derived independently in the linear algebra literature: Boman et al. (2005) establish that symmetric H-matrices with positive diagonals (which is equivalent to WS by part (iv) of Proposition 1) are equivalent to matrices with factor width at most two (PN models). However, the result $PN \Rightarrow WS$ was established earlier by Johnson (2001). Our proof for $WS \Rightarrow PN$ uses the Perron-Frobenius theorem, whereas Boman et al. (2005) use the generalized diagonal dominance property of H-matrices.

Equivalence to pairwise-normalizability gives much insight into the set of walk-summable models. For example, the set of unit-diagonal J matrices that are walk-summable is convex, as it is the intersection of $\mathcal{J}_{PN}$ with an affine space. Also, the set of walk-summable J matrices that are sparse with respect to a particular graph G (with some entries of J are restricted to 0) is convex.

Another important class of models are those that have a *diagonally dominant* information matrix, that is, where for each i it holds that $\sum_{j \neq i} |J_{ij}| < J_{ii}$.

14. An alternative definition of pairwise-normalizability is the existence of a decomposition $J = cI + \sum_{e \in E} [J_e]$, where $c > 0$, and $J_e \succeq 0$. For graphs without isolated nodes, both definitions are equivalent.

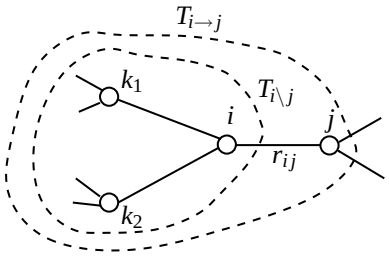


Figure 7: Illustration of the subtree notation, $T_{i\to j}$ and $T_{i\setminus j}$.

Proposition 14 *Diagonally dominant models are pairwise-normalizable (walk-summable).*

A constructive proof is given in Appendix A. The converse does not hold: not all pairwise-normalizable models are diagonally dominant. For instance, in our example of a 4-cycle with a chord shown in Figure 3(a), with $r = .38$ the model is not diagonally dominant but is walk-summable and hence pairwise-normalizable.

4. Walk-sum Interpretation of Belief Propagation

In this section we use the concepts and machinery of walk-sums to analyze belief propagation. We begin with models on trees, for which, as we show, all valid models are walk-summable. Moreover, for these models we show that exact walk-sums over infinite sets of walks for means and variances can be computed efficiently in a recursive fashion. We show that these walk-sum computations map exactly to belief propagation updates. These results (and the computation tree interpretation of LBP recursions) then provide the foundation for our analysis of loopy belief propagation in Section 4.2.

4.1 Walk-Sums and BP on Trees

Our analysis of BP makes use of the following property:

Proposition 15 (Trees are walk-summable) *For tree structured models $J \succ 0 \Leftrightarrow \rho(\bar{R}) \leq 1$ (i.e., all valid trees are walk-summable). Also, for trees $\rho(\bar{R}) = \rho(R) = \lambda_{\max}(R)$.*

Proof. The proof is a special case of the proof of Corollary 3. Trees are non-frustrated (as there are no cycles, let alone frustrated cycles) so they are walk-summable. Negating some variables makes the model attractive and does not change the eigenvalues. □

The proposition shows that walk-sums for means and variances are always defined on tree-structured models, and can be reordered in arbitrary ways without affecting convergence. We rely on this fact heavily in subsequent sections. The next two results identify walk-sum variance and mean computations with the BP update equations. The ingredients for these results are decompositions of the variance and mean walk-sums in terms of sums over walks on subtrees, together with the decomposition in terms of single-revisit and single-visit walks provided in Proposition 9.

Walk-Sum Variance Calculation Let us look first at the computation of the variance at node j, which is equal to the self-return walk-sum $\phi(j \to j)$. This can be computed directly from the single-revisit walk-sum $\alpha_j = \phi(j \stackrel{\setminus j}{\to} j)$ as in Proposition 9. This latter walk-sum can be further decomposed into sums over disjoint subsets of walks each of which corresponds to single-revisit self-return walks that exit node j via a specific one of its neighbors, say i. In particular, as illustrated in Figure 7, the single-revisit self-return walks that do this correspond to walks that live in the subtree $T_{i\to j}$. Using the notation $\mathcal{W}(j \stackrel{\setminus j}{\to} j \mid T_{i\to j})$ for the set of all single-revisit walks which are restricted to stay in subtree $T_{i\to j}$ we see that

$$\alpha_j = \phi(j \stackrel{\setminus j}{\to} j) = \sum_{i\in\mathcal{N}(j)} \phi(j \stackrel{\setminus j}{\to} j \mid T_{i\to j}) \triangleq \sum_{i\in\mathcal{N}(j)} \alpha_{i\to j}\,.$$

Moreover, every single-revisit self-return walk that lives in $T_{i\to j}$ must leave *and* return to node j through the single edge (i,j), and between these first and last steps must execute a (possibly multiple-revisit) self-return walk at node i that is constrained *not* to pass through node j, that is, to live in the subtree $T_{i\setminus j}$ indicated in Figure 7. Thus

$$\alpha_{i\to j} = \phi(j \stackrel{\setminus j}{\to} j \mid T_{i\to j}) = r_{ij}^2 \phi(i \to i \mid T_{i\setminus j}) \triangleq r_{ij}^2 \gamma_{i\setminus j}\,. \tag{14}$$

We next show that the walk-sums α_j and $\alpha_{i\to j}$ (hence variances P_j) can be efficiently calculated by a walk-sum analog of belief propagation. We have the following result:

Proposition 16 *Consider a valid tree model $J = I - R$. Then $\alpha_{i\to j} = -\Delta J_{i\to j}$ and $\gamma_{i\setminus j} = \hat{J}_{i\setminus j}^{-1}$, where $\Delta J_{i\to j}$ and $\hat{J}_{i\setminus j}^{-1}$ are the quantities defined in the Gaussian BP equations (7) and (8).*

See Appendix A for the proof.

Walk-Sum Mean Calculation We extend the above analysis to calculate means in trees. Mean μ_j is the reweighted walk-sum over walks that start anywhere and end at node j, $\mu_j = \phi_h(* \to j)$. Any walk that ends at node j can be expressed as a single-visit walk to node j followed by a multiple-revisit self-return walk from node j: $\phi_h(* \to j) = \left(h_j + \phi_h(* \stackrel{\setminus j}{\to} j)\right)\phi(j \to j)$, where the term h_j corresponds to the length-zero walk that starts and ends at node j.

As we have done for the variances, the single-visit walks to node j can be partitioned into the single-visit walks that reach node j from each of its neighbors, say node i and thus prior to this last step across the edge (i,j), reside in the subtree $T_{i\setminus j}$, so that

$$\beta_{i\to j} \triangleq \phi_h(* \stackrel{\setminus j}{\to} j \mid T_{i\to j}) = r_{ij}\phi_h(* \to i \mid T_{i\setminus j}).$$

Proposition 17 *Consider a valid tree model $J = I - R$. Then $\beta_{i\to j} = \Delta h_{i\to j}$, where $\Delta h_{i\to j}$ is the quantity defined in the Gaussian BP equation (8).*

The proof appears in Appendix A.

4.2 LBP in Walk-Summable Models

In this subsection we use the LBP computation tree to show that LBP includes all the walks for the means, but only a subset of the walks for the variances. This allows us to prove LBP convergence for all walk-summable models. In contrast, for non-walksummable models LBP may or may not converge (and in fact the variances may converge but the means may not). As we will see in Section 5, this can be analyzed by examining walk-summability (and hence validity) of the computation tree, rather than walk-summability of the original model.

