Journal of Machine Learning Research 18 (2018) 1-38 Submitted 11/16; Revised 2/18; Published 4/18

Convergence Analysis of Distributed Inference with Vector-Valued Gaussian Belief Propagation

Jian Du JIAND@ANDREW.CMU.EDU
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Shaodan Ma SHAODANMA@UMAC.MO
Department of Electrical and Computer Engineering
University of Macau
Avenida da Universidade, Taipa, Macau

Yik-Chung Wu YCWU@EEE.HKU.HK
Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong

Soummya Kar SOUMMYAK@ANDREW.CMU.EDU
José M. F. Moura MOURA@ANDREW.CMU.EDU
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Qiang Liu

Abstract

This paper considers inference over distributed linear Gaussian models using factor graphs and Gaussian belief propagation (BP). The distributed inference algorithm involves only local computation of the information matrix and of the mean vector, and message passing between neighbors. Under broad conditions, it is shown that the message information matrix converges to a unique positive definite limit matrix for arbitrary positive semidefinite initialization, and it approaches an arbitrarily small neighborhood of this limit matrix at an exponential rate. A necessary and sufficient convergence condition for the belief mean vector to converge to the optimal centralized estimator is provided under the assumption that the message information matrix is initialized as a positive semidefinite matrix. Further, it is shown that Gaussian BP always converges when the underlying factor graph is given by the union of a forest and a single loop. The proposed convergence condition in the setup of distributed linear Gaussian models is shown to be strictly weaker than other existing convergence conditions and requirements, including the Gaussian Markov random field based walk-summability condition, and applicable to a large class of scenarios.

Keywords: Graphical Model, Large-Scale Networks, Linear Gaussian Model, Markov Random Field, Walk-summability.

1. Introduction

Inference based on a set of measurements from multiple agents on a distributed network is a central issue in many problems. While centralized algorithms can be used in small-scale networks, they face difficulties in large-scale networks, imposing a heavy communication burden when all the data is to be transported to and processed at a central processing unit. Dealing with highly distributed data has been recognized by the U.S. National Research Council as one of the big challenges for processing big data (National Research Council, 2013). Therefore, distributed inference techniques that only involve local communication and computation are important for problems arising in distributed networks.

In large-scale linear parameter learning with Gaussian measurements, Gaussian Belief Propagation (BP) (Weiss and Freeman, 2001a) provides an efficient distributed algorithm for computing the marginal means of the unknown parameters, and it has been adopted in a variety of topics including image interpolation (Xiong et al., 2010), distributed power system state inference (Hu et al., 2011), distributed beamforming (Ng et al., 2008), distributed synchronization (Du and Wu, 2013b), fast solver for system of linear equations (Shental et al., 2008a), distributed rate control in ad-hoc networks (Zhang et al., 2010), factor analyzer network (Frey, 1999), sparse Bayesian learning (Tan and Li, 2010), inter-cell interference mitigation (Lehmann, 2012), and peer-to-peer rating in social networks (Bickson and Malkhi, 2008).

Although with great empirical success (Murphy et al., 1999), it is known that a major challenge that hinders BP is the lack of theoretical guarantees of convergence in loopy networks (Chertkov and Chernyak, 2006; Gómez et al., 2007). Convergence of other forms of loopy BP are analyzed by Ihler et al. (2005), Mooij and Kappen (2005, 2007), Noorshams and Wainwright (2013), and Ravanbakhsh and Greiner (2015), but their analyses are not directly applicable to Gaussian BP. Sufficient convergence conditions for Gaussian BP have been developed in Weiss and Freeman (2001a); Malioutov et al. (2006); Moallemi and Roy (2009a); Su and Wu (2015) when the underlying Gaussian distribution is expressed in terms of pairwise connections between *scalar* variables, i.e., it is a Markov random field (MRF). However, depending on how the underlying joint Gaussian distribution is factorized, Gaussian BP may exhibit different convergence properties as different factorizations (different Gaussian models) lead to fundamentally different recursive update structures. In this paper, we study the convergence of Gaussian BP derived from the distributed linear Gaussian model. The motivation is twofold. From the factorization viewpoint, by specifically employing a factorization based on the linear Gaussian model, we are able to bypass difficulties in existing convergence analyses ((Malioutov et al., 2006) and references therein) based on Gaussian Markov random field factorization. From the distributed inference viewpoint, the linear Gaussian model and associated message passing requirements for implementing the Gaussian BP readily conform to the physical network topology arising in large-scale networks such as in (Hu et al., 2011; Ng et al., 2008; Du and Wu, 2013b; Shental et al., 2008a; Zhang et al., 2010; Frey, 1999; Tan and Li, 2010; Lehmann, 2012; Bickson and Malkhi, 2008), thus it is practically important.

Recently, Giscard et al. (2012, 2013, 2016) present a path-sum method to compute the information matrix inverse of a joint Gaussian distribution. Then, the marginal mean is obtained using the information matrix inverse. The path-sum method converges for an

arbitrary valid Gaussian model, however, it is not clear how to adapt it to the distributed and parallel inference setup. In contrast, Gaussian BP is a parallel and fully distributed method that computes the marginal means by computing only the block diagonal elements of the information matrix inverse. Though the block diagonal elements computed by Gaussian BP may not be correct, it is shown that the belief mean still converges to the correct value once Gaussian BP converges. This explains the popularity of Gaussian BP in distributed inference applications, even though its convergence properties are not fully understood.

To fill this gap, this paper studies the convergence of Gaussian BP for linear Gaussian models. Specifically, for the first time, by establishing certain contractive properties of the distributed information matrix (inverse covariance matrix) updates with respect to the Birkhoff metric, we show that, with arbitrary positive semidefinite (p.s.d.) initial message information matrix, the belief covariance for each local variable converges to a unique positive definite limit, and it approaches an arbitrarily small neighborhood of this limit matrix at an exponential rate. Consequently, the recursive equation for the message mean, which depends on the information matrix, can be reduced to a linear recursive equation. Further, we derive a necessary and sufficient convergence condition for this linear recursive equation under the assumption that the initial message information matrix is p.s.d. Furthermore, we show that, when the structure of the factor graph is the union of a single loop and a forest, Gaussian BP always converges. Finally, it is demonstrated that the proposed convergence condition for the linear Gaussian model encompasses the walk-summable convergence condition for Gaussian MRFs (Malioutov et al., 2006).

Note that there exist other distributed estimation frameworks, e.g., consensus+innovations (Kar and Moura, 2013; Kar et al., 2013) and diffusion algorithms (Cattivelli and Sayed, 2010) that enable distributed estimation of parameters and processes in multi-agent networked environments. The consensus+innovation algorithms converge in mean square sense to the centralized optimal solution under the assumption of global observability of the (aggregate) sensing model and connectivity (on the average) of the inter-agent communication network. In particular, these algorithms allow the communication or message exchange network to be different from the physical coupling network of the field being estimated where either networks can be arbitrarily connected with cycles. The results in Kar and Moura (2013); Kar et al. (2013) imply that the unknown field or parameter can be reconstructed completely at each agent in the network. For large-scale networks with high dimensional unknown variable, it may be impractical though to estimate all the unknowns at every agent. Reference (Kar, 2010, section 3.4) develops approaches to address this problem, where under appropriate conditions, each agent can estimate only a subset of the unknown parameter variables. This paper studies a different distributed inference problem where each agent learns only its own unknown random variables; this leads to lower dimensional data exchanges between neighbors.

The rest of this paper is organized as follows. Section 2 presents the system model for distributed inference. Section 3 derives the vector-valued distributed inference algorithm based on Gaussian BP. Section 4 establishes convergence conditions, and Section 5 discloses the relationship between the derived results and existing convergence conditions of Gaussian BP. Finally, Section 6 presents our conclusions.

Notation: Boldface uppercase and lowercase letters represent matrices and vectors, respectively. For a matrix $\mathbf{A}$, $\mathbf{A}^{-1}$ and $\mathbf{A}^T$ denote its inverse (if it exists) and transpose,

respectively. The symbol $\mathbf{I}_N$ denotes the $N \times N$ identity matrix, and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{R})$ stands for the probability density function (PDF) of a Gaussian random vector $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$. The notation $||\mathbf{x}-\mathbf{y}||^2_{\mathbf{W}}$ stands for $(\mathbf{x}-\mathbf{y})^T \mathbf{W} (\mathbf{x}-\mathbf{y})$. The symbol $\propto$ represents the linear scalar relationship between two real valued functions. For Hermitian matrices $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{X} \succeq \mathbf{Y}$ ($\mathbf{X} \succ \mathbf{Y}$) means that $\mathbf{X}-\mathbf{Y}$ is positive semidefinite (definite). The sets $[\mathbf{A}, \mathbf{B}]$ are defined by $[\mathbf{A}, \mathbf{B}] = \{\mathbf{X} : \mathbf{B} \succeq \mathbf{X} \succeq \mathbf{A}\}$. The symbol $\texttt{Bdiag}\{\cdot\}$ stands for block diagonal matrix with elements listed inside the bracket; $\otimes$ denotes the Kronecker product; and $\mathbf{X}_{i,j}$ denotes the component of matrix $\mathbf{X}$ on the i-th row and j-th column.

2. Problem Statement and Markov Random Field

Consider a general connected network[1] of M agents, with $\mathcal{V} = \{1, \ldots, M\}$ denoting the set of agents, and $\mathcal{E}_{\text{Net}} \subset \mathcal{V} \times \mathcal{V}$ the set of all undirected communication links in the network, i.e., if i and j can communicate or exchange information directly, $(i,j) \in \mathcal{E}_{\text{Net}}$. At every agent $n \in \mathcal{V}$, the local observations are given by a linear Gaussian model:

$$\mathbf{y}_n = \sum_{i \in n \cup \mathcal{I}(n)} \mathbf{A}_{n,i} \mathbf{x}_i + \mathbf{z}_n, \tag{1}$$

where $\mathcal{I}(n)$ denotes the set of neighbors of agent n (i.e., all agents i with $(n,i) \in \mathcal{E}_{\text{Net}}$), $\mathbf{A}_{n,i}$ is a known coefficient matrix with full column rank, $\mathbf{x}_i$ is the local unknown parameter at agent i with dimension $N_i \times 1$ and with prior distribution $\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i|\mathbf{0}, \mathbf{W}_i)$ ($\mathbf{W}_i \succ \mathbf{0}$), and $\mathbf{z}_n$ is the additive noise with distribution $\mathbf{z}_n \sim \mathcal{N}(\mathbf{z}_n|\mathbf{0}, \mathbf{R}_n)$, where $\mathbf{R}_n \succ 0$. It is assumed that $p(\mathbf{x}_i, \mathbf{x}_j) = p(\mathbf{x}_i)\, p(\mathbf{x}_j)$ and $p(\mathbf{z}_i, \mathbf{z}_j) = p(\mathbf{z}_i)\, p(\mathbf{z}_j)$ for $i \neq j$, and the x_i's and z_j's are independent for all i and j. The goal is to learn $\mathbf{x}_i$, based on $\mathbf{y}_n$, $p(\mathbf{x}_i)$, and $p(\mathbf{z}_n)$.[2]

In centralized estimation, all the observations $\mathbf{y}_n$'s at different agents are forwarded to a central processing unit. Define vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ as the stacking of $\mathbf{x}_n$, $\mathbf{y}_n$, and $\mathbf{z}_n$ in ascending order with respect to n, respectively; then, we obtain

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}, \tag{2}$$

where $\mathbf{A}$ is constructed from $\mathbf{A}_{n,i}$, with specific arrangement dependent on the network topology. Assuming $\mathbf{A}$ is of full column rank, and since (2) is a standard linear model, the optimal minimum mean squared error estimate $\widehat{\mathbf{x}} \triangleq \left[\widehat{\mathbf{x}}_1^T, \ldots, \widehat{\mathbf{x}}_M^T\right]^T$ of $\mathbf{x}$ is given by (Murphy, 2012)

$$\widehat{\mathbf{x}} = \int \mathbf{x} \frac{p(\mathbf{x})\, p(\mathbf{y}|\mathbf{x})}{\int p(\mathbf{x})\, p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}} \mathrm{d}\mathbf{x} = \left(\mathbf{W}^{-1} + \mathbf{A}^T \mathbf{R}^{-1} \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{R}^{-1} \mathbf{y}, \tag{3}$$

where $\mathbf{W}$ and $\mathbf{R}$ are block diagonal matrices containing $\mathbf{W}_i$ and $\mathbf{R}_i$ as their diagonal blocks, respectively. Although well-established, centralized estimation in large-scale networks has

1. A connected network is one where any two distinct agents can communicate with each other through a finite number of hops.
2. By slightly modifying (1), the local model would allow two neighboring agents to share a common observation and the analyses in the following sections still apply. Please refer to Du et al. (2017b) for details, and Du et al. (2017a) for the corresponding models and associated (distributed) convergence conditions.

several drawbacks including: 1) the transmission of $\mathbf{y}_n$, $\mathbf{A}_{n,i}$ and $\mathbf{R}_n$ from peripheral agents to the computation center imposes large communication overhead; 2) knowledge of global network topology is needed in order to construct $\mathbf{A}$; 3) the computation burden at the computation center scales up due to the matrix inversion required in (3) with complexity order $\mathcal{O}\left(\left(\sum_{i=1}^{|\mathcal{V}|} N_i\right)^3\right)$, i.e., cubic in the dimension in general.

On the other hand, Gaussian BP running over graphical models representing the joint posterior distribution of all $\mathbf{x}_i$'s provides a distributed way to learn $\mathbf{x}_i$ locally, thereby mitigating the disadvantages of the centralized approach. In particular, with Gaussian MRF, the joint distribution $p(\mathbf{x})\,p(\mathbf{y}|\mathbf{x})$ is expressed in a pairwise form (Malioutov et al., 2006):

$$p(\mathbf{x})\,p(\mathbf{y}|\mathbf{x}) = \prod_{n\in\mathcal{V}} \psi_n\left(\mathbf{x}_n, \{\mathbf{y}_i\}_{i\in\{n\cup\mathcal{I}(n)\}}\right) \prod_{(n,i)\in\mathcal{E}_{\mathrm{MRF}}} \psi_{n,i}(\mathbf{x}_n, \mathbf{x}_i), \tag{4}$$

where

$$\mathcal{E}_{\mathrm{MRF}} \triangleq \mathcal{E}_{\mathrm{Net}} \cup \{(n,i)\,|\exists k, k\neq n, k\neq i, \text{such that } (n,k)\in\mathcal{E}_{\mathrm{Net}}, \text{and } (i,k)\in\mathcal{E}_{\mathrm{Net}}\}; \tag{5}$$

$$\psi_n\left(\mathbf{x}_n, \{\mathbf{y}_i\}_{i\in n\cup\mathcal{I}(n)}\right) = \exp\left\{\frac{1}{2}\left(\mathbf{x}_n^T\mathbf{W}_n^{-1}\mathbf{x}_n + \sum_{i\in n\cup\mathcal{I}(n)} \mathbf{y}_i^T\mathbf{R}_i^{-1}\mathbf{x}_n\right)\right\} \tag{6}$$

is the potential function at agent n, and

$$\begin{aligned}\psi_{n,i}(\mathbf{x}_n,\mathbf{x}_i) = \exp-\Big\{\frac{1}{2}\big[&(\mathbf{A}_{n,n}\mathbf{x}_n)^T\mathbf{R}_n^{-1}(\mathbf{A}_{n,i}\mathbf{x}_i) + (\mathbf{A}_{i,n}\mathbf{x}_n)^T\mathbf{R}_i^{-1}(\mathbf{A}_{i,i}\mathbf{x}_i)\\ &+ \sum_{\substack{k\in\{\widetilde{k}\,|(\widetilde{k},i)\in\mathcal{E}_{\mathrm{Net}},\\ (\widetilde{k},n)\in\mathcal{E}_{\mathrm{Net}}\}}} (\mathbf{A}_{k,n}\mathbf{x}_n)^T\mathbf{R}_k^{-1}(\mathbf{A}_{k,i}\mathbf{x}_i)\big]\Big\}\end{aligned} \tag{7}$$

is the edge potential between $\mathbf{x}_n$ and $\mathbf{x}_i$. After setting up the graphical model representing the joint distribution in (4), messages are exchanged between pairs of agents n and i with $(n,i)\in\mathcal{E}_{\mathrm{MRF}}$. More specifically, according to the standard derivation of Gaussian BP, at the ℓ-th iteration, the message passed from agent n to agent i is

$$w_{n\to i}^{(\ell)}(\mathbf{x}_i) = \int \psi_n\left(\mathbf{x}_n, \{\mathbf{y}_k\}_{k\in n\cup\mathcal{I}(n)}\right)\psi_{n,i}(\mathbf{x}_n,\mathbf{x}_i) \prod_{k\in\mathcal{I}(n)\setminus i} w_{k\to n}^{(\ell-1)}(\mathbf{x}_n)\,\mathrm{d}\mathbf{x}_n. \tag{8}$$

As shown by (8), Gaussian BP is iterative with each agent alternatively receiving messages from its neighbors and forwarding out updated messages. At each iteration, agent i computes its belief on variable $\mathbf{x}_i$ as

$$b_{\mathrm{MRF}}^{(\ell)}(\mathbf{x}_i) \propto \psi_i\left(\mathbf{x}_i, \{\mathbf{y}_n\}_{n\in i\cup\mathcal{I}(i)}\right) \prod_{k\in\mathcal{I}(n)} w_{k\to i}^{(\ell)}(\mathbf{x}_i). \tag{9}$$

It is known that, as the messages (8) converge, the mean of the belief (9) is the exact mean of the marginal distribution of $\mathbf{x}_i$ (Weiss and Freeman, 2001a).