As we have discussed, running LBP for some number of iterations yields identical calculations at any particular node i to the exact inference calculations on the corresponding computation tree rooted at node i. We use the notation $T_i^{(n)}$ for the nth computation tree at node i, T_i for the full computation tree (as $n \to \infty$) and we assign the label 0 to the root node. Then, $P_0(T_i^{(n)})$ denotes the variance at the root node of the nth computation tree rooted at node i in G. The LBP variance estimate at node i after n steps is equal to

$$\hat{P}_i^{(n)} = P_0(T_i^{(n)}) = \phi(0 \to 0 \mid T_i^{(n)}).$$

Similarly, the LBP estimate of the mean μ_i after n steps of LBP is

$$\hat{\mu}_i^{(n)} = \mu_0(T_i^{(n)}) = \phi_h(* \to 0 \mid T_i^{(n)}).$$

As we have mentioned, the definition of the computation trees $T_i^{(n)}$ depend upon the message schedule $\{\mathcal{M}^{(n)}\}$ of LBP, which specifies which subset of messages are updated at iteration n. We say that a message schedule is *proper* if every message is updated infinitely often, that is, if for every $m > 0$ and every directed edge (i, j) in the graph there exists $n > m$ such that $(i, j) \in \mathcal{M}^{(n)}$. Clearly, the fully parallel form is proper since every message is updated at every iteration. Serial forms which iteratively cycle through the directed edges of the graph are also proper. All of our convergence analysis in this section presumes a proper message schedule. We remark that as walk-summability ensures convergence of walk-sums independent of the order of summation, it makes the choice of a particular message schedule unimportant in our convergence analysis. The following result is proven in Appendix A.

Lemma 18 (Walks in G and in T_i) *There is a one-to one correspondence between finite-length walks in G that end at i, and walks in T_i that end at the root node. In particular, for each such walk in G there is a corresponding walk in $T_i^{(n)}$ for n large enough.*

Now, recall that to compute the mean μ_i we need to gather walk-sums over all walks that start anywhere and end at i. We have just shown that LBP gathers all of these walks as the computation tree grows to infinity. The story for the variances is different. The true variance P_{ii} is a walk-sum over all self-return walks that start and end at i in G. However, walks in G that start and end at i may map to walks that start at the root node of $T_i^{(n)}$, but end at a *replica* of the root node instead of the root. These walks are not captured by the LBP variance estimate.[15] The walks for the variance estimate $P_0(T_i^{(n)})$ are self-return walks $\mathcal{W}(0 \to 0 \mid T_i^{(n)})$ that start and end at the root node in the

15. Recall that the computation tree is a representation of the computations seen at the root node of the tree, and it is *only* the computation at *this node*—that is, at *this replica* of node i that corresponds to the LBP computation at node i in G.

computation tree. Consider Figure 2. The walk $(1,2,3,1)$ is a self-return walk in the original graph G but is *not* a self-return walk in the computation tree shown in Figure 2(d). LBP variances capture only those self-return walks of the original graph G that are also self-return walks in the computation tree—for example, the walk $(1,3,2,3,4,3,1)$ is a self-return walk in both Figures 2(a) and (d). We call such walks *backtracking*. Hence,

Lemma 19 (Self-return walks in G and in T_i) *The LBP variance estimate at each node is a sum over the backtracking self-return walks in G, a subset of all self-return walks needed to calculate the correct variance.*

Note that back-tracking walks for the variances have positive weights, since each edge in the walk is traversed an even number of times. With each LBP step the computation tree grows and new back-tracking walks are included, hence variance estimates grow monotonically.[16]

We have shown which walks LBP gathers based on the computation tree. The convergence of the corresponding walk-sums remains to be analyzed. In walk-summable models the answer is simple:

Lemma 20 (Computation trees of WS models are WS) *For a walk-summable model all its computation trees $T_i^{(n)}$ (for all n and i) are walk-summable and hence valid.*

Intuitively, walks in the computation tree $T_i^{(n)}$ are subsets of the walks in G, and hence they converge. This implies that the computation trees are walk-summable, and hence valid. This argument can be made precise, but a shorter formal proof using monotonicity of the spectral radius (11) appears in Appendix A. Next, we use these observations to show convergence of LBP for walk-summable models.

Proposition 21 (Convergence of LBP for walk-summable models) *If a model on a graph G is walk-summable, then LBP is well-posed, the means converge to the true means and the LBP variances converge to walk-sums over the backtracking self-return walks at each node.*

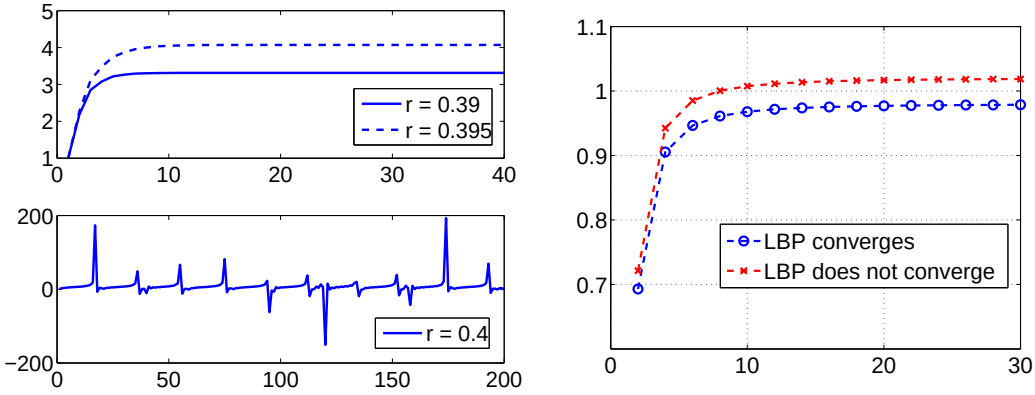
Proof. Let $\mathcal{W}(i \overset{BT}{\to} i)$ denote the back-tracking self-return walks at node i. By Lemmas 18 and 19, we have:

$$\begin{aligned} \mathcal{W}(* \to i) &= \cup_n \mathcal{W}(* \to 0 | T_i^{(n)}) \\ \mathcal{W}(i \overset{BT}{\to} i) &= \cup_n \mathcal{W}(0 \to 0 | T_i^{(n)}). \end{aligned}$$

We note that the computation trees $T_i^{(n)}$ at node i are nested, $T_i^{(n)} \subset T_i^{(n+1)}$ for all n. Hence, $\mathcal{W}(* \to 0|T_i^{(n)}) \subset \mathcal{W}(* \to 0|T_i^{(n+1)})$ and $\mathcal{W}(0 \to 0|T_i^{(n)}) \subset \mathcal{W}(0 \to 0|T_i^{(n+1)})$. Then, by Lemma 7, we obtain the result:

$$\begin{aligned} \mu_i = \phi_h(* \to i) &= \lim_{n\to\infty} \phi_h(* \to 0|T_i^{(n)}) = \lim_{n\to\infty} \hat{\mu}_i^{(n)} \\ P_i^{(BT)} \triangleq \phi(i \overset{BT}{\to} i) &= \lim_{n\to\infty} \phi(0 \to 0|T_i^{(n)}) = \lim_{n\to\infty} \hat{P}_i^{(n)}. \qquad \square \end{aligned}$$

16. Monotonically increasing variance estimates is a characteristic of the particular initialization of LBP that we use, that is, the potential decomposition (4) together with uninformative initial messages. If one instead uses a pairwise-normalized potential decomposition, the variances are then monotonically *decreasing*.



(a) (b)

Figure 8: (a) LBP variances vs. iteration. (b) $\rho(R_n)$ vs. iteration.

Corollary 22 *LBP converges for attractive, non-frustrated, and diagonally dominant models. In attractive and non-frustrated models LBP variance estimates are less than or equal to the true variances (the missing non-backtracking walks all have positive weights).*

In Weiss and Freeman (2001) Gaussian LBP is analyzed for pairwise-normalizable models. They show convergence for the case of diagonally dominant models, and correctness of the means in case of convergence. The class of walk-summable models is strictly larger than the class of diagonally dominant models, so our sufficient condition is stronger. They also show that LBP variances omit some terms needed for the correct variances. These terms correspond to correlations between the root and its replicas in the computation tree. In our framework, each such correlation is a walk-sum over the subset of non-backtracking self-return walks in G that, in the computation tree, begin at a particular replica of the root.