It might seem that our distributed inference problem is now solved, as a solution is readily available. However, there are two serious limitations for the Gaussian MRF approach.

First, messages are passed between pairs of agents in $\mathcal{E}_{\text{MRF}}$, which according to the definition (5) includes not only those direct neighbors, but also pairs that are two hops away but share a common neighbor. This is illustrated in Fig. 1, where Fig. 1(a) shows a network of 4 agents with a line between two neighboring agents indicating the availability of a physical communication link, and Fig. 1(b) shows the equivalent pairwise graph. For this example, in the physical network, there is no direct connection between agents 1 and 4, nor between agents 1 and 3. But in the pairwise representation, those connections are present. We summarize the above observations in the following remark.

Remark 1 *For a network with communication edge set $\mathcal{E}_{Net}$ and local observations following (1), the corresponding MRF graph edge set satisfies $\mathcal{E}_{MRF} \supseteq \mathcal{E}_{Net}$. Thus, Gaussian BP for Gaussian MRFs cannot be applied to the distributed inference problem with the local observation model (1).*[3]

The consequence of the above findings is that, not only does information need to be shared among agents two hops away from each other to construct the edge potential function in (7), but also the messages (8) may be required to be exchanged among non-direct neighbors, where a physical communication link is not available. This complicates significantly the message exchange scheduling.

Secondly, even if the message scheduling between non-neighboring agents can be realized, the convergence of (8) is not guaranteed in loopy networks. For Gaussian MRF with scalar variables, sufficient convergence conditions have been proposed in (Weiss and Freeman, 2001a; Malioutov et al., 2006; Su and Wu, 2015). However, depending on how the factorization of the underlying joint Gaussian distribution is performed, Gaussian BP may exhibit different convergence properties as different factorizations (different Gaussian models) lead to fundamentally different recursive update structures. Furthermore, these results apply only to scalar Gaussian BP, and extension to vector-valued Gaussian BP is nontrivial as we show in this paper.

The next section derives distributed vector inference based on Gaussian BP with high order interactions (beyond pairwise connections), where information sharing and message exchange requirement conform to the physical network topology. Furthermore, convergence conditions will be studied in Section 4, and we show in Section 5 that the convergence condition obtained is strictly weaker than, i.e., subsumes the convergence conditions in (Weiss and Freeman, 2001a; Malioutov et al., 2006; Su and Wu, 2015).

3. In Section 5, we further show that the convergence condition of Gaussian BP obtained in this paper for model (1) encompasses all existing convergence conditions of Gaussian BP for the corresponding Gaussian MRF.

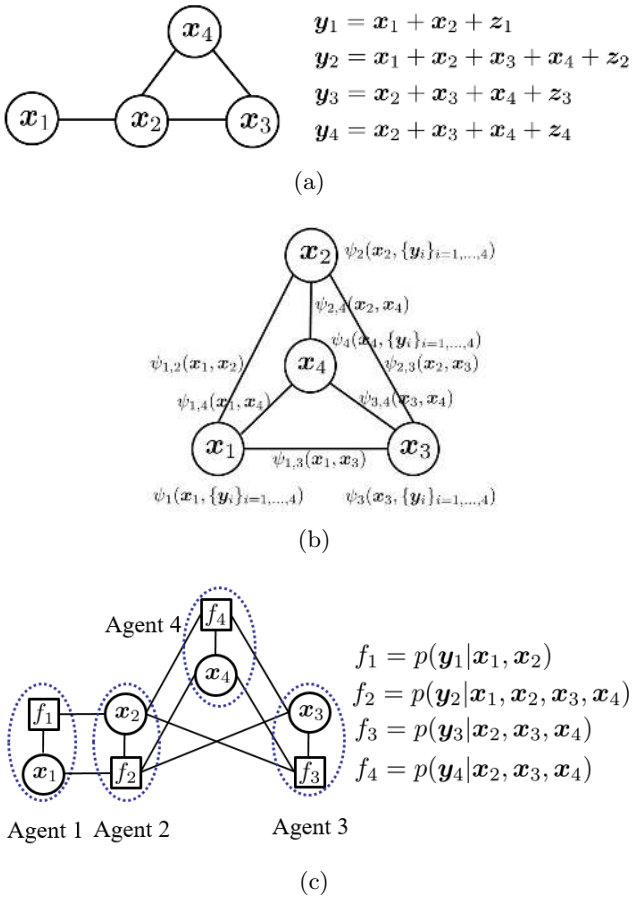


Figure 1: (a) A physical network with 4 agents, where $\{1, 2\}$ and $\{2, 3, 4\}$ are two groups of agents that are within the communication range of each other, respectively. $\mathbf{x}_i$ is the local unknown vector, and $\mathbf{y}_i$ is the local observation at agent i that follows (1); (b) The corresponding MRF of Fig. 1 (a) with $\psi_n\left(\mathbf{x}_n, \{\mathbf{y}_i\}_{i\in n\cup\mathcal{I}(n)}\right)$ and $\psi_{n,i}\left(\mathbf{x}_n, \mathbf{x}_i\right)$ defined in (6) and (7), respectively. (c) The corresponding factor graph of Fig. 1 (a) with f_i defined in (10). Since $p(\mathbf{x}_i)$ does not involve message passing, the $p(\mathbf{x}_i)$ associated to each variable node is not drawn to keep the figure simple.

3. Distributed Inference with Vector-Valued Gaussian BP and Non-Pairwise Interaction

The joint distribution $p(\mathbf{x})\,p(\mathbf{y}|\mathbf{x})$ is first written as the product of the prior distribution and the likelihood function of each local linear Gaussian model in (1) as

$$p(\mathbf{x})\,p(\mathbf{y}|\mathbf{x}) = \prod_{n\in\mathcal{V}} p(\mathbf{x}_n) \prod_{n\in\mathcal{V}} \underbrace{p\left(\mathbf{y}_n|\,\{\mathbf{x}_i\}_{i\in n\cup\mathcal{I}(n)}\right)}_{\triangleq f_n}. \tag{10}$$

To facilitate the derivation of the distributed inference algorithm, the factorization in (10) is expressed in terms of a factor graph (Kschischang et al., 2001), where every vector variable $\mathbf{x}_i$ is represented by a circle (called variable node) and the probability distribution of a vector variable or a group of vector variables is represented by a square (called factor node). A variable node is connected to a factor node if the variable is involved in that particular factor. For example, Fig. 1(c) shows the factor graph representation for the network in Fig. 1(a).

We derive the Gaussian BP algorithm over the corresponding factor graph to learn $\mathbf{x}_n$ for all $n \in \mathcal{V}$ (Kschischang et al., 2001). It involves two types of messages: one is the message from a variable node $\mathbf{x}_j$ to its neighboring factor node f_n, defined as

$$m^{(\ell)}_{j\to f_n}(\mathbf{x}_j) = p(\mathbf{x}_j) \prod_{f_k\in\mathcal{B}(j)\backslash f_n} m^{(\ell-1)}_{f_k\to j}(\mathbf{x}_j), \tag{11}$$

where $\mathcal{B}(j)$ denotes the set of neighbouring factor nodes of $\mathbf{x}_j$, and $m^{(\ell-1)}_{f_k\to j}(\mathbf{x}_j)$ is the message from f_k to $\mathbf{x}_j$ at time $\ell-1$. The second type of message is from a factor node f_n to a neighboring variable node $\mathbf{x}_i$, defined as

$$m^{(\ell)}_{f_n\to i}(\mathbf{x}_i) = \int\cdots\int f_n \times \prod_{j\in\mathcal{B}(f_n)\backslash i} m^{(\ell)}_{j\to f_n}(\mathbf{x}_j)\ \mathrm{d}\,\{\mathbf{x}_j\}_{j\in\mathcal{B}(f_n)\backslash i}, \tag{12}$$

where $\mathcal{B}(f_n)$ denotes the set of neighboring variable nodes of f_n. The process iterates between equations (11) and (12). At each iteration ℓ, the approximate marginal distribution, also referred to as belief, on $\mathbf{x}_i$ is computed locally at $\mathbf{x}_i$ as

$$b^{(\ell)}_{\mathrm{BP}}(\mathbf{x}_i) = p(\mathbf{x}_i) \prod_{f_n\in\mathcal{B}(i)} m^{(\ell)}_{f_n\to i}(\mathbf{x}_i). \tag{13}$$

In the sequel, we derive the exact expressions for the messages $m^{(\ell)}_{j\to f_n}(\mathbf{x}_j)$, $m^{(\ell)}_{f_n\to i}(\mathbf{x}_i)$, and belief $b^{(\ell)}_{\mathrm{BP}}(\mathbf{x}_i)$. First, let the initial messages at each variable node and factor node be in Gaussian function forms as

$$m^{(0)}_{f_n\to i}(\mathbf{x}_i) \propto \exp\left\{-\frac{1}{2}||\mathbf{x}_i - \mathbf{v}^{(0)}_{f_n\to i}||^2_{\mathbf{J}^{(0)}_{f_n\to i}}\right\}. \tag{14}$$

In Appendix A, it is shown that the general expression for the message from variable node j to factor node f_n is

$$m^{(\ell)}_{j\to f_n}(\mathbf{x}_j) \propto \exp\left\{-\frac{1}{2}||\mathbf{x}_j - \mathbf{v}^{(\ell)}_{j\to f_n}||^2_{\mathbf{J}^{(\ell)}_{j\to f_n}}\right\}, \tag{15}$$

with

$$\mathbf{J}_{j\to f_n}^{(\ell)} = \mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\setminus f_n} \mathbf{J}_{f_k\to j}^{(\ell-1)}, \tag{16}$$

$$\mathbf{v}_{j\to f_n}^{(\ell)} = \left[\mathbf{J}_{j\to f_n}^{(\ell)}\right]^{-1}\left[\sum_{f_k\in\mathcal{B}(j)\setminus f_n} \mathbf{J}_{f_k\to j}^{(\ell-1)}\mathbf{v}_{f_k\to j}^{(\ell-1)}\right], \tag{17}$$

where $\mathbf{J}_{f_k\to j}^{(\ell-1)}$ and $\mathbf{v}_{f_k\to j}^{(\ell-1)}$ are the message information matrix (inverse of covariance matrix) and mean vector received at variable node j at the $(\ell-1)$-th iteration, respectively. Furthermore, the message from factor node f_n to variable node i is given by

$$m_{f_n\to i}^{(\ell)}(\mathbf{x}_i) \propto \alpha_{f_n\to i}^{(\ell)} \exp\left\{-\frac{1}{2}||\mathbf{x}_i - \mathbf{v}_{f_n\to i}^{(\ell)}||_{\mathbf{J}_{f_n\to i}^{(\ell)}}^2\right\}, \tag{18}$$

with

$$\mathbf{J}_{f_n\to i}^{(\ell)} = \mathbf{A}_{n,i}^T\left[\mathbf{R}_n + \sum_{j\in\mathcal{B}(f_n)\setminus i}\mathbf{A}_{n,j}\left[\mathbf{J}_{j\to f_n}^{(\ell)}\right]^{-1}\mathbf{A}_{n,j}^T\right]^{-1}\mathbf{A}_{n,i}, \tag{19}$$

$$\mathbf{v}_{f_n\to i}^{(\ell)} = \left[\mathbf{J}_{f_n\to i}^{(\ell)}\right]^{-1}\mathbf{A}_{n,i}^T\left[\mathbf{R}_n + \sum_{j\in\mathcal{B}(f_n)\setminus i}\mathbf{A}_{n,j}\left[\mathbf{J}_{j\to f_n}^{(\ell)}\right]^{-1}\mathbf{A}_{n,j}^T\right]^{-1}\left(\mathbf{y}_n - \sum_{j\in\mathcal{B}(f_n)\setminus i}\mathbf{A}_{n,j}\mathbf{v}_{j\to f_n}^{(\ell)}\right), \tag{20}$$

and

$$\alpha_{f_n\to i}^{(\ell)} \propto \int\cdots\int \exp\left\{-\frac{1}{2}\mathbf{z}^T\mathbf{\Lambda}_{f_n\to i}^{(\ell)}\mathbf{z}\right\}\mathrm{d}\mathbf{z}. \tag{21}$$

In (21), $\mathbf{\Lambda}_{f_n\to i}^{(\ell)}$ is a diagonal matrix containing the eigenvalues of $\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}}^T\mathbf{R}_n^{-1}\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}} + \mathbf{J}_{\{\mathcal{B}(f_n)\setminus i\}\to f_n}^{(\ell)}$, with $\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}}$ denoting a row block matrix containing $\mathbf{A}_{n,j}$ as row elements for all $j\in\mathcal{B}(f_n)\setminus i$ arranged in ascending order, and $\mathbf{J}_{\{\mathcal{B}(f_n)\setminus i\}\to f_n}^{(\ell)}$ denoting a block diagonal matrix with $\mathbf{J}_{j\to f_n}^{(\ell)}$ as its block diagonal elements for all $j\in\mathcal{B}(f_n)\setminus i$ arranged in ascending order.

Obviously, the validity of (18) depends on the existence of $\alpha_{f_n\to i}^{(\ell)}$. It is evident that (21) is the integral of a Gaussian distribution and equals to a constant when $\mathbf{\Lambda}_{f_n\to i}^{(\ell)} \succ \mathbf{0}$ or equivalently $\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}}^T\mathbf{R}_n^{-1}\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}} + \mathbf{J}_{\{\mathcal{B}(f_n)\setminus i\}\to f_n}^{(\ell)} \succ \mathbf{0}$. Otherwise, $\alpha_{f_n\to i}^{(\ell)}$ does not exist. Therefore, the necessary and sufficient condition for the existence of $m_{f_n\to i}^{(\ell)}(\mathbf{x}_i)$ is

$$\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}}^T\mathbf{R}_n^{-1}\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}} + \mathbf{J}_{\{\mathcal{B}(f_n)\setminus i\}\to f_n}^{(\ell)} \succ \mathbf{0}. \tag{22}$$

In general, the necessary and sufficient condition is difficult to be verified, as $\mathbf{J}_{\{\mathcal{B}(f_n)\setminus i\}\to f_n}^{(\ell)}$ changes in each iteration. However, as $\mathbf{R}_n^{-1} \succ \mathbf{0}$, it can be decomposed as $\mathbf{R}_n^{-1} = \widetilde{\mathbf{R}}_n^T\widetilde{\mathbf{R}}_n$. Then

$$\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}}^T\mathbf{R}_n^{-1}\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}} = \left(\widetilde{\mathbf{R}}_n\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}}\right)^T\left(\widetilde{\mathbf{R}}_n\mathbf{A}_{n,\{\mathcal{B}(f_n)\setminus i\}}\right) \succeq \mathbf{0}.$$

Hence, one simple sufficient condition to guarantee (22) is $\mathbf{J}^{(\ell)}_{\{\mathcal{B}(f_n)\setminus i\}\to f_n} \succ \mathbf{0}$ or equivalently its diagonal block matrix $\mathbf{J}^{(\ell)}_{j\to f_n} \succ \mathbf{0}$ for all $j \in \mathcal{B}(f_n) \setminus i$. The following lemma shows that setting the initial message covariances $\mathbf{J}^{(0)}_{f_n\to i} \succeq \mathbf{0}$ for all $(n,i) \in \mathcal{E}_{\text{Net}}$ guarantees $\mathbf{J}^{(\ell)}_{j\to f_n} \succ \mathbf{0}$ for $\ell \geq 1$ and all $(n,j) \in \mathcal{E}_{\text{Net}}$.

Lemma 2 *Let the initial messages at factor node f_k be in Gaussian forms with the initial message information matrix $\boldsymbol{J}^{(0)}_{f_k\to j} \succeq \boldsymbol{0}$ for all $k \in \mathcal{V}$ and $j \in \mathcal{B}(f_k)$. Then $\boldsymbol{J}^{(\ell)}_{j\to f_n} \succ \boldsymbol{0}$ and $\boldsymbol{J}^{(\ell)}_{f_k\to j} \succ \boldsymbol{0}$ for all $\ell \geq 1$ with $j \in \mathcal{V}$ and $f_n, f_k \in \mathcal{B}(j)$. Furthermore, in this case, all messages $m^{(\ell)}_{j\to f_n}(\boldsymbol{x}_j)$ and $m^{(\ell)}_{f_k\to j}(\boldsymbol{x}_i)$ are well defined.*

Proof See Appendix B. ■

For this factor graph based approach, according to the message updating procedure (15) and (18), message exchange is only needed between neighboring agents (an agent refers to a variable-factor pair as shown in Fig. 1 (c)). For example, the messages transmitted from agent n to its neighboring agent i are $m^{(\ell)}_{f_n\to i}(\mathbf{x}_i)$ and $m^{(\ell)}_{n\to f_i}(\mathbf{x}_n)$. Thus, the factor graph does impose a clear messaging schedule, and the message passing scheme given in (11) and (12) conforms with the network topology. Furthermore, if the messages $m^{(\ell)}_{j\to f_n}(\mathbf{x}_j)$ and $m^{(\ell)}_{f_n\to i}(\mathbf{x}_i)$ exist for all ℓ (which can be achieved using Lemma 2), the messages are Gaussian, therefore only the corresponding mean vectors and information matrices (inverse of covariance matrices) are needed to be exchanged.