Example 2. Consider the model in Figure 3(a). We summarize various critical points for this model in Figure 9. For $0 \leq r \leq .39039$ the model is walk-summable and LBP converges; then for a small interval $.39039 \leq r \leq .39865$ the model is not walk-summable but LBP still converges, and for larger r LBP does not converge. We apply LBP to this model with $r = 0.39, 0.395$ and 0.4, and plot the LBP variance estimates for node 1 vs. the iteration number in Figure 8(a). LBP converges in the walk-summable case for $r = .39$, with $\rho(\bar{R}) \approx .9990$. It also converges for $r = 0.395$ with $\rho(\bar{R}) \approx 1.0118$, but soon fails to converge as we increase r to 0.4 with $\rho(\bar{R}) \approx 1.0246$.

Also, for $r = .4$, we note that $\rho(R) = .8 < 1$ and the series $\sum_l R^l$ converges (but $\sum_l \bar{R}^l$ does not) and LBP does not converge. Hence, $\rho(R) < 1$ is not sufficient for LBP convergence showing the importance of the stricter walk-summability condition $\rho(\bar{R}) < 1$.

5. LBP in Non-Walksummable Models

While the condition in Proposition 21 is necessary and sufficient for certain special classes of models—for example, for trees and single cycles—it is only sufficient more generally, and, as

in Example 2, LBP may converge for some non-walksummable models. We extend our analysis to develop a tighter condition for convergence of LBP variances based on a weaker form of walk-summability defined with respect to the computation trees (instead of G). We have shown in Proposition 15 that for trees walk-summability and validity are equivalent, and $\rho(\bar{R}) < 1 \Leftrightarrow \rho(R) < 1 \Leftrightarrow J \succ 0$. Hence, our condition essentially corresponds to validity of the computation tree.

First, we note that when a model on G is valid (J is positive definite) but not walk-summable, then some finite computation trees may be invalid (indefinite). This turns out to be the primary reason why belief propagation can fail to converge. Walk-summability on the original graph implies walk-summability (and hence validity) on all of its computation trees. But if the model is not walk-summable, then its computation tree may or may not be valid.

We characterize walk-summability of the computation trees as follows. Let $T_i^{(n)}$ be the nth computation tree rooted at some node i. We define $R_i^{(n)} \triangleq I - J_i^{(n)}$ where $J_i^{(n)}$ is the normalized information matrix for $T_i^{(n)}$ and I is an identity matrix. The nth computation tree $T_i^{(n)}$ is walk-summable (valid) if and only if $\rho(R_i^{(n)}) < 1$ due to the fact that $\rho(\bar{R}_i^{(n)}) = \rho(R_i^{(n)})$ for trees. We are interested in the validity of all finite computation trees, so we consider the quantity $\lim_{n\to\infty} \rho(R_i^{(n)})$. Lemma 23 guarantees the existence of this limit:

Lemma 23 *The sequence $\{\rho(R_i^{(n)})\}$ is monotonically increasing and bounded above by $\rho(\bar{R})$. Thus, $\lim_{n\to\infty} \rho(R_i^{(n)})$ exists, and is equal to $\sup_n \rho(R_i^{(n)})$.*

In the proof we use *k-fold graphs*, which we introduce in Appendix B. The proof appears in Appendix A. The limit in Lemma 23 is defined with respect to a particular root node and message schedule. The next lemma shows that for connected graphs, as long as the message schedule is proper, they do not matter.

Lemma 24 *For connected graphs and with proper message schedule, $\rho_\infty \triangleq \lim_{n\to\infty} \rho(R_i^{(n)})$ is independent of i. The limit does not change by using any other proper message schedule.*

This independence results from the fact that for large n the computation trees rooted at different nodes overlap significantly. Technical details of the proof appear in Appendix A. Using this lemma we suppress the dependence on the root node i from the notation to simplify matters. The limit ρ_∞ turns out to be critical for convergence of LBP variances:

Proposition 25 (LBP validity/variance convergence) *(i) If $\rho_\infty < 1$, then all finite computation trees are valid and the LBP variances converge to walk-sums over the back-tracking self-return walks. (ii) If $\rho_\infty > 1$, then the computation tree eventually becomes invalid and LBP is ill-posed.*

Proof. (i) Since $\rho_\infty = \lim_{n\to\infty} \rho(R^{(n)}) < 1$ and the sequence $\{\rho(R^{(n)})\}$ is monotonically increasing, then there exists $\delta > 0$ such that $\rho(R^{(n)}) \leq 1 - \delta$ for all n. This implies that all the computation trees $T^{(n)}$ are walk-summable and that variances monotonically increase (since weights of back-tracking walks are positive, see the discussion after Lemma 19). We have that $\lambda_{\max}(R^{(n)}) \leq 1 - \delta$, so $\lambda_{\min}(J^{(n)}) \geq \delta$ and $\lambda_{\max}(P^{(n)}) \leq \frac{1}{\delta}$. The maximum eigenvalue of a matrix is a bound on the maximum entry of the matrix, so $(P^{(n)})_{ii} \leq \lambda_{\max}(P^{(n)}) \leq \frac{1}{\delta}$. The variances are monotonically increasing and bounded above, hence they converge.

(ii) If $\lim_{n\to\infty} \rho(R^{(n)}) > 1$, then there exists an m such that $\rho(R^{(n)}) > 1$ for all $n \geq m$. This means that these computation trees $T^{(n)}$ are invalid, and that the variance estimates at some of the nodes

are negative. □

As discussed in Section 2.2, the LBP computation tree is valid if and only if the information parameters $\hat{J}_{i\backslash j}^{(n)}$ and $\hat{J}_i^{(n)}$ in (7), (9) computed during LBP iterations are strictly positive for all n. Hence, it is easily detected if the LBP computation tree becomes invalid. In this case, continuing to run LBP is not meaningful and will lead to division by zero (if the computation tree is singular) or to negative variances (if it is not positive definite).

Recall that the limit ρ_∞ is invariant to message order by Lemma 24. Hence, by Proposition 25, convergence of LBP variances is likewise invariant to message order (except possibly when $\rho_\infty = 1$). The limit ρ_∞ is bounded above by $\rho(\bar{R})$, hence walk-summability in G is a sufficient condition for well-posedness of the computation tree: $\rho_\infty \leq \rho(\bar{R}) < 1$. However, the bound is not tight in general (except for trees and single cycles). This is related to the phenomenon that the limit of the spectral radius of the finite computation trees can be less than the spectral radius of the infinite computation tree (which has no leaf nodes). See He et al. (2000) for analysis of a related discrepancy.

Means in non-WS models For the case where $\rho_\infty < 1 < \rho(\bar{R})$, the walk-sums for LBP variances converge absolutely (see proof of Proposition 25), but the walk-sums for the means do not. The reason is that LBP only computes a subset of the self-return walks for the variances but captures all the walks for the means. However, the series LBP computes for the means, corresponding to a particular ordering of walks, may still converge.

It is well known (Rusmevichientong and Van Roy, 2001) that once variances converge, the updates for the means follow a linear system. Consider (7) and (8) with $\hat{J}_{i\backslash j}$ fixed, then the LBP messages for the means $\Delta h = (\Delta h_{i\to j} \mid \{i,j\} \in E)$ follow a linear system update. For the parallel message schedule we can express this as:

$$\Delta h^{(n+1)} = L\, \Delta h^{(n)} + b \tag{15}$$

for some matrix L and some vector b. Convergence of this system depends on the spectral radius $\rho(L)$. However, it is difficult to analyze $\rho(L)$ since the matrix L depends on the converged values of the LBP variances. To improve convergence of the means, one can damp the message updates by modifying (8) as follows:

$$\Delta h_{i\to j}^{(n+1)} = (1-\alpha)\Delta h_{i\to j}^{(n)} + \alpha(-J_{ij}(\hat{J}_{i\backslash j}^{(n)})^{-1}\hat{h}_{i\backslash j}^{(n)}) \;\text{ with }\; 0 < \alpha \leq 1. \tag{16}$$

We have observed in experiments that for all the cases where variances converge we also obtain convergence of the means with enough damping of BP messages. We have also tried damping the updates for the ΔJ messages, but whether or not variances converge appears to be independent of damping. Apparently, it is the validity of the computation tree ($\rho_\infty < 1$) that is essential for convergence of both means and variances in damped versions of Gaussian LBP.