Finally, if the Gaussian BP messages exist, according to the definition of belief in (13), $b^{(\ell)}_{\text{BP}}(\mathbf{x}_i)$ at iteration ℓ is computed as

$$\begin{aligned}\mathbf{b}^{(\ell)}_{\text{BP}}(\mathbf{x}_i) &= p(\mathbf{x}_i) \prod_{f_n\in\mathcal{B}(i)} m^{(\ell)}_{f_n\to i}(\mathbf{x}_i),\\ &\propto \mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}^{(\ell)}_i, \mathbf{P}^{(\ell)}_i\right),\end{aligned}$$

where the belief covariance matrix

$$\mathbf{P}^{(\ell)}_i = \left[\mathbf{W}_i^{-1} + \sum_{f_n\in\mathcal{B}(i)} \mathbf{J}^{(\ell)}_{f_n\to i}\right]^{-1}, \tag{23}$$

and mean vector

$$\boldsymbol{\mu}^{(\ell)}_i = \mathbf{P}^{(\ell)}_i \left[\sum_{f_n\in\mathcal{B}(i)} \mathbf{J}^{(\ell)}_{f_n\to i}\mathbf{v}^{(\ell)}_{f_n\to i}\right]. \tag{24}$$

The iterative algorithm based on Gaussian BP is summarized as follows. The algorithm is started by setting the messages from factor nodes to variable nodes as in (14). At each round of message exchange, every variable node computes the output messages to its neighboring factor nodes according to (16) and (17). After receiving the messages from its neighboring variable nodes, each factor node computes its output messages according to (19) and (20). The iterative computation terminates when the iterates in (15) or (18) tend to approach a fixed value or the maximum number of iterations is reached.

Remark 3 *We assume that $\boldsymbol{R}_n \succ \boldsymbol{0}$ in this paper. If, however, some of the observations are noiseless, for example, $\boldsymbol{R}_n = \boldsymbol{0}$, the local observation is $\boldsymbol{y}_n = \sum_{i\in n\cup\mathcal{I}(n)} \boldsymbol{A}_{n,i}\boldsymbol{x}_i$. Then the corresponding local likelihood function is represented by the Dirac measure $\delta(\boldsymbol{y}_n - \sum_{i\in n\cup\mathcal{I}(n)} \boldsymbol{A}_{n,i}\boldsymbol{x}_i)$. Suppose, for example, there is only one agent with $\boldsymbol{R}_n = \boldsymbol{0}$, and all others are $\boldsymbol{R}_i \succ \boldsymbol{0}$. The the joint distribution is written as*

$$p(\boldsymbol{x})\, p(\boldsymbol{y}|\boldsymbol{x}) = \delta(\boldsymbol{y}_n - \sum_{i\in n\cup\mathcal{I}(n)} \boldsymbol{A}_{n,i}\boldsymbol{x}_i) \prod_{j\in\mathcal{V}} p(\boldsymbol{x}_j) \prod_{k\in\mathcal{V}} p\left(\boldsymbol{y}_k | \{\boldsymbol{x}_i\}_{i\in k\cup\mathcal{I}(k)}\right).$$

In this case, if $\boldsymbol{A}_{n,n}$ is invertible, then, by the definition of the Dirac measure, we have $\boldsymbol{x}_n = \boldsymbol{A}_{n,n}^{-1}\left(\boldsymbol{y}_n - \sum_{i\in\mathcal{I}(n)} \boldsymbol{A}_{n,i}\boldsymbol{x}_i\right)$. By substituting this equation into all of the likelihood functions involving $\boldsymbol{x}_n$, we have the equivalent joint distribution as in (10) with all the likelihood functions having a positive definite noise covariance. We thereafter can apply Gaussian BP to this new factorization and the convergence analysis in this paper still applies. Therefore, without loss of generality, we assume all $\boldsymbol{R}_n \succ \boldsymbol{0}$. Note that when $\boldsymbol{R}_n = \boldsymbol{0}$ for all n, this problem is equivalent to solving algebraic equations, which has been studied in (Shental et al., 2008b) using Gaussian BP.

4. Convergence Analysis

The challenge of deploying the Gaussian BP algorithm for large-scale networks is in determining whether it will converge or not. In particular, it is generally known that if the factor graph contains cycles, the Gaussian BP algorithm may diverge. Thus, determining convergence conditions for the Gaussian BP algorithm is very important. Sufficient conditions for the convergence of Gaussian BP with scalar variables in loopy graphs are available in (Weiss and Freeman, 2001a; Malioutov et al., 2006; Su and Wu, 2015). However, these conditions are derived based on pairwise graphs with local functions in the form of (6) and (7). This contrasts with the model considered in this paper, where the f_n in (10) involves high-order interactions between vector variables, and thus the convergence results in (Weiss and Freeman, 2001a; Malioutov et al., 2006; Su and Wu, 2015) cannot be applied to the factor graph based vector-form Gaussian BP.

Due to the recursive updating property of $m_{j\to f_n}^{(\ell)}(\mathbf{x}_j)$ and $m_{f_n\to i}^{(\ell)}(\mathbf{x}_i)$ in (15) and (18), the message evolution can be simplified by combining these two kinds of messages into one. By substituting $\mathbf{J}_{j\to f_n}^{(\ell)}$ in (16) into (19), the updating of the message covariance matrix inverse, referred to as message information matrix in the following, can be denoted as

$$\begin{aligned}\mathbf{J}_{f_n\to i}^{(\ell)} &= \mathbf{A}_{n,i}^T\left[\mathbf{R}_n + \sum_{j\in\mathcal{B}(f_n)\setminus i} \mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\setminus f_n} \mathbf{J}_{f_k\to j}^{(\ell-1)}\right]^{-1}\mathbf{A}_{n,j}^T\right]^{-1}\mathbf{A}_{n,i} \\ &\triangleq \mathcal{F}_{n\to i}\left(\left\{\mathbf{J}_{f_k\to j}^{(\ell-1)}\right\}_{(f_k,j)\in\widetilde{\mathcal{B}}(f_n,i)}\right), \end{aligned} \tag{25}$$

where $\widetilde{\mathcal{B}}(f_n,i) = \{(f_k,j)\,|j\in\mathcal{B}(f_n)\setminus i, f_k\in\mathcal{B}(j)\setminus f_n\}$. Observing that $\mathbf{J}_{f_n\to i}^{(\ell)}$ in (25) is independent of $\mathbf{v}_{j\to f_n}^{(\ell)}$ and $\mathbf{v}_{f_n\to i}^{(\ell)}$ in (17) and (18), so we can first focus on the convergence property of $\mathbf{J}_{f_n\to i}^{(\ell)}$ alone and then later on that of $\mathbf{v}_{f_n\to i}^{(\ell)}$. With the convergence characteri-

zation of $\mathbf{J}_{f_n\to i}^{(\ell)}$ and $\mathbf{v}_{f_n\to i}^{(\ell)}$, we will further investigate the convergence of belief covariances and means in (23) and (24), respectively.

Note that computing $\mathbf{P}_j^{(\ell)}$ requires all the incoming messages from neighboring nodes including $\mathbf{J}_{f_n\to j}^{(\ell)}$ as shown in (23) by replacing the subscript i with j in (23). However, according to (25), when computing $\mathbf{J}_{f_n\to i}^{(\ell)}$ the quantity $\mathbf{J}_{f_n\to j}^{(\ell-1)}$ is excluded, i.e., the quantity inside the inner square brackets equals $[\mathbf{P}_j^{(\ell-1)}]^{-1} - \mathbf{J}_{f_n\to j}^{(\ell-1)}$. Therefore, one cannot compute $\mathbf{J}_{f_n\to i}^{(\ell)}$ from $\mathbf{P}_j^{(\ell)}$ alone.

4.1 Convergence of Message Information Matrices

To efficiently represent the updates of all message information matrices, we introduce the following definitions. Let

$$\mathbf{J}^{(\ell-1)} \triangleq \mathtt{Bdiag}\left(\left\{\mathbf{J}_{f_n\to i}^{(\ell-1)}\right\}_{n\in\mathcal{V},i\in\mathcal{B}(f_n)}\right)$$

be a block diagonal matrix with diagonal blocks being the message information matrices in the network at time $\ell-1$ with index arranged in ascending order first on n and then on i. Using the definition of $\mathbf{J}^{(\ell-1)}$, the term $\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}_{f_k\to j}^{(\ell-1)}$ in (25) can be written as $\mathbf{\Xi}_{n,j}\mathbf{J}^{(\ell-1)}\mathbf{\Xi}_{n,j}^T$, where $\mathbf{\Xi}_{n,j}$ is for selecting appropriate components from $\mathbf{J}^{(\ell-1)}$ to form the summation. Further, define $\mathbf{H}_{n,i} = \left[\{\mathbf{A}_{n,j}\}_{j\in B(f_n)\backslash i}\right]$, $\mathbf{\Psi}_{n,i} = \mathtt{Bdiag}\left(\left\{\mathbf{W}_j^{-1}\right\}_{j\in B(f_n)\backslash i}\right)$ and $\mathbf{K}_{n,i} = \mathtt{Bdiag}\left(\{\mathbf{\Xi}_{n,j}\}_{j\in B(f_n)\backslash i}\right)$, all with component blocks arranged with ascending order on j. Then (25) can be written as

$$\mathbf{J}_{f_n\to i}^{(\ell)} = \mathbf{A}_{n,i}^T\left\{\mathbf{R}_n + \mathbf{H}_{n,i}\left[\mathbf{\Psi}_{n,i} + \mathbf{K}_{n,i}\left(\mathbf{I}_{|\mathcal{B}(f_n)|-1}\otimes\mathbf{J}^{(\ell-1)}\right)\mathbf{K}_{n,i}^T\right]^{-1}\mathbf{H}_{n,i}^T\right\}^{-1}\mathbf{A}_{n,i}. \quad (26)$$

Now, we define the function $\mathcal{F} \triangleq \{\mathcal{F}_{1\to k},\ldots,\mathcal{F}_{n\to i},\ldots,\mathcal{F}_{n\to M}\}$ that satisfies $\mathbf{J}^{(\ell)} = \mathcal{F}\left(\mathbf{J}^{(\ell-1)}\right)$. Then, by stacking $\mathbf{J}_{f_n\to i}^{(\ell)}$ on the left side of (26) for all n and i as the block diagonal matrix $\mathbf{J}^{(\ell)}$, we obtain

$$\begin{aligned}\mathbf{J}^{(\ell)} &= \mathbf{A}^T\{\mathbf{\Omega} + \mathbf{H}\left[\mathbf{\Psi} + \mathbf{K}\left(\mathbf{I}_\varphi\otimes\mathbf{J}^{(\ell-1)}\right)\mathbf{K}^T\right]^{-1}\mathbf{H}^T\}^{-1}\mathbf{A},\\ &\triangleq \mathcal{F}\left(\mathbf{J}^{(\ell-1)}\right), \end{aligned} \quad (27)$$

where $\mathbf{A}$, $\mathbf{H}$, $\mathbf{\Psi}$, and $\mathbf{K}$ are block diagonal matrices with block elements $\mathbf{A}_{n,i}$, $\mathbf{H}_{n,i}$, $\mathbf{\Psi}_{n,i}$, and $\mathbf{K}_{n,i}$, respectively, arranged in ascending order, first on n and then on i (i.e., the same order as $\mathbf{J}_{f_n\to i}^{(\ell)}$ in $\mathbf{J}^{(\ell)}$). Furthermore, $\varphi = \sum_{n=1}^M |\mathcal{B}(f_n)|\,(|\mathcal{B}(f_n)|-1)$ and $\mathbf{\Omega}$ is a block diagonal matrix with diagonal blocks $\mathbf{I}_{|B(f_n)|}\otimes\mathbf{R}_n$ with ascending order on n. We first present some properties of the updating operator $\mathcal{F}(\cdot)$, the proofs being provided in Appendix C.

Proposition 4 *The updating operator $\mathcal{F}(\cdot)$ satisfies the following properties:*

P 4.1: $\mathcal{F}\left(\mathbf{J}^{(\ell)}\right) \succeq \mathcal{F}\left(\mathbf{J}^{(\ell-1)}\right)$, if $\mathbf{J}^{(\ell)} \succeq \mathbf{J}^{(\ell-1)} \succeq \mathbf{0}$.

P 4.2: $\alpha\mathcal{F}\left(\mathbf{J}^{(\ell)}\right) \succ \mathcal{F}\left(\alpha\mathbf{J}^{(\ell)}\right)$ and $\mathcal{F}\left(\alpha^{-1}\mathbf{J}^{(\ell)}\right) \succ \alpha^{-1}\mathcal{F}\left(\mathbf{J}^{(\ell)}\right)$, if $\mathbf{J}^{(\ell)} \succ \mathbf{0}$ and $\alpha > 1$.

P 4.3: Define $\mathbf{U} \triangleq \mathbf{A}^T\boldsymbol{\Omega}^{-1}\mathbf{A}$ and $\mathbf{L} \triangleq \mathbf{A}^T\left[\boldsymbol{\Omega} + \mathbf{H}\boldsymbol{\Psi}^{-1}\mathbf{H}^T\right]^{-1}\mathbf{A}$. With arbitrary $\mathbf{J}^{(0)} \succeq \mathbf{0}$, $\mathcal{F}\left(\mathbf{J}^{(\ell)}\right)$ is bounded by $\mathbf{U} \succeq \mathcal{F}\left(\mathbf{J}^{(\ell)}\right) \succeq \mathbf{L} \succ \mathbf{0}$ for $\ell \geq 1$.

Based on the above properties of $\mathcal{F}(\cdot)$, we can establish the convergence of the information matrices.

Theorem 5 *There exists a unique positive definite fixed point $\boldsymbol{J}^*$ for the mapping $\mathcal{F}(\cdot)$.*

Proof The set $[\mathbf{L}, \mathbf{U}]$ is a compact set. Further, according to Proposition 4, P 4.3, for arbitrary $\mathbf{J}^{(0)} \succeq \mathbf{0}$, $\mathcal{F}$ maps $[\mathbf{L}, \mathbf{U}]$ into itself starting from $\ell \geq 1$. Next, we show that $[\mathbf{L}, \mathbf{U}]$ is a convex set. Suppose that $\mathbf{X}$, $\mathbf{Y} \in [\mathbf{L}, \mathbf{U}]$, and $0 \leq t \leq 1$, then $t\mathbf{X} - t\mathbf{L}$ and $(1-t)\mathbf{Y} - (1-t)\mathbf{L}$ are positive semidefinite (p.s.d.) matrices. Since the sum of two p.s.d. matrices is a p.s.d. matrix, $t\mathbf{X} + (1-t)\mathbf{Y} \succeq \mathbf{L}$. Likewise, it can be shown that $t\mathbf{X} + (1-t)\mathbf{Y} \preceq \mathbf{U}$. Thus, the continuous function $\mathcal{F}$ maps a compact convex subset of the Banach space of positive definite matrices into itself. Therefore, the mapping $\mathcal{F}$ has a fixed point in $[\mathbf{L}, \mathbf{U}]$ according to Brouwer's Fixed-Point Theorem (Zeidler, 1985), and the fixed point is positive definite (p.d.).

Next, we prove the uniqueness of the fixed point. Suppose that there exist two fixed points $\mathbf{J}^* \succ \mathbf{0}$ and $\widetilde{\mathbf{J}}^* \succ \mathbf{0}$. Since $\mathbf{J}^*$ and $\widetilde{\mathbf{J}}^*$ are p.d., their components $\mathbf{J}^*_{f_n \to i}$ and $\widetilde{\mathbf{J}}^*_{f_n \to i}$ are also p.d. matrices. For the component blocks of $\mathbf{J}^*$ and $\widetilde{\mathbf{J}}^*$, there are two possibilities: 1) $\widetilde{\mathbf{J}}^*_{f_n \to i} - \mathbf{J}^*_{f_n \to i} \succ \mathbf{0}$ or $\widetilde{\mathbf{J}}^*_{f_n \to i} - \mathbf{J}^*_{f_n \to i}$ is indefinite for some $n, i \in \mathcal{V}$, and 2) $\widetilde{\mathbf{J}}^*_{f_n \to i} - \mathbf{J}^*_{f_n \to i} \preceq 0$ for all $n, i \in \mathcal{V}$.

For the first case, there must exist $\xi_{f_n,i} > 1$ such that $\xi_{f_n,i}\mathbf{J}^*_{f_n \to i} - \widetilde{\mathbf{J}}^*_{f_n \to i}$ has one or more zero eigenvalues, while all other eigenvalues are positive. Pick the component matrix with the maximum $\xi_{f_n,i}$ among those falling into this case, say $\xi_{f_\varrho,\tau}$, then, we can write

$$\xi_{f_\varrho,\tau}\mathbf{J}^*_{f_\varrho \to \tau} - \widetilde{\mathbf{J}}^*_{f_\varrho \to \tau} \succeq \mathbf{0}, \tag{28}$$

or in terms of the information matrices for the whole network

$$\xi_{f_\varrho,\tau}\mathbf{J}^* \succeq \widetilde{\mathbf{J}}^* \succ \mathbf{0}, \quad \xi_{f_\varrho,\tau} > 1. \tag{29}$$

Applying $\mathcal{F}$ on both sides of (29), according to the monotonic property of $\mathcal{F}(\cdot)$ as shown in Proposition 4, P 4.1, we have

$$\mathcal{F}\left(\xi_{f_\varrho,\tau}\mathbf{J}^*\right) \succeq \mathcal{F}\left(\widetilde{\mathbf{J}}^*\right) = \widetilde{\mathbf{J}}^*, \tag{30}$$

where the equality is due to $\widetilde{\mathbf{J}}^*$ being a fixed point. According to Proposition 4, P 4.2, $\xi_{f_\varrho,\tau}\mathcal{F}(\mathbf{J}^*) \succ \mathcal{F}\left(\xi_{f_\varrho,\tau}\mathbf{J}^*\right)$. Therefore, from (30), we obtain $\xi_{f_\varrho,\tau}\mathbf{J}^* \succ \widetilde{\mathbf{J}}^*$. Consequently,

$$\xi_{f_\varrho,\tau}\mathbf{J}^*_{f_\varrho,\tau} \succ \widetilde{\mathbf{J}}^*_{f_\varrho,\tau}.$$

But this contradicts with $\xi_{f_\varrho,\tau}\mathbf{J}^*_{f_\varrho,\tau} - \widetilde{\mathbf{J}}^*_{f_\varrho,\tau}$ having one or more zero eigenvalues as discussed before (28). Therefore, we must have $\mathbf{J}^* = \widetilde{\mathbf{J}}^*$.