Example 3. We illustrate Proposition 25 on a simple example. Consider the 5-node cycle model from Figure 3(b). In Figure 8(b), for $\rho = .49$ we plot $\rho(R_n)$ vs. n (lower curve) and observe that $\lim_{n\to\infty}\rho(R_n) \approx .98 < 1$, and LBP converges. For $\rho = .51$ (upper curve), the model defined on the 5-node cycle is still valid but $\lim_{n\to\infty}\rho(R_n) \approx 1.02 > 1$ so LBP is ill-posed and does not converge.

As we mentioned, in non-walksummable models the series that LBP computes for the means is not absolutely convergent and may diverge even when variances converge. For our 4-cycle with a chord example in Figure 3(a), the region where variances converge but means diverge is very

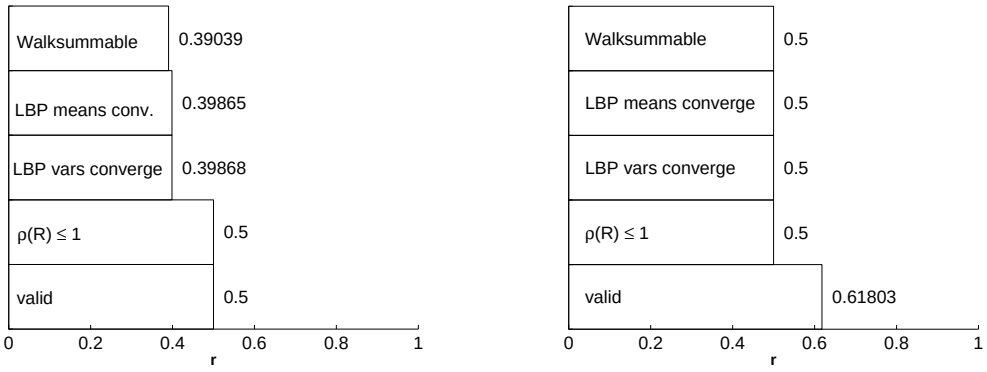


Figure 9: Critical regions for example models from Figure 3. (a) 4-cycle with a chord. (b) 5-cycle.

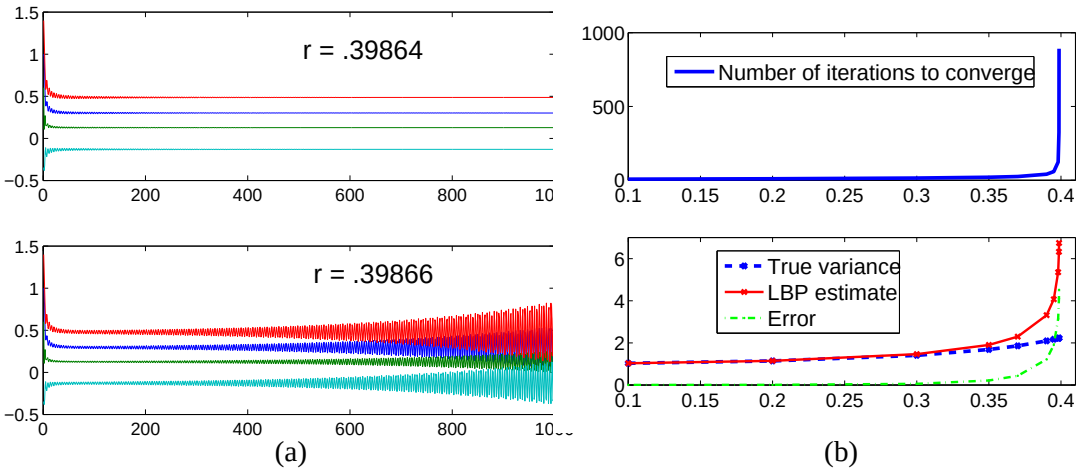


Figure 10: The 4-cycle with a chord example. (a) Convergence and divergence of the means near the LBP mean critical point. (b) Variance near the LBP variance critical point: (top) number of iterations for variances to converge, (bottom) true variance, LBP estimate and the error at node 1.

narrow, $r \approx .39865$ to $r \approx .39867$ (we use the parallel message schedule here; the critical point for the means is slightly higher using a serial schedule). In Figure 10(a) we show mean estimates vs. the iteration number on both sides of the LBP mean critical point for $r = 0.39864$ and for $r = 0.39866$. In the first case the means converge, while in the latter they slowly but very definitely diverge. The spectral radius of the linear system for mean updates in (15) for the two cases is $\rho(L) = 0.99717 < 1$ and $\rho(L) = 1.00157 > 1$ respectively. In the divergent example, all the eigenvalues of L have real components less than 1 (the maximum such real component is $0.8063 < 1$). Thus by damping we can force all the eigenvalues of L to enter the unit circle: the damped linear system is $(1-\alpha)I + \alpha L$. Using $\alpha = 0.9$ in (16) the means converge.

In Figure 10(b) we illustrate that near the LBP variance critical point, the LBP estimates become more difficult to obtain and their quality deteriorates dramatically. We consider the graph in Figure

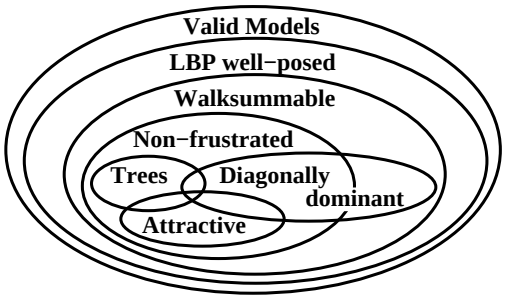


Figure 11: Venn diagram summarizing various subclasses of Gaussian models.

3(a) again as r approaches 0.39867, the critical point for the convergence of the variances. The picture shows that the number of iterations as well as the error in LBP variance estimates explode near the critical point. In the figure we show the variance at node 1, but similar behavior occurs at every node. In Figure 9, we summarize the critical points of both models from Figure 3.

6. Conclusion

We have presented a walk-sum interpretation of inference in Gaussian graphical models, which holds for a wide class of models that we call walk-summable. We have shown that walk-summability encompasses many classes of models which are considered "easy" for inference—trees, attractive, non-frustrated and diagonally dominant models—but also includes many models outside of these classes. A Venn diagram summarizing relations between these sets appears in Figure 11. We have also shown the equivalence of walk-summability to pairwise-normalizability.

We have established that in walk-summable models LBP is guaranteed to converge, for both means and variances, and that upon convergence the means are correct, whereas the variances only capture walk-sums over back-tracking walks. We have also used the walk-summability of valid (i.e., positive definite) models on trees to develop a more complete picture of LBP for non-walksummable models, relating variance convergence to validity of the LBP computation tree.

There are a variety of directions in which these results can be extended. One involves developing improved walk-sum algorithms that gather more walks than LBP does, to yield better variance estimates. Results along these lines—involving vectors of variables at each node as well as factor graph versions of LBP that group larger sets of variables—will be presented in a future publication. Another direction is to apply walk-sum analysis to other algorithms for Gaussian inference, for example, Chandrasekaran et al. are applying walk-sums to better understand the embedded trees algorithm (Sudderth et al., 2004).

Our current work is limited to Gaussian models, as walk-sums arise from the power series expansion for the matrix inverse. However, related expansions of correlations in terms of walks have been investigated for other models. Fisher (1967) developed an approximation to the pairwise correlations in Ising models based on self-avoiding walks. Brydges et al. (1983) use walk-sums for non-Gaussian classical and quantum spin-systems, where the weights of walks involve complicated multi-dimensional integrals. It would be very useful to develop ways to compute or approximate self-avoiding or non-Gaussian walk-sums efficiently and extend the walk-sum perspective to inference in a broader class of models.