On the other hand, if we have case two, which is $\widetilde{\mathbf{J}}^*_{f_n\to i} - \mathbf{J}^*_{f_n\to i} \preceq 0$ for all $n, i \in \mathcal{V}$, we can repeat the above derivation with the roles of $\widetilde{\mathbf{J}}^*$ and $\mathbf{J}^*$ reversed, and we would again obtain $\mathbf{J}^* = \widetilde{\mathbf{J}}^*$. Consequently, $\mathbf{J}^*$ is unique. ■

Lemma 2 states that with arbitrary p.s.d. initial message information matrices, the message information matrices will be kept as p.d. at every iteration. On the other hand, Theorem 5 indicates that there exists a unique fixed point for the mapping $\mathcal{F}$. Next, we will show that, with arbitrary initial value $\mathbf{J}^{(0)} \succeq 0$, $\mathbf{J}^{(\ell)}$ converges to a unique p.d. matrix.

Theorem 6 *The matrix sequence* $\left\{\boldsymbol{J}^{(\ell)}\right\}_{l=0,1,\ldots}$ *defined by (27) converges to a unique positive definite matrix* $\boldsymbol{J}^*$ *for any initial covariance matrix* $\boldsymbol{J}^{(0)} \succeq \mathbf{0}$.

Proof With arbitrary initial value $\mathbf{J}^{(0)} \succeq \mathbf{0}$, following Proposition 4, P 4.3, we have $\mathbf{U} \succeq \mathbf{J}^{(1)} \succeq \mathbf{L} \succ \mathbf{0}$. On the other hand, according to Theorem 5, (27) has a unique fixed point $\mathbf{J}^* \succ \mathbf{0}$. Notice that we can always choose a scalar $\alpha > 1$ such that

$$\alpha\mathbf{J}^* \succeq \mathbf{J}^{(1)} \succeq \mathbf{L}. \tag{31}$$

Applying $\mathcal{F}(\cdot)$ to (31) ℓ times, and using Proposition 4, P 4.1, we have

$$\mathcal{F}^\ell(\alpha\mathbf{J}^*) \succeq \mathcal{F}^{\ell+1}\left(\mathbf{J}^{(0)}\right) \succeq \mathcal{F}^\ell(\mathbf{L}), \tag{32}$$

where $\mathcal{F}^\ell(\mathbf{X})$ denotes applying $\mathcal{F}$ on $\mathbf{X}$ ℓ times.

We start from the left inequality in (32). According to Proposition 4, P 4.2, $\alpha\mathbf{J}^* \succ \mathcal{F}(\alpha\mathbf{J}^*)$. Applying $\mathcal{F}$ again gives $\mathcal{F}(\alpha\mathbf{J}^*) \succ \mathcal{F}^2(\alpha\mathbf{J}^*)$. Applying $\mathcal{F}(\cdot)$ repeatedly, we can obtain $\mathcal{F}^2(\alpha\mathbf{J}^*) \succ \mathcal{F}^3(\alpha\mathbf{J}^*) \succ \mathcal{F}^4(\alpha\mathbf{J}^*)$, etc. Thus $\mathcal{F}^\ell(\alpha\mathbf{J}^*)$ is a non-increasing sequence with respect to the partial order induced by the cone of p.s.d. matrices as ℓ increases. Furthermore, since $\mathcal{F}(\cdot)$ is bounded below by $\mathbf{L}$, $\mathcal{F}^\ell(\alpha\mathbf{J}^*)$ converges. Finally, since there exists only one fixed point for $\mathcal{F}(\cdot)$, $\lim_{l\to\infty}\mathcal{F}^\ell(\alpha\mathbf{J}^*) = \mathbf{J}^*$. On the other hand, for the right hand side of (32), as $\mathcal{F}(\cdot) \succeq \mathbf{L}$, we have $\mathcal{F}(\mathbf{L}) \succeq \mathbf{L}$. Applying $\mathcal{F}$ repeatedly gives successively $\mathcal{F}^2(\mathbf{L}) \succeq \mathcal{F}(\mathbf{L})$, $\mathcal{F}^3(\mathbf{L}) \succeq \mathcal{F}^2(\mathbf{L})$, etc. So, $\mathcal{F}^\ell(\mathbf{L})$ is an non-decreasing sequence (with respect to the partial order induced by the cone of p.s.d. matrices). Since $\mathcal{F}(\cdot)$ is upper bounded by $\mathbf{U}$, $\mathcal{F}^\ell(\mathbf{L})$ is a convergent sequence. Again, due to the uniqueness of the fixed point, we have $\lim_{l\to\infty}\mathcal{F}^\ell(\mathbf{L}) = \mathbf{J}^*$. Finally, taking the limit with respect to ℓ on (32), we have $\lim_{l\to\infty}\mathcal{F}^\ell\left(\mathbf{J}^{(0)}\right) = \mathbf{J}^*$, for arbitrary initial $\mathbf{J}^{(0)} \succeq \mathbf{0}$. ■

Remark 7 *According to Theorem 6, the information matrix* $\boldsymbol{J}^{(\ell)}_{f_n\to i}$ *converges if all initial information matrices are p.s.d., i.e.,* $\boldsymbol{J}^{(0)}_{f_n\to i} \succeq \boldsymbol{0}$ *for all* $i \in \mathcal{V}$ *and* $f_n \in \mathcal{B}(i)$. *However, for the pairwise model, the messages are derived based on the classical Gaussian MRF based factorization (in the form of equations (6) and (7)) of the joint distribution. This differs from the model considered in this paper, where the factor* f_n *follows equation (10), which leads to intrinsically different recursive equations. More specifically, for BP on the*

Gaussian MRF based factorization, the information matrix does not necessarily converge for all initial nonnegative values (for the scalar variable case) as shown in (Malioutov et al., 2006; Moallemi and Roy, 2009a).

Remark 8 *Due to the computation of $\boldsymbol{J}_{f_n\to i}^{(\ell)}$ being independent of the local observations $\boldsymbol{y}_n$, as long as the network topology does not change, the converged value $\boldsymbol{J}_{f_n\to i}^{*}$ can be precomputed offline and stored at each agent, and there is no need to re-compute $\boldsymbol{J}_{f_n\to i}^{*}$ even if $\boldsymbol{y}_n$ varies.*

Another fundamental question is how fast the convergence is, and this is the focus of the discussion below. Since the convergence of a dynamic system is often studied with respect to the part metric (Chueshov, 2002), in the following, we start by introducing the part metric.

Definition 9 *Part (Birkhoff) Metric (Chueshov, 2002): For arbitrary symmetric matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ with the same dimension, if there exists $\alpha \geq 1$ such that $\alpha\boldsymbol{X} \succeq \boldsymbol{Y} \succeq \alpha^{-1}\boldsymbol{X}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are called the parts, and $\mathrm{d}(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \inf\left\{\log\alpha : \alpha\boldsymbol{X} \succeq \boldsymbol{Y} \succeq \alpha^{-1}\boldsymbol{X}, \alpha \geq 1\right\}$ defines a metric called the part metric.*

As it is useful to have an estimate of the convergence rate of $\mathbf{J}^{(\ell)}$ in terms of the more standard induced matrix norms, we further introduce the notion of monotone norms. The norms $||\cdot||_2$ and $||\cdot||_F$ (Frobenus norm) are monotone norms.

Definition 10 *Monotone Norm (Ciarlet, 1989, 2.2-10): A matrix norm $\|\cdot\|$ is monotone if*

$$\boldsymbol{X} \succeq \boldsymbol{0}, \boldsymbol{Y} \succeq \boldsymbol{X} \Rightarrow \|\boldsymbol{Y}\| \geq \|\boldsymbol{X}\|.$$

Next, for arbitrary $\epsilon > 0$, we will show that $\left\{\mathbf{J}^{(\ell)}\right\}_{l=1,..}$ approaches the ϵ-neighborhood of the fixed point $\mathbf{J}^{\star}$ exponentially fast with respect to the monotone norm. To this end, for a fixed $\epsilon > 0$, define the set

$$\mathcal{C} = \left\{\mathbf{J}^{(\ell)} | \mathbf{U} \succeq \mathbf{J}^{(\ell)} \succeq \mathbf{J}^{*} + \epsilon\mathbf{I}\right\} \cup \left\{\mathbf{J}^{(\ell)} | \mathbf{J}^{*} - \epsilon\mathbf{I} \succeq \mathbf{J}^{(\ell)} \succeq \mathbf{L}\right\}. \tag{33}$$

Theorem 11 *With the initial message information matrix set to be an arbitrary p.s.d. matrix, i.e., $\boldsymbol{J}_{f_n\to i}^{(0)} \succeq \boldsymbol{0}$, the sequence $\left\{\boldsymbol{J}^{(\ell)}\right\}_{l=0,1,...}$ approaches an arbitrarily small neighborhood of the fixed positive definite matrix $\boldsymbol{J}^{*}$ at an exponential rate with respect to any matrix norm.*

Proof Fix $\epsilon > 0$ and consider the set $\mathcal{C}$ defined in (33). It suffices to show that the quantity $\|\mathbf{J}^{(\ell)} - \mathbf{J}^{*}\|$, where $\|\cdot\|$ is a monotone norm as defined in Definition 10, decays exponentially as long as $\mathbf{J}^{(s)} \in \mathcal{C}$ for all $s \in \{0, 1, \cdots \ell\}$. To this end, for $\mathbf{J}^{(\ell)} \in \mathcal{C}$, and $\mathbf{J}^{*} \notin \mathcal{C}$ (necessarily), according to Definition 9, we have $\mathrm{d}\left(\mathbf{J}^{(\ell)}, \mathbf{J}^{*}\right) \triangleq \inf\left\{\log\alpha : \alpha\mathbf{J}^{(\ell)} \succeq \mathbf{J}^{*} \succeq \alpha^{-1}\mathbf{J}^{(\ell)}\right\}$. Since $\mathrm{d}\left(\mathbf{J}^{(\ell)}, \mathbf{J}^{*}\right)$ is the smallest number satisfying $\alpha\mathbf{J}^{(\ell)} \succeq \mathbf{J}^{*} \succeq \alpha^{-1}\mathbf{J}^{(\ell)}$, this is equivalent to

$$\exp\left\{\mathrm{d}\left(\mathbf{J}^{(\ell)}, \mathbf{J}^{*}\right)\right\}\mathbf{J}^{(\ell)} \succeq \mathbf{J}^{*} \succeq \exp\left\{-\mathrm{d}\left(\mathbf{J}^{(\ell)}, \mathbf{J}^{*}\right)\right\}\mathbf{J}^{(\ell)}. \tag{34}$$

Applying Proposition 4, P 4.1 to (34), we have

$$\mathcal{F}\left(\exp\left\{\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}\mathbf{J}^{(\ell)}\right)\succ\mathcal{F}\left(\mathbf{J}^{*}\right)\succ\mathcal{F}\left(\exp\left\{-\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}\mathbf{J}^{(\ell)}\right).$$

Then applying Proposition 4, P 4.2 and considering that $\exp\left\{\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}>1$ and $\exp\left\{-\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}<1$, we obtain

$$\exp\left\{\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}\mathcal{F}\left(\mathbf{J}^{(\ell)}\right)\succeq\mathcal{F}\left(\mathbf{J}^{*}\right)\succeq\exp\left\{-\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}\mathcal{F}\left(\mathbf{J}^{(\ell)}\right).$$

Notice that, for arbitrary p.d. matrices $\mathbf{X}$ and $\mathbf{Y}$, if $\mathbf{X}-k\mathbf{Y}\succ\mathbf{0}$, then, by definition, we have $\mathbf{x}^{T}\mathbf{X}\mathbf{x}-k\mathbf{x}^{T}\mathbf{Y}\mathbf{x}>0$ for arbitrary $\mathbf{x}\neq\mathbf{0}$. Then, there must exist $o>0$ that is small enough such that $\mathbf{x}^{T}\mathbf{X}\mathbf{x}-(k+o)\,\mathbf{x}^{T}\mathbf{Y}\mathbf{x}>0$ or equivalently $\mathbf{X}\succ(k+o)\,\mathbf{Y}$. Thus, as $\exp(\cdot)$ is a continuous function, there must exist some $\triangle\mathrm{d}>0$ such that

$$\exp\left\{-\triangle\mathrm{d}+\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}\mathcal{F}\left(\mathbf{J}^{(\ell)}\right)\succ\mathcal{F}\left(\mathbf{J}^{*}\right)\succ\exp\left\{\triangle\mathrm{d}-\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}\mathcal{F}\left(\mathbf{J}^{(\ell)}\right).\quad(35)$$

Now, using the definition of the part metric, (35) is equivalent to

$$-\triangle\mathrm{d}+\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\geq\mathrm{d}\left(\mathcal{F}\left(\mathbf{J}^{(\ell)}\right),\mathcal{F}\left(\mathbf{J}^{*}\right)\right).$$

Hence, we obtain $\mathrm{d}\left(\mathcal{F}\left(\mathbf{J}^{(\ell)}\right),\mathcal{F}\left(\mathbf{J}^{*}\right)\right)<\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)$. Since this result holds for any $\mathbf{J}^{(\ell)}\in\mathcal{C}$, we also have $\mathrm{d}\left(\mathcal{F}\left(\mathbf{J}^{(\ell)}\right),\mathcal{F}\left(\mathbf{J}^{*}\right)\right)<c\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)$, where $c=\sup_{\mathbf{J}^{(\ell)}\in\mathcal{C}}\frac{\mathrm{d}\left(\mathcal{F}\left(\mathbf{J}^{(\ell)}\right),\mathcal{F}\left(\mathbf{J}^{*}\right)\right)}{\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)}<1$. Since $\mathbf{J}^{(\ell+1)}=\mathcal{F}\left(\mathbf{J}^{(\ell)}\right)$ and $\mathbf{J}^{*}=\mathcal{F}\left(\mathbf{J}^{*}\right)$, we have

$$\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)<c^{\ell}\mathrm{d}\left(\mathbf{J}^{(0)},\mathbf{J}^{*}\right).\quad(36)$$

According to (Krause and Nussbaum, 1993, Lemma 2.3), the convergence rate of $||\mathbf{J}^{(\ell)}-\mathbf{J}^{*}||$ can be determined by that of $\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)$. More specifically,

$$||\mathbf{J}^{(\ell)}-\mathbf{J}^{*}||\leq\left(2\exp\left\{\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}-\exp\left\{-\mathrm{d}\left(\mathbf{J}^{(\ell)},\mathbf{J}^{*}\right)\right\}-1\right)\min\left\{||\mathbf{J}^{(\ell)}||,||\mathbf{J}^{*}||\right\},\quad(37)$$

where $||\cdot||$ is a monotone norm defined on the p.s.d. cone:

As we show in Proposition 4, P 4.3 that $\mathbf{J}^{(\ell)}$ is bounded, then $||\mathbf{J}^{(\ell)}||$ and $||\mathbf{J}^{*}||$ must be finite. Let ζ be the largest value of $\min\left\{||\mathbf{J}^{(\ell)}||,||\mathbf{J}^{*}||\right\}$ for all $\{\mathbf{J}^{(\ell)}\}$ with $\ell\geq0$, then $\zeta>0$. According to (36) and (37), we have that

$$||\mathbf{J}^{(\ell)}-\mathbf{J}^{*}||<\zeta\left(2\exp\left\{c^{\ell}d_{0}\right\}-\exp\left\{-c^{\ell}d_{0}\right\}-1\right),\quad(38)$$

with $0<c<1$ and $d_{0}=\mathrm{d}\left(\mathbf{J}^{(0)},\mathbf{J}^{*}\right)$, which is a constant. The above inequality is equivalent to

$$||\mathbf{J}^{(\ell)}-\mathbf{J}^{*}||<\zeta\left(3\exp\left\{c^{\ell}d_{0}\right\}-\exp\left\{c^{\ell}d_{0}\right\}-\exp\left\{-c^{\ell}d_{0}\right\}-1\right).\quad(39)$$

Since both $\exp\left\{c^\ell d_0\right\}$ and $\exp\left\{-c^\ell d_0\right\}$ are positive and $\exp\left\{c^\ell d_0\right\}\exp\left\{-c^\ell d_0\right\} = 1$, according to the arithmetic-geometric mean inequality, we have $\exp\left\{c^\ell d_0\right\} + \exp\left\{-c^\ell d_0\right\} \geq 2\left(\exp\left\{c^\ell d_0\right\}\exp\left\{-c^\ell d_0\right\}\right)^{1/2} = 2$. Then, the right-hand side of (39) is further amplified, and we obtain

$$||\mathbf{J}^{(\ell)} - \mathbf{J}^*|| < \zeta\left(3\exp\left\{c^\ell d_0\right\} - 3\right) = 3\zeta\left(\exp\left\{c^\ell d_0\right\} - 1\right).$$

Therefore, the sequence $\left\{\mathbf{J}^{(\ell)}\right\}_{l=0,1,\ldots}$ approaches the ϵ-neighborhood (and hence any arbitrarily small neighborhood) of the fixed positive definite matrix $\mathbf{J}^*$ at an exponential rate with respect to any matrix norm. ■

The physical meaning of Theorem 11 is that the distance between $\mathbf{J}^{(\ell)}$ and $\mathbf{J}^*$ decreases exponentially fast before $\mathbf{J}^{(\ell)}$ enters $\mathbf{J}^*$'s neighborhood, which can be chosen to be arbitrarily small. Next, we study how to choose the initial value $\mathbf{J}^{(0)}$ so that $\mathbf{J}^{(\ell)}$ converges faster.