Appendix A. Detailed Proofs

Proof of Proposition 1 *Proof of (i) ⇒ (ii).* We examine convergence of the matrix series in (ii) element-wise. First note that $(\bar{R}^l)_{ij}$ is an absolute walk-sum over all walks of length l from i to j:

$$(\bar{R}^l)_{ij} = \sum_{w:i\xrightarrow{l}j} |\phi(w)|$$

(there are a finite number of these walks so the sum is well-defined). Now, if (i) holds then using properties of absolute convergence we can order the sum $\sum_{w:i\to j}|\phi(w)|$ however we wish and it still converges. If we order walks by their length and then group terms for walks of equal lengths (each group has a finite number of terms) we obtain:

$$\sum_{w:i\to j}|\phi(w)| = \sum_l \sum_{w:i\xrightarrow{l}j}|\phi(w)| = \sum_l (\bar{R}^l)_{ij}\,. \tag{17}$$

Therefore, the series $\sum_l(\bar{R}^l)_{ij}$ converges for all i,j.

Proof of (ii) ⇒ (i). To show convergence of the sum $\sum_{w:i\to j}|\phi(w)|$ it is sufficient to test convergence for any convenient ordering of the walks. As shown in (17), $\sum_l(\bar{R}^l)_{ij}$ corresponds to one particular ordering of the walks which converges by (ii). Therefore, the walk-sums in (i) converge absolutely.

Proof of (ii) ⇔ (iii). This is a standard result in matrix analysis (Varga, 2000).

Proof of (iii) ⇔ (iv). Note that λ is an eigenvalue of $\bar{R}$ if and only if $1-\lambda$ is an eigenvalue of $I-\bar{R}$ ($\bar{R}x=\lambda x \Leftrightarrow (I-\bar{R})x=(1-\lambda)x$). Therefore, $\lambda_{\min}(I-\bar{R}) = 1-\lambda_{\max}(\bar{R})$. According to the Perron-Frobenius theorem, $\rho(\bar{R})=\lambda_{\max}(\bar{R})$ because $\bar{R}$ is non-negative. Thus, $\rho(\bar{R}) = 1-\lambda_{\min}(I-\bar{R})$ and we have that $\rho(\bar{R})<1 \Leftrightarrow \lambda_{\min}(I-\bar{R})>0$. □

Proof of Corollary 3 We will show that for any non-frustrated model there exists a diagonal D with $D_{ii}=\pm 1$, that is, a signature matrix, such that $DRD=\bar{R}$. Hence, R and $\bar{R}$ have the same eigenvalues, because $DRD=DRD^{-1}$ is a similarity transform which preserves the eigenvalues of a matrix. It follows that $I-R\succ 0$ implies $I-\bar{R}\succ 0$ and walk-summability of J by Proposition 1(iv).

Now we describe how to construct a signature similarity which makes R attractive for non-frustrated models. We show how to split the vertices into two sets V^+ and V^- such that negating V^- makes the model attractive. Find a spanning tree T of the graph G. Pick a node i. Assign it to V^+. For any other node j, there is a unique path to i in T. If the product of edge weights along the path is positive, then assign j to V^+, otherwise to V^-. Now, since the model is non-frustrated, all edges $\{j,k\}$ in G such that $j,k\in V^+$ are positive, all edges with $j,k\in V^-$ are positive, and all edges with $j\in V^+$ and $k\in V^-$ are negative. This can be seen by constructing the cycle that goes from j to i to k in T and crosses the edge $\{k,j\}$ to close itself. If $j,k\in V^+$ then the paths j to i and i to k have a positive weight, hence in order for the cycle to have a positive weight, the last step $\{k,j\}$ must also have a positive weight. The other two cases are similar. Now let D be diagonal with $D_{ii}=1$ for $i\in V^+$, and $D_{ii}=-1$ for $i\in V^-$. Then $DRD=\begin{bmatrix} R_{V^+} & -R_{V^+,V^-} \\ -R_{V^-,V^+} & R_{V^-}\end{bmatrix}\geq 0$, that is, $DRD=\bar{R}$. □

Proof of Proposition 4 *Proof of WS ⇒ (i).* WS is equivalent to $\rho(\bar{R})<1$ by Proposition 1. But $\rho(R)\leq\rho(\bar{R})$ by (11). Hence, $\rho(\bar{R})<1\Rightarrow\rho(R)<1$.

Proof of (i) $\Rightarrow$ (ii). Given $J = I - R$, it holds that $\lambda_{\min}(J) = 1 - \lambda_{\max}(R)$. Also, $\lambda_{\max}(R) \leq \rho(R)$. Hence, $\lambda_{\min}(J) = 1 - \lambda_{\max}(R) \geq 1 - \rho(R) > 0$ for $\rho(R) < 1$.

Proof of (i) $\Rightarrow$ (iii). This is a standard result in matrix analysis. $\square$

Proof of Proposition 10 Assume that G is connected (otherwise we apply the proof to each connected component, and the spectral radii are the maxima over the respective connected components). We prove that $\rho(\bar{R}) = \rho(\hat{R})$. By the Perron-Frobenius theorem, there exists a positive vector x such that $\bar{R}x = \rho(\bar{R})x$. Let $\hat{x} = (x;x)$. Then $\hat{R}\hat{x} = \rho(\bar{R})\hat{x}$ because

$$(\hat{R}\hat{x})_{\pm} = (R_+ + R_-)x = \bar{R}x = \rho(\bar{R})x.$$

Hence, $\rho(\bar{R})$ is an eigenvalue of $\hat{R}$ with positive eigenvector $\hat{x}$. First suppose that $\hat{G}$ is connected. Then, by the Perron-Frobenius theorem, $\rho(\bar{R}) = \rho(\hat{R})$ because $\hat{R}$ has a unique positive eigenvector which has eigenvalue equal to $\rho(\hat{R})$. Now, $\hat{J} = I - \hat{R} \succ 0 \Leftrightarrow \hat{J}$ is WS $\Leftrightarrow \rho(\hat{R}) < 1 \Leftrightarrow \rho(\bar{R}) \leq 1 \Leftrightarrow J = I - R$ is WS. If $\hat{G}$ is disconnected then $\hat{R}$ is a block-diagonal matrix with two copies of $\bar{R}$ (after relabeling the nodes), so $\rho(\hat{R}) = \rho(\bar{R})$. $\square$

Proof of Proposition 11 We partition walk-sums into sums over "even" and "odd" walks according to the number of negative edges crossed by the walk. Thus a walk w is even if $\phi(w) > 0$ and is odd if $\phi(w) < 0$. The graph $\hat{G}$ is defined so that every walk from i_+ to j_+ is even and every walk from i_+ to j_- is odd. Thus,

$$\begin{aligned} P_{ij} &= \sum_{\text{even } w:i\to j} \phi(w) + \sum_{\text{odd } w:i\to j} \phi(w) \\ &= \sum_{w:i_+\to j_+} \hat{\phi}(w) - \sum_{w:i_+\to j_-} \hat{\phi}(w) \\ &= \hat{P}_{i_+,j_+} - \hat{P}_{i_+,j_-}\,. \end{aligned}$$

The second part of the the proposition follows by similar logic. Now we classify a walk as even if $h_{w_0}\phi(w) > 0$ and as odd if $h_{w_0}\phi(w) < 0$. Note also that setting $\hat{h} = (h_+;h_-)$ has the effect that all walks with $h_{w_0} > 0$ begin in V_+ and all walks with $h_{w_0} < 0$ begin in V_-. Consequently, every even walk ends in V_+ and every odd walk ends in V_-. Thus,

$$\begin{aligned} \mu_i &= \sum_{\text{even } w:*\to i} h_*\phi(w) + \sum_{\text{odd } w:*\to i} h_*\phi(w) \\ &= \sum_{w:*\to i_+} \hat{h}_*\hat{\phi}(w) - \sum_{w:*\to i_-} \hat{h}_*\hat{\phi}(w) \\ &= \hat{\mu}_{i_+} - \hat{\mu}_{i_-} \;\square \end{aligned}$$

Proof of Proposition 12 Take J_1 and J_2 pairwise-normalizable. Take any $\alpha, \beta \geq 0$ such that at least one of them is positive. Then $\alpha J_1 + \beta J_2$ is also pairwise-normalizable simply by taking the same weighted combinations of each of the J_e matrices for J_1 and J_2. Setting $\beta = 0$ shows that $\mathcal{J}_{PN}$ is a cone, and setting $\beta = 1 - \alpha$ shows convexity. The cone is pointed since it is a subset of the cone of semidefinite matrices, which is pointed. $\square$

Proof of Proposition 13 *Proof of PN ⇒WS.* It is evident that any J matrix which is pairwise-normalizable is positive definite. Furthermore, reversing the sign of the partial correlation coefficient on edge e simply negates the off-diagonal element of J_e which does not change the value of $\det J_e$ so that we still have $J_e \succeq 0$. Thus, we can make all the negative coefficients positive and the resulting model $I - \bar{R}$ is still pairwise-normalizable and hence positive definite. Then, by Proposition 1(iv), $J = I - R$ is walk-summable.