Theorem 12 *With $\boldsymbol{0} \preceq \boldsymbol{J}^{(0)} \preceq \boldsymbol{L}$, $\boldsymbol{J}^{(\ell)}$ is a monotonic increasing sequence, and $\boldsymbol{J}^{(\ell)}$ converges most rapidly with $\boldsymbol{J}^{(0)} = \boldsymbol{L}$. Moreover, with $\boldsymbol{J}^{(0)} \succeq \boldsymbol{U}$, $\boldsymbol{J}^{(\ell)}$ is a monotonic decreasing sequence, and $\boldsymbol{J}^{(\ell)}$ converges most rapidly with $\boldsymbol{J}^{(0)} = \boldsymbol{U}$.*

Proof Following Proposition 4, P 4.3, it can be verified that for $\mathbf{0} \preceq \mathbf{J}^{(0)} \preceq \mathbf{L}$, we have $\mathbf{J}^{(1)} \succeq \mathbf{J}^{(0)}$. Then, according to Proposition 4, P 4.1, and by induction, this relationship can be extended to $\mathbf{J}^{(\ell)} \succeq \ldots \mathbf{J}^{(1)} \succeq \mathbf{J}^{(0)}$, which states that $\mathbf{J}^{(\ell)}$ is a monotonic increasing sequence. Now, suppose that there are two sequences $\mathbf{J}^{(\ell)}$ and $\widetilde{\mathbf{J}}^{(\ell)}$ that are started with different initial values $\mathbf{0} \preceq \mathbf{J}^{(0)} \prec \mathbf{L}$ and $\mathbf{0} \preceq \widetilde{\mathbf{J}}^{(0)} \prec \mathbf{L}$, respectively. Then these two sequences are monotonically increasing and bounded by $\mathbf{J}^*$. To prove that $\mathbf{J}^{(0)} = \mathbf{L}$ leads to the fastest convergence, it is sufficient to prove that $\mathbf{J}^{(\ell)} \succ \widetilde{\mathbf{J}}^{(\ell)}$ for $\ell = 0, 1 \ldots$. First, note that $\mathbf{J}^{(0)} \succ \widetilde{\mathbf{J}}^{(0)}$. Assume $\mathbf{J}^{(n)} \succ \widetilde{\mathbf{J}}^{(n)}$ for some $n \geq 0$. According to Proposition 4, P 4.1, we have $\mathcal{F}\left(\mathbf{J}^{(n)}\right) \succeq \mathcal{F}\left(\widetilde{\mathbf{J}}^{(n)}\right)$, or equivalently $\mathbf{J}^{(n+1)} \succeq \widetilde{\mathbf{J}}^{(n+1)}$. Therefore, by induction, we have proven that, with $\mathbf{J}^{(0)} = \mathbf{L}$, $\mathbf{J}^{(\ell)}$ converges more rapidly than with any other initial value $\mathbf{0} \preceq \mathbf{J}^{(0)} \prec \mathbf{L}$.

With similar logic, we can show that, with $\mathbf{J}^{(0)} \succeq \mathbf{U}$, $\mathbf{J}^{(\ell)}$ is a monotonic decreasing sequence; and, with $\mathbf{J}^{(0)} = \mathbf{U}$, $\mathbf{J}^{(\ell)}$ converges more rapidly than that with any other initial value $\mathbf{J}^{(0)} \succ \mathbf{U}$. ■

Notice that it is a common practice in the Gaussian BP literature that the initial information matrix (or inverse variance for the scalar case) is set to be $\mathbf{0}$, i.e., $\mathbf{J}^{(0)}_{f_n \to i} = \mathbf{0}$ (Weiss and Freeman, 2001a; Malioutov et al., 2006). Theorem 12 reveals that there is a better choice to guarantee faster convergence.

4.2 Convergence of Message Mean Vector

According to Theorems 6 and 11, as long as we choose $\mathbf{J}^{(0)}_{f_k \to j} \succeq \mathbf{0}$ for all $j \in \mathcal{V}$ and $f_k \in \mathcal{B}(j)$, the distance between $\mathbf{J}^{(\ell)}_{f_k \to j}$ and $\mathbf{J}^*_{f_k \to j}$ decreases exponentially fast before $\mathbf{J}^{(\ell)}_{f_k \to j}$ enters $\mathbf{J}^*_{f_k \to j}$'s neighborhood, which can be chosen to be arbitrarily small. Furthermore,

according to (16), $\left[\mathbf{J}_{j\to f_n}^{(\ell)}\right]^{-1}$ also converges to a p.d. matrix once $\mathbf{J}_{f_k\to j}^{(\ell)}$ converges, and the converged value for $\left[\mathbf{J}_{j\to f_n}^{(\ell)}\right]^{-1}$ is denoted by $\left[\mathbf{J}_{j\to f_n}^{*}\right]^{-1}$. Then for arbitrary initial value $\mathbf{v}_{f_k\to j}^{(0)}$, the evolution of $\mathbf{v}_{j\to f_n}^{(\ell)}$ in (17) can be written in terms of the converged message information matrices, which is

$$\mathbf{v}_{j\to f_n}^{(\ell)} = \left[\mathbf{J}_{j\to f_n}^{*}\right]^{-1} \sum_{f_k\in\mathcal{B}(j)\setminus f_n} \mathbf{J}_{f_k\to j}^{*}\mathbf{v}_{f_k\to j}^{(\ell-1)}. \tag{40}$$

Using (20), and replacing indices j, i, n with z, j, k respectively, $\mathbf{v}_{f_k\to j}^{(\ell-1)}$ is given by

$$\mathbf{v}_{f_k\to j}^{(\ell-1)} = [\mathbf{J}_{f_k\to j}^{*}]^{-1}\mathbf{A}_{k,j}^{T}\underbrace{\left[\mathbf{R}_k + \sum_{z\in\mathcal{B}(f_k)\setminus j} \mathbf{A}_{k,z}\left[\mathbf{J}_{z\to f_k}^{*}\right]^{-1}\mathbf{A}_{k,z}^{T}\right]}_{\triangleq\mathbf{M}_{k,j}}{}^{-1}\left(\mathbf{y}_k - \sum_{z\in\mathcal{B}(f_k)\setminus j}\mathbf{A}_{k,z}\mathbf{v}_{z\to f_k}^{(\ell-1)}\right). \tag{41}$$

Putting (41) into (40), we have

$$\mathbf{v}_{j\to f_n}^{(\ell)} = \mathbf{b}_{j\to f_n} - \sum_{f_k\in\mathcal{B}(j)\setminus f_n}\sum_{z\in\mathcal{B}(f_k)\setminus j}[\mathbf{J}_{j\to f_n}^{*}]^{-1}\mathbf{A}_{k,j}^{T}\mathbf{M}_{k,j}^{-1}\mathbf{A}_{k,z}\mathbf{v}_{z\to f_k}^{(\ell-1)}, \tag{42}$$

where $\mathbf{b}_{j\to f_n} = [\mathbf{J}_{j\to f_n}^{*}]^{-1}\sum_{f_k\in\mathcal{B}(j)\setminus f_n}\mathbf{A}_{k,j}^{T}\mathbf{M}_{k,j}^{-1}\mathbf{y}_k$. The above equation can be further written in compact form as

$$\mathbf{v}_{j\to f_n}^{(\ell)} = \mathbf{b}_{j\to f_n} - \mathbf{Q}_{j\to f_n}\mathbf{v}^{(\ell-1)},$$

with the column vector $\mathbf{v}^{(\ell-1)}$ containing $\mathbf{v}_{z\to f_k}^{(\ell-1)}$ for all $z\in\mathcal{V}$ and $f_k\in\mathcal{B}(z)$ as subvector with ascending index first on z and then on k. The matrix $\mathbf{Q}_{j\to f_n}$ is a row block matrix with component block $[\mathbf{J}_{j\to f_n}^{*}]^{-1}\mathbf{A}_{k,j}^{T}\mathbf{M}_{k,j}^{-1}\mathbf{A}_{k,z}$ if $f_k\in\mathcal{B}(j)\setminus f_n$ and $z\in\mathcal{B}(f_k)\setminus j$, and $\mathbf{0}$ otherwise. Let $\mathbf{Q}$ be the block matrix that stacks $\mathbf{Q}_{j\to f_n}$ with the order first on j and then on n, and $\mathbf{b}$ be the vector containing $\mathbf{b}_{j\to f_n}$ with the same stacking order as $\mathbf{Q}_{j\to f_n}$. We have

$$\mathbf{v}^{(\ell)} = -\mathbf{Q}\mathbf{v}^{(\ell-1)} + \mathbf{b}, \quad \ell\geq 1,2,\ldots. \tag{43}$$

It is known that for arbitrary initial value $\mathbf{v}^{(0)}$, $\mathbf{v}^{(\ell)}$ converges if and only if the spectral radius $\rho(\mathbf{Q})<1$ (Demmel, 1997, pp. 280). Since the elements of $\mathbf{v}^{(0)}$, i.e., $\mathbf{v}_{j\to f_n}^{(0)}$, depends on $\mathbf{v}_{f_k\to j}^{(0)}$, we can choose arbitrary $\mathbf{v}_{f_k\to j}^{(0)}$. Furthermore, as $\mathbf{v}^{(\ell)}$ depends on the convergence of $\mathbf{J}^{(\ell)}$, we have the following result.

Theorem 13 *The vector sequence $\left\{\boldsymbol{v}^{(\ell)}\right\}_{l=1,2,\ldots}$ defined by (43) converges to a unique value under any initial value $\left\{\boldsymbol{v}_{f_k\to j}^{(0)}\right\}_{k\in\mathcal{V},j\in\mathcal{B}(f_k)}$ and initial message information matrix $\boldsymbol{J}^{(0)}\succeq\mathbf{0}$ if and only if $\rho(\mathbf{Q})<1$.*

The row block matrix $\mathbf{Q}_j$, a row block of $\mathbf{Q}$, contains only block entries $\mathbf{0}$ and $\mathbf{Q}_{j\to f_n}$. When the observation model (1) reduces to the pairwise model, where only two unknown variables are involved in each local observation, it can be shown that $\mathbf{Q}_j$ and $\mathbf{Q}_i$ are orthogonal if $i \neq j$. A distributed convergence condition is obtained utilizing this orthogonal property in Du et al. (2017a). However, for the more general case studied in this paper, properties of $\mathbf{Q}_j$ and $\mathbf{Q}$ need to be further exploited to show when $\rho(\mathbf{Q}) < 1$ is satisfied.

In the sequel, we will show that $\rho(\mathbf{Q}) < 1$ is satisfied for a single loop factor graph with multiple chains/trees (an example is shown in Fig. 2), thus Gaussian BP converges in such a topology. Although Weiss (2000) shows the convergence of Gaussian BP on the MRF with a single loop, the analysis cannot be applied here since the local observations model (1) is different from the pairwise model in (Weiss, 2000).

Theorem 14 *For any factor graph that is the union of a single loop and a forest, with arbitrary positive semi-definite initial information matrix, i.e., $\boldsymbol{J}^{(0)}_{f_n\to i} \succeq \boldsymbol{0}$ for all $i \in \mathcal{V}$ and $f_n \in \mathcal{B}(i)$, the message information matrix $\boldsymbol{J}^{(\ell)}_{f_n\to i}$ and mean vector $\boldsymbol{v}^{(\ell)}_{i\to f_n}$ is guaranteed to converge to their corresponding unique points.*

Proof In this proof, Fig. 2 is being used as a reference throughout. For a single loop factor graph with chains/trees as shown in Fig. 2 (a), there are two kinds of nodes. One is the factors/variables in the loop, and they are denoted by $f_n/\mathbf{x}_j$. The other is the factors/variables on the chains/trees but outside the loop, denoted as $\widetilde{f}_k/\widetilde{\mathbf{z}}_i$. Then message from a variable node to a neighboring factor node on the graph can be categorized into three groups:

1) message on a tree/chain passing towards the loop, e.g., $m^{*}_{\widetilde{z}\to f_k}(\widetilde{\mathbf{x}}_z)$ and $m^{*}_{\widetilde{s}\to \widetilde{f}_k}(\widetilde{\mathbf{x}}_s)$;

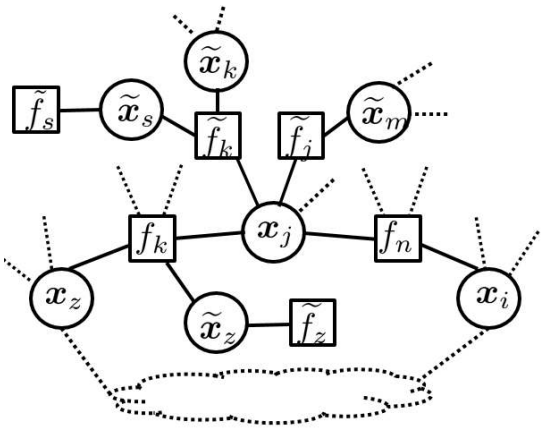
2) message on a tree/chain passing away from the loop, e.g., $m^{(\ell)}_{j\to \widetilde{f}_k}(\mathbf{x}_j)$, $m^{(\ell)}_{\widetilde{s}\to \widetilde{f}_s}(\mathbf{x}_s)$ and $m^{(\ell)}_{\widetilde{z}\to \widetilde{f}_z}(\mathbf{x}_z)$;

3) message in the loop, e.g., $m^{(\ell)}_{j\to f_n}(\mathbf{x}_j)$, $m^{(\ell)}_{z\to f_k}(\mathbf{x}_z)$ and $m^{(\ell)}_{i\to f_n}(\mathbf{x}_i)$.

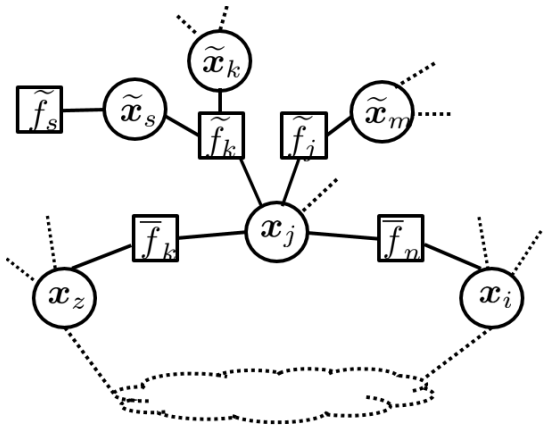
According to (11), computation of the messages in the first group does not depend on messages in the loop and is thus convergence guaranteed. Therefore, the message iteration number is replaced with a $*$ to denote the converged message. Also, from the definition of message computation in (11), if messages in the third group converge, the second group messages should also converge. Therefore, we next focus on showing the convergence of messages in the third group.

For a factor node f_k in the loop with $\mathbf{x}_z$ and $\mathbf{x}_j$ being its two neighboring variable nodes in the loop and $\widetilde{\mathbf{x}}_z$ being its neighboring variable node outside the loop, according to the definition of message computation in (12), we have

$$\begin{aligned} m^{(\ell)}_{f_k\to j}(\mathbf{x}_j) &= \int\int f_k \times m^{(\ell)}_{z\to f_k}(\mathbf{x}_z) \prod_{\widetilde{z}\in\mathcal{B}(f_k)\backslash j} m^{*}_{\widetilde{z}\to f_k}(\widetilde{\mathbf{x}}_z)\, \mathrm{d}\{\widetilde{\mathbf{x}}_z\}_{\widetilde{z}\in\mathcal{B}(f_k)\backslash j}\, \mathrm{d}\mathbf{x}_z, \\ &= \int m^{(\ell)}_{z\to f_k}(\mathbf{x}_z) \left[\int f_k \times \prod_{\widetilde{z}\in\mathcal{B}(f_k)\backslash j} m^{*}_{\widetilde{z}\to f_k}(\widetilde{\mathbf{x}}_z)\, \mathrm{d}\{\widetilde{\mathbf{x}}_z\}_{\widetilde{z}\in\mathcal{B}(f_k)\backslash j}\right] \mathrm{d}\mathbf{x}_z. \end{aligned} \tag{44}$$



(a)



(b)

Figure 2: (a) An example of factor graph with a single loop and chains/trees topology, where the dashed line indicates possible chains/trees; (b) The equivalent factor graph of Fig 2 (a) with new factor functions that do not have neighboring variable nodes except those in the loop.