Proof of WS ⇒PN. Given a walk-summable model $J = I - R$ we construct a pairwise-normalized representation of the information matrix. We may assume the graph is connected (otherwise, we may apply the following construction for each connected component of the graph). Hence, by the Perron-Frobenius theorem there exists a positive eigenvector $x > 0$ of $\bar{R}$ such that $\bar{R}x = \lambda x$ and $\lambda = \rho(\bar{R}) > 0$. Given (x, λ) we construct a representation $J = \sum_e [J_e]$ where for $e = \{i, j\}$ we set:

$$J_e = \begin{pmatrix} \frac{|r_{ij}|x_j}{\lambda x_i} & -r_{ij} \\ -r_{ij} & \frac{|r_{ij}|x_i}{\lambda x_j} \end{pmatrix}.$$

This is well-defined (there is no division by zero) since x and λ are positive. First, we verify that $J = \sum_{e \in E} [J_e]$. It is evident that the off-diagonal elements of the edge matrices sum to $-R$. We check that the diagonal elements sum to one:

$$\sum_e [J_e]_{ii} = \frac{1}{\lambda x_i} \sum_j |r_{ij}| x_j = \frac{(\bar{R}x)_i}{\lambda x_i} = \frac{(\lambda x)_i}{\lambda x_i} = 1.$$

Next, we verify that each J_e is positive definite. This matrix has positive diagonal and determinant

$$\det J_e = \left(\frac{|r_{ij}|x_j}{\lambda x_i} \right) \left(\frac{|r_{ij}|x_i}{\lambda x_j} \right) - (-r_{ij})^2 = r_{ij}^2 \left(\frac{1}{\lambda^2} - 1 \right) > 0.$$

The inequality follows from walk-summability because $0 < \lambda < 1$ and hence $\left(\frac{1}{\lambda^2} - 1\right) > 0$. Thus, $J_e \succ 0$. □

Proof of Proposition 14 Let $a_i = J_{ii} - \sum_{j \neq i} |J_{ij}|$. Note that $a_i > 0$ follows from diagonal dominance. Let $\deg(i)$ denote the degree of node i in G. Then, $J = \sum_{e \in E} [J_e]$ where for edge $e = \{i, j\}$ we set

$$J_e = \begin{pmatrix} |J_{ij}| + \frac{a_i}{\deg(i)} & J_{ij} \\ J_{ij} & |J_{ij}| + \frac{a_j}{\deg(j)} \end{pmatrix}$$

with all other elements of $[J_e]$ set to zero. Note that:

$$\sum_e [J_e]_{ii} = \sum_{j \in \mathcal{N}(i)} \left(|J_{ij}| + \frac{a_i}{\deg(i)} \right) = a_i + \sum_{j \in \mathcal{N}(i)} |J_{ij}| = J_{ii}\,.$$

Also, J_e has positive diagonal elements and has determinant $\det(J_e) > 0$. Hence, $J_e \succ 0$. Thus, J is pairwise-normalizable. □

Proof of Proposition 16 To calculate the walk-sum for multiple-revisit self-return walks in $T_{i\backslash j}$, we can use the single-revisit counterpart:

$$\gamma_{i\backslash j} = \phi(i \to i \mid T_{i\backslash j}) = \frac{1}{1 - \phi\left(i \overset{\backslash i}{\to} i \mid T_{i\backslash j}\right)} \ . \tag{18}$$

Now, we decompose the single-revisit walks in the subtree $T_{i\backslash j}$ in terms of the possible first step of the walk (i,k), where $k \in \mathcal{N}(i)\backslash j$. Hence,

$$\phi(i \overset{\backslash i}{\to} i \mid T_{i\backslash j}) = \sum_{k \in \mathcal{N}(i)\backslash j} \phi(i \overset{\backslash i}{\to} i \mid T_{k\to i}). \tag{19}$$

Using (14), (18), and (19), we are able to represent the walk-sum $\phi(j \overset{\backslash j}{\to} j \mid T_{i\to j})$ in $T_{i\to j}$ in terms of the walk-sums $\phi(i \overset{\backslash i}{\to} i \mid T_{k\to i})$ on smaller subtrees $T_{k\to i}$. This is the basis of the recursive calculation:

$$\alpha_{i\to j} = r_{ij}^2 \frac{1}{1 - \sum_{k \in \mathcal{N}(i)\backslash j} \alpha_{k\to i}} \ .$$

These equations look strikingly similar to the belief propagation updates. Combining (7) and (8) from Section 2.1 we have:

$$-\Delta J_{i\to j} = J_{ij}^2 \frac{1}{J_{ii} + \sum_{k \in \mathcal{N}(i)\backslash j} \Delta J_{k\to i}} \ .$$

It is evident that the recursive walk-sum equations can be mapped exactly to belief propagation updates. In normalized models $J_{ii} = 1$. We have the message update $\alpha_{i\to j} = -\Delta J_{i\to j}$, and the variance estimate in the subtree $T_{i\backslash j}$ is $\gamma_{i\backslash j} = \hat{J}^{-1}_{i\backslash j}$. □

Proof of Proposition 17 A multiple-revisit walk in $T_{i\backslash j}$ can be written in terms of single-visit walks:

$$\phi_h(* \to i \mid T_{i\backslash j}) = \left(h_i + \phi_h(* \overset{\backslash i}{\to} i \mid T_{i\backslash j})\right) \phi(i \to i \mid T_{i\backslash j}).$$

We already have $\gamma_{i\backslash j} = \phi(i \to i \mid T_{i\backslash j})$ from (18). The remaining term $\phi_h(* \overset{\backslash i}{\to} i \mid T_{i\backslash j})$ can be decomposed by the subtrees in which the walk lives:

$$\phi_h(* \overset{\backslash i}{\to} i \mid T_{i\backslash j}) = \sum_{k \in \mathcal{N}(i)\backslash j} \phi_h(* \overset{\backslash i}{\to} i \mid T_{k\to i}).$$

Thus we have the recursion:

$$\beta_{i\to j} = r_{ij}\gamma_{i\backslash j}(h_i + \sum_{k \in \mathcal{N}(i)\backslash j} \beta_{k\to i}).$$

To compare this to the Gaussian BP updates, let us combine (7) and (8) in Section 2.2:

$$\Delta h_{i\to j} = -J_{ij}\hat{J}^{-1}_{i\backslash j}\left(h_i + \sum_{k \in \mathcal{N}(i)\backslash j} \Delta h_{k\to i}\right).$$

Thus BP updates for the means can also be mapped exactly into recursive walk-sum updates via $\beta_{i\to j} = \Delta h_{i\to j}$. □

Proof of Lemma 18 First, we note that for every walk w which ends at the root node of $T_i^{(n)}$ there is a corresponding walk in G which ends at i. The reason is that the neighbors of a given node j in $T_i^{(n)}$ correspond to a subset of the neighbors of j in G. Hence, for each step (w_k, w_{k+1}) of the walk in $T_i^{(n)}$ there is a corresponding step in G.

Next, we show that every walk $w = (w_0, \ldots, w_l)$ in G is contained in $T_{w_l}^{(n)}$ for some n. First consider the parallel message schedule, for which the computation tree $T_{w_l}^{(n)}$ grows uniformly. Then for any walk in G that ends at w_l and has length n there is a walk in $T_{w_l}^{(n)}$ that ends at the root.

The intuition for other message schedules is that every step (i, j) of the walk will appear eventually in any proper message schedule $\mathcal{M}$. A formal proof is somewhat technical. First we unwrap the walk w into a tree T_w rooted at w_l in the following way: start at w_l, the end of the walk, and traverse the walk in reverse. First add the edge $\{w_l, w_{l-1}\}$ to T_w. Now, suppose we are at node w_k in T_w and the next step in w is $\{w_k, w_{k-1}\}$. If w_{k-1} is already a neighbor of w_k in T_w then set the current node in T_w to w_{k-1}. Otherwise create a new node w_{k-1} and add the edge to T_w. It is clear that loops are never made in this procedure, so T_w is a tree.

We now show for any proper message schedule $\mathcal{M}$ that T_w is part of the computation tree $T_{w_l}^{(n)}$ for some n. Pick a leaf edge $\{i_1, j_1\}$ of T_w. Since $\{\mathcal{M}^{(n)}\}$ is proper, there exist n_1 such that $(i_1, j_1) \in \mathcal{M}^{(n_1)}$. Now $(i_1, j_1) \in T_{i_1 \to j_1}^{(n_1)}$, and the edge appears at the root of $T_{i_1 \to j_1}^{(n_1)}$. Also, $T_{i_1 \to j_1}^{(n_1)} \subset T_{i_1 \to j_1}^{(m)}$ for $m > n_1$, so this holds for all subsequent steps as well. Now remove $\{i_1, j_1\}$ from T_w and pick another leaf edge $\{i_2, j_2\}$. Again, since $\{\mathcal{M}^{(n)}\}$ is proper, there exist $n_2 > n_1$ such that $(i_2, j_2) \in \mathcal{M}^{(n_2)}$. Remove $\{i_2, j_2\}$ from T_w, and continue similarly. At each such point n_k of eliminating some new edge $\{i_k, j_k\}$ of T_w, the whole eliminated subtree of T_w extending from $\{i_k, j_k\}$ has to belong to $T_{i_k \to j_k}^{(n_k)}$. Continue until just the root of T_w remains at step n. Now the computation tree $T_{w_l}^{(n)}$ (which is created by splicing together $T_{i \to j}^{(n)}$ for all edges (i, j) coming into the root of T_w) contains T_w, and hence it contains the walk w. □

Proof of Lemma 20 This result comes as an immediate corollary of Proposition 28, which states that $\rho(R_i^{(n)}) \leq \rho(\bar{R})$ (here $R_i^{(n)}$ is the partial correlation matrix for $T_i^{(n)}$). For WS models, $\rho(\bar{R}) < 1$ and the result follows. □

Proof of Lemma 23 The fact that the sequence $\{\rho(R_i^{(n)})\}$ is bounded by $\rho(\bar{R})$ is a nontrivial fact, proven in Appendix B using a *k-fold graph* construction. To prove monotonicity, note first that for trees $\rho(R_i^{(n)}) = \rho(\bar{R}_i^{(n)})$. Also, note that all of the variables in the computation tree $T_i^{(n)}$ are also present in T_i^{n+1}. We zero-pad $\bar{R}_i^{(n)}$ to make it the same size as $\bar{R}_i^{(n+1)}$ (this does not change the spectral radius). Then it holds that $\bar{R}_i^{(n)} \leq \bar{R}_i^{(n+1)}$ element-wise. Using (11), it follows that $\rho(\bar{R}_i^{(n)}) \leq \rho(\bar{R}_i^{(n+1)})$, establishing monotonicity. □

Proof of Lemma 24 Let $T_i^{(n)}(\mathcal{M})$ denote the nth computation tree under a proper message schedule $\mathcal{M}$ rooted at node i. We use the following simple extension of Lemma 18: Let $T_i^{(n)}(\mathcal{M}_1)$ be the nth computation tree rooted at i under message schedule $\mathcal{M}_1$. Take any node in $T_i^{(n)}(\mathcal{M}_1)$ which is a replica of node j in G. Then there exists m such that $T_i^{(n)}(\mathcal{M}_1) \subset T_j^{(m)}(\mathcal{M}_2)$, where $\mathcal{M}_2$ is another

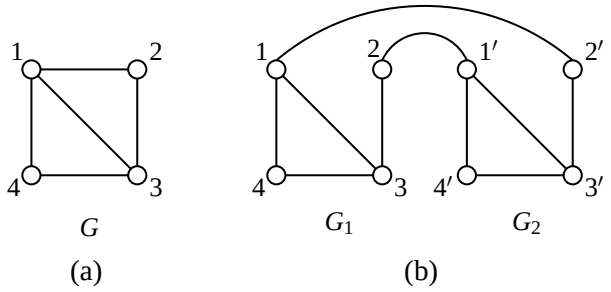


Figure 12: Illustration of (a) graph G and (b) a 2-fold graph of G.

message schedule. The proof parallels that of Lemma 18: the tree $T_i^{(n)}(\mathcal{M}_1)$ has a finite number of edges, and we use induction adding one edge at a time.

Consider message schedule $\mathcal{M}_1$. By Lemma 23, $\rho_i \triangleq \lim_{n\to\infty} \rho(R_i^{(n)}(\mathcal{M}_1))$ exists. For any ε pick an L such that for $n \geq L$ it holds that $|\rho(R_i^{(n)}(\mathcal{M}_1)) - \rho_i| \leq \frac{\varepsilon}{2}$. Pick a replica of node j inside $T_i^{(L)}(\mathcal{M}_1)$. Then using the property from the previous paragraph, there exists M such that $T_i^{(L)}(\mathcal{M}_1) \subset T_j^{(M)}(\mathcal{M}_2)$. Similarly there exists N such that $T_j^{(M)}(\mathcal{M}_2) \subset T_i^{(N)}(\mathcal{M}_1)$. It follows that $\overline{R}_i^{(L)}(\mathcal{M}_1) \leq \overline{R}_j^{(M)}(\mathcal{M}_2) \leq \overline{R}_i^{(N)}(\mathcal{M}_1)$, where we zero-pad the first two matrices to have the same size as the last one. Then, $\rho(\overline{R}_i^{(L)}(\mathcal{M}_1)) \leq \rho(\overline{R}_j^{(M)}(\mathcal{M}_2)) \leq \rho(\overline{R}_i^{(N)}(\mathcal{M}_1))$. Then it holds that $\rho_i - \frac{\varepsilon}{2} \leq \rho(\overline{R}_j^{(M)}(\mathcal{M}_2)) \leq \rho_i + \frac{\varepsilon}{2}$. Hence, $|\rho(\overline{R}_j^{(M)}(\mathcal{M}_2)) - \rho_i| \leq \varepsilon$, and $\lim_{n\to\infty} \rho(\overline{R}_j^{(n)}(\mathcal{M}_2)) = \rho_i$. □

Appendix B. K-fold Graphs and Proof of Boundedness of $\rho(R_i^{(n)})$.

Consider an arbitrary graph $G = (V, E)$. Suppose that we have a pairwise MRF defined on G with self potentials $\psi_i(x_i)$, for $v_i \in V$ and pairwise potentials $\psi_{ij}(x_i, x_j)$ for $(v_i, v_j) \in E$. We construct a family of K-fold graphs based on G as follows:

1. Create K disconnected copies G_k, $k \in \{1, .., K\}$ of G, with nodes $v_i^{(k)}$, and edges $(v_i^{(k)}, v_j^{(k)})$. The nodes and the edges of G_k are labeled in the same way as the ones of G. The potentials ψ_i and ψ_{ij} are copied to the corresponding nodes and edges in all G_k.