As shown in Lemma 2, $m^*_{\widetilde{z}\to f_k}(\widetilde{\mathbf{x}}_z)$ must be in Gaussian function form, which is denoted by $m^*_{\widetilde{z}\to f_k}(\widetilde{\mathbf{x}}_z) \propto \mathcal{N}\left(\widetilde{\mathbf{x}}_z|\mathbf{v}^*_{\widetilde{z}\to f_k}, \left[\mathbf{J}^*_{\widetilde{z}\to f_k}\right]^{-1}\right)$. Besides, from (1) we obtain

$$f_k = \mathcal{N}\left(\mathbf{y}_k|\mathbf{A}_{k,z}\mathbf{x}_z + \mathbf{A}_{k,j}\mathbf{x}_j + \sum_{\widetilde{z}\in\mathcal{B}(f_k)} \mathbf{A}_{k,\widetilde{z}}\widetilde{\mathbf{x}}_z, \mathbf{R}_k\right).$$

It can be shown that the inner integration in the second line of (44) is given by

$$\mathcal{N}\left(\overline{\mathbf{y}}_k|\overline{\mathbf{A}}_{k,z}\mathbf{x}_z + \overline{\mathbf{A}}_{k,j}\mathbf{x}_j, \overline{\mathbf{R}}_k\right) \triangleq \overline{f}_k,$$

where the overbar is used to denote the new constant matrix or vector. Then (44) can be written as

$$m^{(\ell)}_{f_k\to j}(\mathbf{x}_j) = \int \overline{f}_k \times m^{(\ell)}_{z\to f_k}(\mathbf{x}_z)\, \mathrm{d}\mathbf{x}_z. \tag{45}$$

Comparing (45) with (12), we obtain $m^{(\ell)}_{f_k\to j}(\mathbf{x}_j)$ as if $m^{(\ell)}_{\overline{f}_k\to j}(\mathbf{x}_j)$ is being passed to a factor node $\overline{f}_k$. Therefore, a factor graph with a single loop and multiple trees/chains is equivalent to a single loop factor graph in which each factor node has no neighboring variable node outside the loop. As a result, the example of Fig. 2 (a) is equivalent to Fig. 2 (b). In the following, we focus on this equivalent topology for the convergence analysis.

Note that, for arbitrary variable node j in the loop, there are two neighboring factor nodes in the loop. Further, using the notation for the equivalent topology, (42) is reduced to

$$\begin{aligned}\mathbf{v}^{(\ell)}_{j\to\overline{f}_n} = &-\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1}\overline{\mathbf{A}}^T_{k,j}\mathbf{T}^{-1}_{k,j}\overline{\mathbf{A}}_{k,z}\mathbf{v}^{(\ell-1)}_{z\to\overline{f}_k}\\ &\underbrace{+\overline{\mathbf{b}}_{j\to\overline{f}_n} - \sum_{\widetilde{f}_k\in\mathcal{B}(j)\setminus f_n}\sum_{\widetilde{s}\in\mathcal{B}\left(\widetilde{f}_k\right)\setminus j}\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1}\overline{\mathbf{A}}^T_{k,j}\overline{\mathbf{M}}^{-1}_{k,j}\overline{\mathbf{A}}_{k,\widetilde{s}}\mathbf{v}^*_{\widetilde{s}\to\widetilde{f}_k}}_{\triangleq\mathbf{c}_{j\to\overline{f}_n}},\end{aligned} \tag{46}$$

where $\mathbf{v}^*_{\widetilde{s}\to\widetilde{f}_k}$ is the converged mean vector on the chain/tree;

$$\overline{\mathbf{b}}_{j\to\overline{f}_n} = \left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1}\sum_{\overline{f}_k\in\mathcal{B}(j)\setminus\overline{f}_n}\overline{\mathbf{A}}^T_{k,j}\overline{\mathbf{M}}^{-1}_{k,j}\overline{\mathbf{y}}_k$$

with $\overline{\mathbf{M}}_{k,j} = \overline{\mathbf{R}}_k + \sum_{\widetilde{s}\in\mathcal{B}\left(\widetilde{f}_k\right)\setminus j}\overline{\mathbf{A}}_{k,\widetilde{s}}\left[\mathbf{J}^*_{\widetilde{s}\to\widetilde{f}_k}\right]^{-1}\overline{\mathbf{A}}^T_{k,\widetilde{s}}$, and

$$\mathbf{T}_{k,j} = \overline{\mathbf{R}}_k + \overline{\mathbf{A}}_{k,z}\left[\mathbf{J}^*_{z\to\overline{f}_k}\right]^{-1}\overline{\mathbf{A}}^T_{k,z}, \tag{47}$$

with $\mathbf{x}_z$ and $\overline{f}_k$ in the loop where $\overline{f}_k \in \mathcal{B}(j)\setminus\overline{f}_n$ and $\mathbf{x}_z \in \mathcal{B}\left(\overline{f}_k\right)\setminus j$. By multiplying $\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{1/2}$ on both sides of (46), and defining $\boldsymbol{\beta}^{(\ell)}_{j\to\overline{f}_n} = \left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{1/2}\mathbf{v}^{(\ell)}_{j\to\overline{f}_n}$, we have

$$\boldsymbol{\beta}^{(\ell)}_{j\to\overline{f}_n} = -\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1/2}\overline{\mathbf{A}}^T_{k,j}\mathbf{T}^{-1}_{k,j}\overline{\mathbf{A}}_{k,z}\left[\mathbf{J}^*_{z\to\overline{f}_k}\right]^{-1/2}\boldsymbol{\beta}^{(\ell-1)}_{z\to\overline{f}_k} + \left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{1/2}\mathbf{c}_{j\to\overline{f}_n}, \tag{48}$$

Let $\boldsymbol{\beta}^{(\ell-1)}$ contain $\boldsymbol{\beta}^{(\ell-1)}_{z\to\overline{f}_k}$ for all $\mathbf{x}_z$ with $z\in\mathcal{B}\left(\overline{f}_k\right)$ and $\overline{f}_k$ being in the loop, and the index is arranged first on k and then on z. Then, the above equation is written in a compact form as

$$\boldsymbol{\beta}^{(\ell)}_{j\to\overline{f}_n}=-\mathbf{Q}_{j\to\overline{f}_n}\boldsymbol{\beta}^{(\ell-1)}+\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{1/2}\mathbf{c}_{j\to\overline{f}_n}, \tag{49}$$

where $\mathbf{Q}_{j\to\overline{f}_n}$ is a row block matrix with the only nonzero block

$$\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1/2}\overline{\mathbf{A}}^T_{k,j}\mathbf{T}^{-1}_{k,j}\overline{\mathbf{A}}_{k,z}\left[\mathbf{J}^*_{z\to\overline{f}_k}\right]^{-1/2}$$

located at the position corresponding to the position $\boldsymbol{\beta}^{(\ell)}_{z\to\overline{f}_k}$ in $\boldsymbol{\beta}^{(\ell)}$. Then let $\mathbf{Q}$ be a matrix that stacks $\mathbf{Q}_{j\to\overline{f}_n}$ as its row, where j and $\overline{f}_n$ are in the loop with $j\in\mathcal{B}\left(\overline{f}_n\right)$. Besides, let $\mathbf{c}$ be the vector containing the subvector $\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{1/2}\mathbf{c}_{j\to\overline{f}_n}$ with the same order as $\mathbf{Q}_{j\to\overline{f}_n}$ in $\mathbf{Q}$. We have

$$\boldsymbol{\beta}^{(\ell)}=-\mathbf{Q}\boldsymbol{\beta}^{(\ell-1)}+\mathbf{c}. \tag{50}$$

Since $\mathbf{Q}$ is a square matrix, $\rho\left(\mathbf{Q}\right)\leq\sqrt{\rho\left(\mathbf{Q}\mathbf{Q}^T\right)}$ and therefore $\rho\left(\mathbf{Q}\mathbf{Q}^T\right)<1$ is the sufficient condition for the convergence of $\boldsymbol{\beta}^{(\ell)}$. We next investigate the elements in $\mathbf{Q}\mathbf{Q}^T$.

Due to the single loop structure of the graph, every $\boldsymbol{\beta}^{(\ell)}_{j\to\overline{f}_n}$ in (48) would be dependent on a unique $\boldsymbol{\beta}^{(\ell)}_{z\to\overline{f}_k}$, where $\overline{f}_k\in\mathcal{B}\left(j\right)\setminus\overline{f}_n$ and $z\in\mathcal{B}\left(\overline{f}_k\right)\setminus j$ (i.e., the message two hops backward along the loop in the factor graph). Thus, the position of the non-zero block in $\mathbf{Q}_{j\to\overline{f}_n}$ will be different and non-overlapping for different combinations of $(j,\overline{f}_n)$. As a result, there exists a column permutation matrix $\boldsymbol{\Xi}$ such that $\mathbf{Q}\boldsymbol{\Xi}$ is a block diagonal matrix. Therefore, $\left(\mathbf{Q}\boldsymbol{\Xi}\right)\left(\mathbf{Q}\boldsymbol{\Xi}\right)^T=\mathbf{Q}\mathbf{Q}^T$ is also a diagonal matrix, and we can write

$$\mathbf{Q}\mathbf{Q}^T=\texttt{Bdiag}\left\{\mathbf{Q}_{j\to\overline{f}_n}\mathbf{Q}^T_{j\to\overline{f}_n}\right\}_{j\,\in\,\mathcal{B}\left(\overline{f}_n\right)}.$$

As a consequence, $\rho\left(\mathbf{Q}\mathbf{Q}^T\right)<1$ is equivalent to $\rho\left(\mathbf{Q}_{j\to\overline{f}_n}\mathbf{Q}^T_{j\to\overline{f}_n}\right)<1$ for all j and $\overline{f}_n$ in the loop with $j\in\mathcal{B}\left(\overline{f}_n\right)$. Following the definition of $\mathbf{Q}_{j\to\overline{f}_n}$ below (49), we obtain

$$\begin{aligned}\mathbf{Q}_{j\to\overline{f}_n}\mathbf{Q}^T_{j\to\overline{f}_n}&=\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1/2}\overline{\mathbf{A}}^T_{k,j}\mathbf{T}^{-1}_{k,j}\overline{\mathbf{A}}_{k,z}\left[\mathbf{J}^*_{z\to\overline{f}_k}\right]^{-1}\overline{\mathbf{A}}^T_{k,z}\mathbf{T}^{-1}_{k,j}\overline{\mathbf{A}}_{k,j}\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1/2}\\&=\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1/2}\overline{\mathbf{A}}^T_{k,j}\mathbf{T}^{-1}_{k,j}\left(\mathbf{T}_{k,j}-\overline{\mathbf{R}}_k\right)\mathbf{T}^{-1}_{k,j}\overline{\mathbf{A}}_{k,j}\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1/2},\end{aligned} \tag{51}$$

where the second equation follows from the definition of $\mathbf{T}_{k,j}$ in (47). Besides, since $\overline{\mathbf{R}}_k\succ\mathbf{0}$, we have $\mathbf{T}_{k,j}-\overline{\mathbf{R}}_k\prec\mathbf{T}_{k,j}$. Following P B.2 in Appendix B, and due to $\mathbf{T}_{k,j}=\mathbf{T}^T_{k,j}$, we have

$$\mathbf{T}^{-1/2}_{k,j}\left(\mathbf{T}_{k,j}-\overline{\mathbf{R}}_k\right)\mathbf{T}^{-1/2}_{k,j}\prec\mathbf{I}. \tag{52}$$

Applying P B.2 in Appendix B again to (52), and making use of (51), we obtain

$$\mathbf{Q}_{j\to\overline{f}_n}\mathbf{Q}^T_{j\to\overline{f}_n}\prec\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1/2}\overline{\mathbf{A}}^T_{k,j}\mathbf{T}^{-1}_{k,j}\overline{\mathbf{A}}_{k,j}\left[\mathbf{J}^*_{j\to\overline{f}_n}\right]^{-1/2}. \tag{53}$$

According to (47), we have

$$\overline{\mathbf{A}}_{k,j}^T \mathbf{T}_{k,j}^{-1} \overline{\mathbf{A}}_{k,j} = \overline{\mathbf{A}}_{k,j}^T \left[\overline{\mathbf{R}}_k + \overline{\mathbf{A}}_{k,z} \left[\mathbf{J}^*_{z \to \overline{f}_k} \right]^{-1} \overline{\mathbf{A}}_{k,z}^T \right]^{-1} \overline{\mathbf{A}}_{k,j}. \tag{54}$$

On the other hand, using (19), due to $\mathcal{B}\left(\overline{f}_k\right) \setminus j = \mathbf{x}_z$ in the considered topology, the right hand side of (54) is $\mathbf{J}^*_{\overline{f}_k \to j}$. Therefore, (53) is further written as

$$\mathbf{Q}_{j \to \overline{f}_n} \mathbf{Q}_{j \to \overline{f}_n}^T \prec \left[\mathbf{J}^*_{j \to \overline{f}_n} \right]^{-1/2} \mathbf{J}^*_{\overline{f}_k \to j} \left[\mathbf{J}^*_{j \to \overline{f}_n} \right]^{-1/2}. \tag{55}$$

From (16), $\mathbf{J}^*_{j \to \overline{f}_n} = \mathbf{W}_j^{-1} + \mathbf{J}^*_{\overline{f}_k \to j} + \sum_{\widetilde{f}_k \in \mathcal{B}(j) \setminus \overline{f}_n} \mathbf{J}^*_{\widetilde{f}_k \to j}$, thus $\mathbf{J}^*_{\overline{f}_k \to j} \preceq \mathbf{J}^*_{j \to \overline{f}_n}$. Therefore, $\left[\mathbf{J}^*_{j \to \overline{f}_n} \right]^{-1/2} \mathbf{J}^*_{\overline{f}_k \to j} \left[\mathbf{J}^*_{j \to \overline{f}_n} \right]^{-1/2} \preceq \mathbf{I}$, and, together with (55), we have

$$\mathbf{Q}_{j \to \overline{f}_n} \mathbf{Q}_{j \to \overline{f}_n}^T \prec \mathbf{I}.$$

Hence $\rho\left(\mathbf{Q}_{j \to \overline{f}_n} \mathbf{Q}_{j \to \overline{f}_n}^T \right) < 1$ for all j and $\overline{f}_n$ in the loop and $j \in \mathcal{B}\left(\overline{f}_n\right)$, and equivalently $\rho(\mathbf{Q}) < 1$. This completes the proof. ■

4.3 Convergence of Belief Covariance and Mean Vector

As the computation of the belief covariance $\mathbf{P}_i^{(\ell)}$ depends on the message information matrix $\mathbf{J}^{(\ell)}_{f_n \to i}$, using Theorems 6 and 11, we can derive the convergence and uniqueness properties of $\mathbf{P}_i^{(\ell)}$.

Before we present the main result, we first present some properties of the part metric $\mathrm{d}(\mathbf{X}, \mathbf{Y})$, with positive definite arguments $\mathbf{X}$, $\mathbf{Y}$, and $\triangle\mathbf{X}$. The proofs are provided in Appendix D.

Proposition 15 *The part metric* $\mathrm{d}(\boldsymbol{X}, \boldsymbol{Y})$ *satisfies the following properties*

P 15.1: $\mathrm{d}(\mathbf{X}_1 + \mathbf{X}_2, \mathbf{Y}_1 + \mathbf{Y}_2) \leq \mathrm{d}(\mathbf{X}_1, \mathbf{Y}_1) + \mathrm{d}(\mathbf{X}_2, \mathbf{Y}_2)$;
P 15.2: $\mathrm{d}(\mathbf{X}, \mathbf{Y}) = \mathrm{d}\left(\mathbf{X}^{-1}, \mathbf{Y}^{-1}\right)$.

We now have the following result.

Corollary 16 *With arbitrary initial message information matrix* $\boldsymbol{J}^{(0)}_{f_n \to i} \succeq \boldsymbol{0}$ *for all* $i \in \mathcal{V}$ *and* $f_n \in \mathcal{B}(i)$*, the belief covariance matrix* $\boldsymbol{P}_i^{(\ell)}$ *converges to a unique p.d. matrix at an exponential rate with respect to any matrix norm before* $\boldsymbol{P}_i^{(\ell)}$ *enters* $\boldsymbol{P}_i^*$*'s neighborhood, which can be chosen to be arbitrarily small.*

Proof Since $\mathbf{J}^{(\ell)}_{f_n \to i}$ converges to a p.d. matrix, and according to (23), $\mathbf{P}_i^{(\ell)}$ also converges. Below, we study the convergence rate of $\mathbf{P}_i^{(\ell)}$. According to the definition of $\mathbf{P}_i^{(\ell)}$ in (23)

and part metric in Definition 9, we have

$$\mathrm{d}\left([\mathbf{P}_i^{(\ell)}]^{-1},[\mathbf{P}_i^{*}]^{-1}\right)=\mathrm{d}\left(\mathbf{W}_i^{-1}+\sum_{f_n\in\mathcal{B}(i)}\mathbf{J}_{f_n\to i}^{(\ell)},\mathbf{W}_i^{-1}+\sum_{f_n\in\mathcal{B}(i)}\mathbf{J}_{f_n\to i}^{*}\right).$$

By applying P 15.1 to the above equation, we obtain

$$\mathrm{d}\left([\mathbf{P}_i^{(\ell)}]^{-1},[\mathbf{P}_i^{*}]^{-1}\right)\leq\mathrm{d}\left(\mathbf{W}_i^{-1},\mathbf{W}_i^{-1}\right)+\sum_{f_n\in\mathcal{B}(i)}\mathrm{d}\left(\mathbf{J}_{f_n\to i}^{(\ell)},\mathbf{J}_{f_n\to i}^{*}\right)=\sum_{f_n\in\mathcal{B}(i)}\mathrm{d}\left(\mathbf{J}_{f_n\to i}^{(\ell)},\mathbf{J}_{f_n\to i}^{*}\right).$$

According to (36), for all $i\in\mathcal{V}$ and $f_n\in\mathcal{B}(i)$, there exist a $c<1$ such that

$$\mathrm{d}\left(\mathbf{J}_{f_n\to i}^{(\ell)},\mathbf{J}_{f_n\to i}^{*}\right)<c^{\ell}\mathrm{d}\left(\mathbf{J}_{f_n\to i}^{(0)},\mathbf{J}_{f_n\to i}^{*}\right).$$

Applying the above inequality to compute $\left[\mathbf{P}_i^{(\ell)}\right]^{-1}$ in (24), we obtain

$$\mathrm{d}\left([\mathbf{P}_i^{(\ell)}]^{-1},[\mathbf{P}_i^{*}]^{-1}\right)<c^{\ell}\sum_{f_n\in\mathcal{B}(i)}\mathrm{d}\left(\mathbf{J}_{f_n\to i}^{(0)},\mathbf{J}_{f_n\to i}^{*}\right).$$

Following P 15.2, the above inequality is equivalent to

$$\mathrm{d}\left(\mathbf{P}_i^{(\ell)},\mathbf{P}_i^{*}\right)<c^{\ell}\sum_{f_n\in\mathcal{B}(i)}\mathrm{d}\left(\mathbf{J}_{f_n\to i}^{(0)},\mathbf{J}_{f_n\to i}^{*}\right),$$

where $\sum_{f_n\in\mathcal{B}(i)}\mathrm{d}\left(\mathbf{J}_{f_n\to i}^{(0)},\mathbf{J}_{f_n\to i}^{*}\right)$ is a constant. Following the same procedure as that from (36) to (38), we can prove that $\mathbf{P}_i^{(\ell)}$ converges at an exponential rate with respect to the monotone norm before $\mathbf{P}_i^{(\ell)}$ enters $\mathbf{P}_i^{*}$'s neighborhood, which can be chosen to be arbitrarily small. ■

On the other hand, as shown in (24), the computation of the belief mean $\boldsymbol{\mu}_i^{(\ell)}$ depends on the belief covariance $\mathbf{P}_i^{(\ell)}$ and the message mean $\mathbf{v}_{f_n\to i}^{(\ell)}$. Thus, under the same condition as in Theorem 13, $\boldsymbol{\mu}_i^{(\ell)}$ is convergence guaranteed. Moreover, it is shown in (Weiss and Freeman, 2001b, Appendix) that, for Gaussian BP over a factor graph, the converged value of belief mean equals the optimal estimate in (3). Together with the convergence guaranteed topology revealed in Theorem 14, we have the following Corollary.