2. Pick some pair of graphs G_k, G_l, and choose an edge (v_i, v_j) in G. We flip the corresponding edges in G_k and G_l, edges $(v_i^{(k)}, v_j^{(k)})$ and $(v_i^{(l)}, v_j^{(l)})$ become $(v_i^{(k)}, v_j^{(l)})$ and $(v_i^{(l)}, v_j^{(k)})$. The pairwise potentials are adjusted accordingly.

3. Repeat step 2 an arbitrary number of times for a different pair of graphs G_k, or a different edge in G.

An illustration of the procedure appears in Figure 12. The original graph G is a 4-cycle with a chord. We create a 2-fold graph based on G by flipping the edges $(1,2)$ in G_1 and $(1', 2')$ in G_2.

Now we apply the K-fold graph construction to Gaussian MRF models. Suppose that we have a model with information parameters J and h on G. Suppose that J is normalized to have unit-diagonal. Let G^K be a K-fold graph based on G with the information matrix J^K (which is also

unit-diagonal by construction). Also, let $T_i^{(n)}$ be the nth computation tree for the original graph, and $J_i^{(n)}$ the corresponding information matrix (also unit-diagonal). Let $R = I - J$, $R^K = I^K - J^K$, and $R_i^{(n)} = I^{(n)} - J_i^{(n)}$ (here I, I^K, and $I^{(n)}$ are identity matrices of appropriate dimensions).

Lemma 26 (Spectral radii of R and R^K) *For any K-fold graph G^K based on G: $\rho(\bar{R}^K) = \rho(\bar{R})$.*

Proof. Suppose that G is connected (otherwise apply the proof to each connected component of G, and the spectral radius for G will be the maximum of the spectral radii for the connected components).

Then, by the Perron-Frobenius theorem there exists a vector $x > 0$ such that $\bar{R}x = \rho(\bar{R})x$. Create a K-fold vector x^K by copying entry x_i into each of the K corresponding entries of x^K. Then x^K is positive, and it also holds that $\bar{R}^K x^K = \rho(\bar{R})x^K$ (since the local neighborhoods in G and G^K are the same). Now $\bar{R}^K$ is a non-negative matrix, and x^K is a positive eigenvector, hence it achieves the spectral radius of $\bar{R}^K$ by the Perron-Frobenius theorem. Thus, $\rho(\bar{R}) = \rho(\bar{R}^K)$. □

The construction of a K-fold graph based on G has parallels with the computation tree on G. The K-fold graph is locally equivalent to G and the computation tree, except for its leaf nodes, is also locally equivalent to G. We show next that the computation tree $T_i^{(n)}$ is contained in some G^K for K large enough.

Lemma 27 (K-fold graphs and computation trees) *Consider a computation tree $T_i^{(n)}$ corresponding to graph G. There exists a K-fold graph G^K, which contains $T_i^{(n)}$ as a subgraph, for K large enough.*

Proof. We provide a simple construction of a K-fold graph, making no attempt to minimize K. Let $T_i^{(n)} = (V_n, E_n)$. Each node $v' \in V_n$ corresponds to some node $v \in V$ in G. We create a K-fold graph G^K by making a copy $G_{v'}$ of G for every node $v' \in T_i^{(n)}$. Hence $K = |V_n|$. For each edge $(u', v') \in E_n$ in the computation tree, we make an edge flip between nodes in graphs $G_{u'}$ and $G_{v'}$ that correspond to u and v in G. This operation is well-defined because edges in $T_i^{(n)}$ that map to the same edge in G do not meet. Thus, the procedure creates G^K which contains $T_i^{(n)}$ as a subgraph. □

Finally, we use the preceding lemmas to prove a bound on the spectral radii of the matrices $R_i^{(n)}$ for the computation tree $T_i^{(n)}$.

Proposition 28 (Bound on $\rho(R_i^{(n)})$) *For computation tree $T_i^{(n)}$: $\rho(R_i^{(n)}) \leq \rho(\bar{R})$.*

Proof. Consider a computation tree $T_i^{(n)}$. Recall that $\rho(R_i^{(n)}) = \rho(\bar{R}_i^{(n)})$, since $T_i^{(n)}$ is a tree. Use Lemma 27 to construct a K-fold graph G^K which has $T_i^{(n)}$ as a subgraph. Zero-padding $\bar{R}_i^{(n)}$ to have the same size as $\bar{R}^K$, it holds that $\bar{R}_i^{(n)} \leq \bar{R}^K$. Since $\bar{R}_i^{(n)} \leq \bar{R}^K$, using (11) and Lemma 26 we have: $\rho(R_i^{(n)}) \leq \rho(\bar{R}^K) = \rho(\bar{R})$. □

References

E. G. Boman, D. Chen, O. Parekh, and S. Toledo. On factor width and symmetric H-matrices. *Linear Algebra and its Applications*, 405, 2005.

D. C. Brydges, J. Frohlich, and A. D. Sokal. The random-walk representation of classical spin systems and correlation inequalities. *Communications in mathematical physics*, 91, 1983.

V. Chandrasekaran, J. Johnson, and A. Willsky. Walk-sum analysis and convergence of embedded subgraph algorithms. In preparation.

R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.

M. Fisher. Critical temperatures of anisotropic Ising lattices II, general upper bounds. *Physical Review*, 1967.

R. Godement. *Analysis I*. Springer-Verlag, 2004.

J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished manuscript, 1971.

L. He, X. Liu, and G. Strang. Laplacian eigenvalues of growing tees. In *Conf. on Math. Theory of Networks and Systems*, 2000.

R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

R. Horn and C. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

A. T. Ihler, J. W. Fisher III, and A. S. Willsky. Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research*, 6, May 2005.

J. Johnson. Walk-summable Gauss-Markov random fields. Unpublished manuscript, available at `http://www.mit.edu/people/jasonj`, December 2001.

J. Johnson, D. Malioutov, and A. Willsky. Walk-sum interpretation and analysis of Gaussian belief propagation. In *Advances in Neural Information Processing Systems*, 2006.

B. Jones and M. West. Covariance decomposition in undirected Gaussian graphical models. *Biometrika*, 92(4), 2005.

S. Kirkland, J. J. McDonald, and M. Tsatsomeros. Sign patterns that require positive eigenvalues. *Linear and Multilinear Algebra*, 41, 1996.

S. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, 1996.

C. Moallemi and B. Van Roy. Consensus propagation. In *Advances in Neural Information Processing Systems*, 2006a.

C. Moallemi and B. Van Roy. Convergence of the min-sum message passing algorithm for quadratic optimization. Technical Report, March 2006b.

J. Mooij and H. Kappen. Sufficient conditions for convergence of loopy belief propagation. In *Proc. Uncertainty in Artificial Intelligence*, 2005.

J. Pearl. *Probabilistic inference in intelligent systems*. Morgan Kaufmann, 1988.

K. Plarre and P. Kumar. Extended message passing algorithm for inference in loopy Gaussian graphical models. In *Ad Hoc Networks*, 2004.

W. Rudin. *Principles of Mathematical Analysis*. McGraw Hill, 3rd edition, 1976.

P. Rusmevichientong and B. Van Roy. An analysis of belief propagation on the turbo decoding graph with Gaussian densities. *IEEE Trans. Information Theory*, 48(2), 2001.

T. Speed and H. Kiiveri. Gaussian Markov distributions over finite graphs. *The Annals of Statistics*, 14(1), 1986.

E. Sudderth, M. Wainwright, and A. Willsky. Embedded trees: Estimation of Gaussian processes on graphs with cycles. *IEEE Trans. Signal Processing*, 52(11), November 2004.

S. Tatikonda and M. Jordan. Loopy belief propagation and Gibbs measures. In *Uncertainty in Artificial Intelligence*, 2002.

R. S. Varga. *Matrix iterative analysis*. Springer-Verlag, 2000.

M. Wainwright, T. Jaakkola, and A. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Trans. Information Theory*, 49(5), 2003.

Y. Weiss and W. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, 13, 2001.

J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring AI in the new millennium*, 2003.