Corollary 17 *With arbitrary $\boldsymbol{J}_{f_n\to i}^{(0)}\succeq\boldsymbol{0}$ and arbitrary $\boldsymbol{v}_{f_n\to i}^{(0)}$ for all $i\in\mathcal{V}$ and $f_n\in\mathcal{B}(i)$, the mean vector $\boldsymbol{\mu}_i^{(\ell)}$ in (24) converges to the optimal estimate $\widehat{\boldsymbol{x}}_i$ in (3) if and only if $\rho(\boldsymbol{Q})<1$, where $\boldsymbol{Q}$ is defined in (43). Furthermore, a sufficient condition to guarantee $\rho(\boldsymbol{Q})<1$ is when the factor graph contains only one single loop connected to multiple chains/trees.*

5. Relationships with Existing Convergence Conditions

In this section, we show the relationship between our convergence condition for Gaussian BP and the recent proposed path-sum method (Giscard et al., 2016). We also show that our convergence condition is more general than the walk-summable condition (Malioutov et al., 2006) for the scalar case.

5.1 Relationship with the Path-sum Method

The path-sum method is proposed in (Giscard et al., 2012, 2013, 2016) to compute $\left(\mathbf{W}^{-1}+\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A}\right)^{-1}$ in (3), in which the matrix inverse $\left(\mathbf{W}^{-1}+\mathbf{A}^T\mathbf{R}^{-1}\mathbf{A}\right)^{-1}$ is interpreted as the sum of simple paths and simple cycles on a weighted graph. The resulting formulation is guaranteed to converge to the correct value for any valid multivariate Gaussian distribution.

The BP message update equations (16), (17), (19), and (20) can be seen as a cut-off of path-sum by retaining only self-loops and backtracks (simple cycles of lengths one and two). In the presence of a graph with one or more loops, equations (16), (17), (19), and (20) do not include the terms related to simple cycles with length larger than 2. This may be a potential cause for the possible divergence of the Gaussian BP algorithm. From this perspective, the divergence can be averted if none of the walks going around the loop(s) have weight greater than one, or equivalently, that the spectral radius of the block matrix representing the loop(s) is strictly less than one. This is an intuitive explanation of the condition $\rho(\mathbf{Q})<1$ obtained in Theorem 13. It also immediately follows from these considerations that the convergence rate is at least geometric, with a cut-off of order ℓ yielding an $\mathcal{O}(\rho(\mathbf{Q})^{\ell})$ error[4].

While the path-sum framework provides an insightful interpretation of the results obtained in this paper, the path-sum algorithm may not be efficiently implementable in distributed and parallel settings, as it requires the summation over all the paths of any length. In contrast, Gaussian BP, while paying the price of non-convergence in general loopy models, makes it possible to realize parallel and fully distributed inference. In summary, though the path-sum method converges for arbitrary valid Gaussian models, it is difficult to be adapted to a distributed and parallel inference setup as the Gaussian BP method.

5.2 Relationship with the Walk-Summable Condition

We show next that, in the setup of linear Gaussian models, the condition $\rho(\mathbf{Q})<1$ as in Corollary 17 encompasses the Gaussian MRF based walk-summable (Malioutov et al. (2006)) in terms of convergence. As all existing results on Gaussian BP convergence (Malioutov et al., 2006; Moallemi and Roy, 2009b) only apply to scalar variables, we restrict the following discussion to only the scalar case. In (Malioutov et al., 2006), the starting point for the convergence analysis for Gaussian MRF is a joint multivariate Gaussian distribution

$$q(\mathbf{x})\propto\exp\Big\{-\frac{1}{2}\mathbf{x}^T\mathbf{J}\mathbf{x}+\mathbf{h}^T\mathbf{x}\Big\},\tag{56}$$

4. We thank an anonymous reviewer for this interpretation.

expressed in the normalized information form such that $\mathbf{J}_{i,i} = 1$ for all i. The underlying Gaussian distribution is factorized as (Malioutov et al. (2006))

$$q(\mathbf{x}) \propto \prod_{i\in\mathcal{V}} \psi_i(\mathbf{x}_i) \prod_{J_{i,j}\neq 0; i\leq j} \psi_{i,j}(\mathbf{x}_i, \mathbf{x}_j), \tag{57}$$

where

$$\psi_i(\mathbf{x}_i) = \exp\left\{-\frac{1}{2}\mathbf{J}_{i,i}\mathbf{x}_i^2 + \mathbf{h}_i\right\} \quad \text{and} \quad \psi_{i,j}(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\mathbf{x}_i\mathbf{J}_{i,j}\mathbf{x}_j\}.$$

In Malioutov et al. (2006, Proposition 1), based on the interpretation that $\left[\mathbf{J}^{-1}\right]_{i,j}$ is the sum of the weights of all the walks from variable j to variable i on the corresponding Gaussian MRF, a sufficient Gaussian BP convergence condition known as walk-summability is provided, which is equivalent to

$$\mathbf{I} - |\mathbf{R}| \succ \mathbf{0}, \tag{58}$$

together with the initial message variance inverse being set to 0, where $\mathbf{R} = \mathbf{I} - \mathbf{J}$ and $|\mathbf{R}|$ is matrix of entrywise absolute values of $\mathbf{R}$. In the following, we establish the relationship between walk-summable Gaussian MRF and linear Gaussian model by utilizing properties of H-matrices (Boman et al., 2005). We show that, with Gaussian MRF satisfying the walk-summable condition, the joint distribution $q(\mathbf{x})$ in (57) can be reformulated as a special case of the linear Gaussian model based factorization in (10). Moreover, Gaussian BP on this particular linear Gaussian model always converges.

Definition 18 *H-Matrices (Boman et al., 2005): A matrix $\boldsymbol{X}$ is an H-matrix if all eigenvalues of the matrix $\mathcal{H}(\boldsymbol{X})$, where $[\mathcal{H}(\boldsymbol{X})]_{i,i} = |\boldsymbol{X}_{i,i}|$, and $[\mathcal{H}(\boldsymbol{X})]_{i,j} = -|\boldsymbol{X}_{i,j}|$ have positive real parts.*

Proposition 19 *Factor width at most 2 factorization (Boman et al., 2005, Theorem 9): A symmetric H-matrix $\boldsymbol{X}$ that has non-negative diagonals can always be factorized as $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{V}^T$, where $\boldsymbol{V}$ is a real matrix with each column of $\boldsymbol{V}$ containing at most 2 non-zeros.*

Let ω be an arbitrary positive value that is smaller than the minimum eigenvalue of $\mathbf{I} - |\mathbf{R}|$ and also satisfies $0 < \omega < 1$. According to (58), we have $(1-\omega)\mathbf{I} - |\mathbf{R}| \succ \mathbf{0}$. Furthermore, by applying $\mathcal{H}(\cdot)$ to $(1-\omega)\mathbf{I} - \mathbf{R}$, we have $[\mathcal{H}((1-\omega)\mathbf{I} - \mathbf{R})]_{i,i} = |(1-\omega)\mathbf{I} - \mathbf{R}|_{i,i} = 1 - \omega$ and $[\mathcal{H}((1-\omega)\mathbf{I} - \mathbf{R})]_{i,j} = -|[(1-\omega)\mathbf{I} - \mathbf{R})]_{i,j}| = -|\mathbf{R}_{i,j}|$. Thus, $\mathcal{H}((1-\omega)\mathbf{I} - \mathbf{R}) = (1-\omega)\mathbf{I} - |\mathbf{R}| \succ \mathbf{0}$, and we conclude that $(1-\omega)\mathbf{I} - \mathbf{R}$ is an H-matrix. According to Proposition 19, $(1-\omega)\mathbf{I} - \mathbf{R} = \mathbf{J} - \omega\mathbf{I} = \mathbf{V}\mathbf{V}^T$, where each column of $\mathbf{V}$ contains at most 2 non-zeros. Now, we can rewrite the joint distribution in (57) as

$$\begin{aligned} q(\mathbf{x}) &\propto \exp\left\{-\frac{1}{2}\mathbf{x}^T(\mathbf{J} - \omega\mathbf{I})\mathbf{x} - \frac{1}{2}\omega\mathbf{x}^T\mathbf{x} + \mathbf{h}^T\mathbf{x}\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\mathbf{V}^T\mathbf{x}\right)^T\left(\mathbf{V}^T\mathbf{x}\right) - \frac{1}{2}\left(\omega\mathbf{x}^T\mathbf{x} - 2\mathbf{h}^T\mathbf{x}\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\sum_{n=1}^{M}\left(V_{n,n_i}x_{n_i} + V_{n,n_j}x_{n_j}\right)^2 - \frac{1}{2}\sum_{n=1}^{M}\omega(x_n - \frac{h_n}{\omega})^2\right\}, \end{aligned} \tag{59}$$

where V_{n,n_i} and V_{n,n_j} denote the two possible non-zero elements on the n-th column and n_i-th and n_j-th rows, and M is the dimension of $\mathbf{x}$. Thus, a walk-summable Gaussian MRF in (56) (or equivalently (57)) can always be written as

$$q(\mathbf{x}) \propto \prod_{n=1}^{M} \underbrace{\mathcal{N}(x_n|\frac{1}{\omega}h_n, \frac{1}{\omega})}_{p(x_n)} \prod_{n=1}^{M} \underbrace{\mathcal{N}(0|V_{n,n_i}x_{n_i} + V_{n,n_j}x_{n_j}, 1)}_{f_n}. \tag{60}$$

Note that, in the above equation, $p(x_n)$ serves as the prior distribution for x_n as that in (10) and f_n is the local likelihood function with local observation being $y_n = 0$ and noise distribution $z_n \sim \mathcal{N}(z_n|0,1)$[5]. Thus the above equation is a special case of the linear Gaussian model based factorization in (10) with the local likelihood function f_n containing only a pair of variables. For this pairwise linear Gaussian model with scalar variables, it is shown in (Moallemi and Roy, 2009b) that $\rho(\mathbf{Q}) < 1$ is fulfilled. Thus by Corollary 17, Gaussian BP always converges. In summary, for the factorization based on Gaussian MRF, if the walk-summable convergence condition is fulfilled, there is an equivalent joint distribution factorization based on linear Gaussian model; and Gaussian BP is convergence guaranteed for this linear Gaussian model.

In the following, we further demonstrate through an example that there exist Gaussian MRFs, in which the information matrix $\mathbf{J}$ fails to satisfy the walk-summable condition, but a convergence guaranteed Gaussian BP update based on the distributed linear Gaussian model representation can still be obtained. More specifically, consider the following information matrix $\mathbf{J}$ in a Gaussian MRF:

$$\mathbf{J} = \begin{bmatrix} 1 & \frac{1}{3\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{3} \\ \frac{1}{3\sqrt{2}} & 1 & 0 & \frac{1}{3} \\ \frac{1}{\sqrt{3}} & 0 & 1 & \frac{1}{\sqrt{6}} \\ \frac{\sqrt{2}}{3} & \frac{1}{3} & \frac{1}{\sqrt{6}} & 1 \end{bmatrix}. \tag{61}$$

The eigenvalues of $\mathbf{I} - |\mathbf{R}|$ to 4 decimal places are -0.0754, 0.9712, 1.4780, and 1.6262. According to the walk-summable definition in (58), it is non walk-summable and the convergence condition in (Malioutov et al., 2006) is inconclusive as to whether Gaussian BP converges. On the other hand, we can study the Gaussian BP convergence of this example by employing a linear Gaussian model representation, and rewriting $\mathbf{J}$ as $\mathbf{J} = \mathbf{A}^T\mathbf{R}^{-1}\mathbf{A} + \mathbf{W}^{-1}$, where

$$\mathbf{A} = \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 6 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

and $\mathbf{R} = \mathbf{I}$. In Fig. 3, the joint distribution of this example with Gaussian MRF and the corresponding linear Gaussian model are represented by factor graphs. As it is shown in Corollary 17, for a factor graph that is the union of a forest and a single loop, as in Fig. 3(b), Gaussian BP always converges to the exact value. This is in sharp contrast to

5. For a particular f_n, if there is only one non-zero coefficient, $f_n \times \mathcal{N}(x_n|\frac{1}{\omega}h_n, \frac{1}{\omega})$ is also proportional to a Gaussian distribution, which can be seen as a prior distribution in (10).

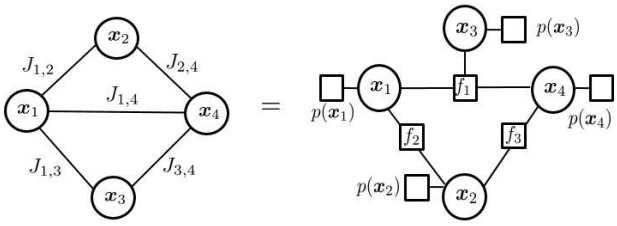


Figure 3: The Gaussian MRF corresponding to $\mathbf{J}$ in (61) with the factorization following (4); (b) The factor graph corresponding to $\mathbf{J}$ in (61) with the factorization following (10).

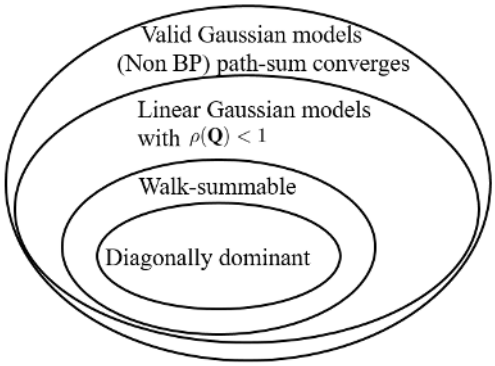


Figure 4: Venn diagram summarizing various subclasses of Gaussian models: the three inner most conditions are all for the BP algorithm while the path-sum in general does not constitute a BP algorithm.

the inconclusive convergence property when the same joint distribution is expressed using the classical Gaussian MRF in (4).

In summary, we have shown that the linear Gaussian model with $\rho(\mathbf{Q}) < 1$ encompasses walk-summable Gaussian MRF. Further, it is shown in (Malioutov et al., 2006) that the diagonally dominant convergence condition in (Weiss and Freeman, 2001a) for Gaussian BP is a special case of the walk-summable condition. Also, the convergence condition in (Su and Wu, 2015) is encompassed by walk-summable condition. Therefore, we have the Venn diagram in Fig. 4 summarizing the relations (in terms of convergence guarantees) between the convergence condition proposed in this paper and existing conditions.

6. Conclusions

This paper shows that, depending on how the factorization of the underlying joint Gaussian distribution is performed, Gaussian belief propagation (BP) may exhibit different convergence properties as different factorizations lead to fundamentally different recursive update structures. The paper studies the convergence of Gaussian BP derived from the factorization

based on the distributed linear Gaussian model. We show that the condition we present for convergence of the marginal mean based on factorizations using the linear Gaussian model is more general than the walk-summable condition (Malioutov et al., 2006) (and references therein) that is based on the Gaussian Markov random field factorization. Further, the linear Gaussian model that is studied in this paper readily conforms to the physical network topology arising in large-scale networks.

Further, the paper analyzes the convergence of the Gaussian BP based distributed inference algorithm. In particular, we show analytically that, with arbitrary positive semidefinite matrix initialization, the message information matrix exchanged among agents converges to a unique positive definite matrix, and it approaches an arbitrarily small neighborhood of this unique positive definite matrix at an exponential rate (with respect to any matrix norm). Regarding the initial information matrix, there exist positive definite initializations that guarantee faster convergence than the commonly used all-zero matrix. Moreover, under the positive semidefinite initial message information matrix condition, we present a necessary and sufficient condition of the belief mean vector to converge to the optimal centralized estimate. We also prove that Gaussian BP always converges if the corresponding factor graph is a union of a single loop and a forest. In particular, we show that the proposed convergence condition for Gaussian BP based on the linear Gaussian model leads to a strictly larger class of models in which Gaussian BP converges than those postulated by the Gaussian Markov random field based walk-summable condition. Finally, we discuss connections of Gaussian BP with the general path-sum algorithm. In the future, it will be interesting to explore if these path-sum interpretations can lead to modifications of the standard Gaussian BP algorithm that guarantee the convergence of Gaussian BP for larger classes of topologies while being also parallel and fully distributed.

Acknowledgments

We thank the reviewers for giving detailed comments and suggestions that have been helpful in improving this paper.

We would like to acknowledge support for this project from the National Science Foundation (NSF grant # CCF1513936), National Natural Science Foundation of China (NSFC grant 61601524), Macau Science and Technology Development Fund under grant FDCT 091/2015/A3, Research Committee of University of Macau under Grant MYRG2014-00146-FST and Grant MYRG2016-00146-FST, and the General Research Fund (GRF) from Hong Kong Research Grant Council (Project No.: 17212416).

Appendix A.

We first compute the first round updating message from variable node to factor node. Substituting $m^{(0)}_{f_n\to i}(\mathbf{x}_i) \propto \exp\left\{-\frac{1}{2}||\mathbf{x}_j - \mathbf{v}^{(0)}_{f_n\to i}||^2_{\mathbf{J}^{(0)}_{f_n\to i}}\right\}$ into (11) and, after algebraic manipulations, we obtain

$$m^{(1)}_{j\to f_n}(\mathbf{x}_j) \propto \exp\left\{-\frac{1}{2}||\mathbf{x}_j - \mathbf{v}^{(1)}_{j\to f_n}||^2_{\mathbf{J}^{(1)}_{j\to f_n}}\right\}, \tag{62}$$

with

$$\mathbf{J}^{(1)}_{j\to f_n} = \mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n} \mathbf{J}^{(0)}_{f_k\to j},$$

and

$$\mathbf{v}^{(1)}_{j\to f_n} = \left[\mathbf{J}^{(1)}_{j\to f_n}\right]^{-1}\left[\sum_{f_k\in\mathcal{B}(j)\backslash f_n} \mathbf{J}^{(0)}_{f_k\to j}\mathbf{v}^{(0)}_{f_k\to j}\right].$$

Next, we evaluate $m^{(1)}_{f_n\to i}(\mathbf{x}_i)$. By substituting $m^{(1)}_{j\to f_n}(\mathbf{x}_j)$ in (62) into (12), we obtain

$$\begin{aligned} m^{(1)}_{f_n\to i}(\mathbf{x}_i) \propto \int\ldots\int \exp\Big\{-\frac{1}{2}\big(\mathbf{y}_n - \sum_{j\in\mathcal{B}(f_n)}\mathbf{A}_{n,j}\mathbf{x}_j\big)^T\mathbf{R}^{-1}\big(\mathbf{y}_n - \sum_{j\in\mathcal{B}(f_n)}\mathbf{A}_{n,j}\mathbf{x}_j\big)\Big\}\times \\ \prod_{j\in\mathcal{B}(f_n)\backslash i}\exp\Big\{-\frac{1}{2}||\mathbf{x}_j - \mathbf{v}^{(1)}_{j\to f_n}||^2_{\mathbf{J}^{(1)}_{j\to f_n}}\Big\}\,\mathrm{d}\{\mathbf{x}_j\}_{j\in\mathcal{B}(f_n)\backslash i}. \end{aligned} \tag{63}$$

Let $\mathbf{x}_{\{\mathcal{B}(f_n)\backslash i\}}$ and $\mathbf{v}^{(1)}_{\{\mathcal{B}(f_n)\backslash i\}\to f_n}$ be stacked vectors containing $\mathbf{x}_j$ and $\mathbf{v}^{(1)}_{j\to f_n}$ as vector elements for all $j\in\mathcal{B}(f_n)\setminus i$ arranged in ascending order on j, respectively; $\mathbf{A}_{n,\{\mathcal{B}(f_n)\backslash i\}}$ denotes a row block matrix containing $\mathbf{A}_{n,j}$ as row elements for all $j\in\mathcal{B}(f_n)\setminus i$ arranged in ascending order; and $\mathbf{J}^{(1)}_{\{\mathcal{B}(f_n)\backslash i\}\to f_n}$ is a block diagonal matrix with $\mathbf{J}^{(1)}_{j\to f_n}$ as its block diagonal elements for all $j\in\mathcal{B}(f_n)\setminus i$ arranged in ascending order. Then, (63) can be reformulated as

$$\begin{aligned} m^{(1)}_{f_n\to i}(\mathbf{x}_i) \propto & \int\ldots\int \exp\left\{-\frac{1}{2}||\mathbf{y}_n - \mathbf{A}_{n,i}\mathbf{x}_i - \mathbf{A}_{n,\{\mathcal{B}(f_n)\backslash i\}}\mathbf{x}^T_{\{\mathcal{B}(f_n)\backslash i\}}||^2_{\mathbf{R}^{-1}}\right\} \\ & \times\exp\left\{-\frac{1}{2}||\mathbf{x}_{\{\mathcal{B}(f_n)\backslash i\}} - \mathbf{v}^{(1)}_{\{\mathcal{B}(f_n)\backslash i\}\to f_n})||^2_{\widetilde{\mathbf{J}}^{(1)}_{\{\mathcal{B}(f_n)\backslash i\}\to f_n}}\right\}\mathrm{d}\mathbf{x}_{\{\mathcal{B}(f_n)\backslash i\}} \\ \propto & \exp\left\{-\frac{1}{2}||\mathbf{y}_n - \mathbf{A}_{n,i}\mathbf{x}_i||^2_{\mathbf{R}^{-1}}\right\} \\ & \times\int\ldots\int\exp\left\{-\frac{1}{2}(\mathbf{x}^T_{\{\mathcal{B}(f_n)\backslash i\}}\mathbf{K}^{(1)}_{f_n\to i}\mathbf{x}_{\{\mathcal{B}(f_n)\backslash i\}} - 2[\mathbf{h}^{(1)}_{f_n\to i}]^T\mathbf{x}_{\{\mathcal{B}(f_n)\backslash i\}})\right\}\mathrm{d}\mathbf{x}_{\{\mathcal{B}(f_n)\backslash i\}}, \end{aligned} \tag{64}$$

where

$$\mathbf{K}^{(1)}_{f_n\to i} = \mathbf{A}^T_{n,\{\mathcal{B}(f_n)\backslash i\}}\mathbf{R}^{-1}\mathbf{A}_{n,\{\mathcal{B}(f_n)\backslash i\}} + \mathbf{J}^{(1)}_{\{\mathcal{B}(f_n)\backslash i\}\to f_n}$$

and

$$\mathbf{h}^{(1)}_{f_n\to i} = \mathbf{A}^T_{n,\{\mathcal{B}(f_n)\backslash i\}}\mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{A}_{n,i}\mathbf{x}_i) + \mathbf{J}^{(1)}_{\{\mathcal{B}(f_n)\backslash i\}\to f_n}\mathbf{v}^{(1)}_{\{\mathcal{B}(f_n)\backslash i\}\to f_n}.$$

By completing the square for the integrand of (64), we obtain

$$m^{(1)}_{f_n\to i}(\mathbf{x}_i) \propto \alpha^{(1)}_{f_n\to i}\exp\left\{-\frac{1}{2}||\mathbf{y}_n-\mathbf{A}_{n,i}\mathbf{x}_i||^2_{\mathbf{R}_n^{-1}}+\frac{1}{2}\left[\mathbf{h}^{(1)}_{f_n\to i}\right]^T\left[\mathbf{K}^{(1)}_{f_n\to i}\right]^{-1}\mathbf{h}^{(\ell)}_{f_n\to i}\right\}, \quad (65)$$

with

$$\alpha^{(1)}_{f_n\to i}=\int\ldots\int\exp\left\{-\frac{1}{2}||\mathbf{x}_{\mathcal{B}(f_n)\backslash i\}}-\left[\mathbf{K}^{(1)}_{f_n\to i}\right]^{-1}\mathbf{h}^{(1)}_{f_n\to i}||^2_{\mathbf{K}^{(1)}_{f_n\to i}}\right\}\mathrm{d}\mathbf{x}_{\mathcal{B}(f_n)\backslash i\}}.$$

Next, by applying the spectral theorem to $\mathbf{K}^{(1)}_{f_n\to i}$ and after some algebraic manipulations, we simplify (65) as

$$m^{(1)}_{f_n\to i}(\mathbf{x}_i)\propto\alpha^{(1)}_{f_n\to i}\exp\left\{-\frac{1}{2}||\mathbf{x}_i-\mathbf{v}^{(1)}_{f_n\to i}||_{\mathbf{J}^{(1)}_{f_n\to i}}\right\},$$

with the inverse of the covariance, the information matrix

$$\mathbf{J}^{(1)}_{f_n\to i}=\mathbf{A}^T_{n,i}\left[\mathbf{R}_n+\sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{J}^{(1)}_{j\to f_n}\right]^{-1}\mathbf{A}^T_{n,j}\right]^{-1}\mathbf{A}_{n,i},$$

and the mean vector

$$\mathbf{v}^{(1)}_{f_n\to i}=\left[\mathbf{J}^{(1)}_{f_n\to i}\right]^{-1}\mathbf{A}^H_{n,i}\left[\mathbf{R}_n+\sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{J}^{(1)}_{j\to f_n}\right]^{-1}\mathbf{A}^T_{n,j}\right]^{-1}\left(\mathbf{y}_n-\sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\mathbf{v}^{(1)}_{j\to f_n}\right),$$

and

$$\alpha^{(1)}_{f_n\to i}\propto\int\ldots\int\exp\left\{-\frac{1}{2}\mathbf{z}^T\mathbf{\Lambda}^{(1)}_{f_n\to i}\mathbf{z}\right\}\mathrm{d}\mathbf{z}, \quad (66)$$

where $\mathbf{\Lambda}^{(1)}_{f_n\to i}$ is a diagonal matrix containing the eigenvalues of $\mathbf{A}^T_{n,\{\mathcal{B}(f_n)\backslash i\}}\mathbf{R}_n^{-1}\mathbf{A}_{n,\{\mathcal{B}(f_n)\backslash i\}}+\mathbf{J}^{(1)}_{\{\mathcal{B}(f_n)\backslash i\}\to f_n}$.

By induction, and following similar derivations as in (62) to (66), we obtain the general updating expressions as in (15) to (21).

Appendix B.

Before going into the proof of Lemma 2, we note the following properties of positive definite (p.d.) matrices. If $\mathbf{X}\succ\mathbf{0}$, $\mathbf{Y}\succ\mathbf{0}$, $\mathbf{Z}\succeq\mathbf{0}$ are of the same dimension, then we have (Du and Wu, 2013a):

P B.1: $\mathbf{X}+\mathbf{Y}\succ\mathbf{0}$ and $\mathbf{X}+\mathbf{Z}\succ\mathbf{0}$.

P B.2: $\mathbf{A}^T\mathbf{X}\mathbf{A}\succ\mathbf{0}$, $\mathbf{A}^T\mathbf{Z}\mathbf{A}\succeq\mathbf{0}$, $\mathbf{A}\mathbf{X}\mathbf{A}^T\succeq\mathbf{0}$ and $\mathbf{A}\mathbf{Z}\mathbf{A}^T\succeq\mathbf{0}$ for any full column rank matrix $\mathbf{A}$ with compatible dimension.

Now, we prove Lemma 2. If $\mathbf{J}^{(\ell-1)}_{f_k\to j}\succeq\mathbf{0}$ for all $f_k\in\mathcal{B}(j)\setminus f_n$, according to P B.1, $\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell-1)}_{f_k\to j}\succeq\mathbf{0}$. As $\mathbf{W}_j^{-1}\succ\mathbf{0}$, we have $\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell-1)}_{f_k\to j}\succ\mathbf{0}$, which, according to (16), is equivalent to $\mathbf{J}^{(\ell)}_{j\to f_n}\succ\mathbf{0}$. Besides, as $\mathbf{A}_{n,j}$ is full column rank, if $\left[\mathbf{J}^{(\ell)}_{j\to f_n}\right]^{-1}\succ\mathbf{0}$

for all $j \in \mathcal{B}(f_n)\backslash i$, according to P B.2, $\mathbf{A}_{n,j}\left[\mathbf{J}^{(\ell)}_{j\to f_n}\right]^{-1}\mathbf{A}^T_{n,j} \succeq \mathbf{0}$. With $\mathbf{R}_n \succ \mathbf{0}$, following P B.1, we have $\left[\mathbf{R}_n + \sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{J}^{(\ell)}_{j\to f_n}\right]^{-1}\mathbf{A}^T_{n,j}\right]^{-1} \succ \mathbf{0}$. As $\mathbf{A}_{n,i}$ is of full column rank, by applying P B.2 again, we have $\mathbf{A}^T_{n,i}\left[\mathbf{R}_n + \sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{J}^{(\ell)}_{j\to f_n}\right]^{-1}\mathbf{A}^T_{n,j}\right]^{-1}\mathbf{A}_{n,i} \succ \mathbf{0}$, which according to (19) is equivalent to $\mathbf{J}^{(\ell)}_{f_n\to i} \succ \mathbf{0}$.

In summary, we have proved that 1) if $\mathbf{J}^{(\ell-1)}_{f_k\to j} \succeq \mathbf{0}$ for all $f_k \in \mathcal{B}(j)\setminus f_n$, then $\mathbf{J}^{(\ell)}_{j\to f_n} \succ \mathbf{0}$; 2) if $\left[\mathbf{J}^{(\ell)}_{j\to f_n}\right]^{-1} \succ \mathbf{0}$ for all $j \in \mathcal{B}(f_n)\setminus i$, then $\mathbf{J}^{(\ell)}_{f_n\to i} \succ \mathbf{0}$. Therefore, by setting $\mathbf{J}^{(0)}_{f_k\to j} \succeq \mathbf{0}$ for all $k \in \mathcal{V}$ and $j \in \mathcal{B}(f_k)$, according to the results of the first case, we have $\mathbf{J}^{(1)}_{j\to f_n} \succ \mathbf{0}$ for all $j \in \mathcal{V}$ and $f_n \in \mathcal{B}(j)$. Then, applying the second case, we further have $\mathbf{J}^{(1)}_{f_n\to i} \succ \mathbf{0}$ for all $n \in \mathcal{V}$ and $i \in \mathcal{B}(f_n)$. By repeatedly using the above arguments, it follows readily that $\mathbf{J}^{(\ell)}_{f_k\to j} \succ \mathbf{0}$ and $\mathbf{J}^{(\ell)}_{j\to f_n} \succ \mathbf{0}$ for $\ell \geq 1$ and with $j \in \mathcal{V}$, $f_n, f_k \in \mathcal{B}(j)$. Furthermore, according to the discussion before Lemma 2, all messages $m^{(\ell)}_{j\to f_n}(\mathbf{x}_j)$ and $m^{(\ell)}_{f_n\to i}(\mathbf{x}_i)$ exist, and are in Gaussian form as in (15) and (18).

Appendix C.

First, Proposition 4, P 4.1 is proved. Suppose that $\mathbf{J}^{(\ell)} \succeq \mathbf{J}^{(\ell-1)} \succeq \mathbf{0}$, i.e., $\mathbf{J}^{(\ell)}_{f_k\to j} \succeq \mathbf{J}^{(\ell-1)}_{f_k\to j} \succeq \mathbf{0}$ for all $(f_k, j) \in \widetilde{\mathcal{B}}(f_n, i)$, we have

$$\mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell)}_{f_k\to j} \succeq \mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell-1)}_{f_k\to j} \succ \mathbf{0}.$$

Then, according to the fact that if $\mathbf{X} \succeq \mathbf{Y} \succ \mathbf{0}$, $\mathbf{Y}^{-1} \succeq \mathbf{X}^{-1} \succ \mathbf{0}$, we have

$$\left[\mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell-1)}_{f_k\to j}\right]^{-1} \succeq \left[\mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell)}_{f_k\to j}\right]^{-1} \succ \mathbf{0}.$$

Since $\mathbf{A}_{n,j}$ is of full column rank and following P B.2 in Appendix B, we have

$$\mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell-1)}_{f_k\to j}\right]^{-1}\mathbf{A}^T_{n,j} \succeq \mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell)}_{f_k\to j}\right]^{-1}\mathbf{A}^T_{n,j} \succ \mathbf{0}.$$

Following the same procedure of the proof above and due to $\mathbf{R} \succ \mathbf{0}$, we can further prove that

$$\begin{aligned}&\mathbf{A}^T_{n,i}\left[\mathbf{R}_n + \sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell)}_{f_k\to j}\right]^{-1}\mathbf{A}^T_{n,j}\right]^{-1}\mathbf{A}_{n,i}\\ \succeq&\mathbf{A}^T_{n,i}\left[\mathbf{R}_n + \sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1} + \sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}^{(\ell-1)}_{f_k\to j}\right]^{-1}\mathbf{A}^T_{n,j}\right]^{-1}\mathbf{A}_{n,i},\end{aligned}$$

which is equivalent to

$$\mathcal{F}_{n\to i}\left(\left\{\mathbf{J}_{f_k\to j}^{(\ell)}\right\}_{(f_k,j)\in\widetilde{\mathcal{B}}(f_n,i)}\right)\succeq\mathcal{F}_{n\to i}\left(\left\{\mathbf{J}_{f_k\to j}^{(\ell-1)}\right\}_{(f_k,j)\in\widetilde{\mathcal{B}}(f_n,i)}\right).$$

Since $\mathcal{F}$ contains $\mathcal{F}_{n\to i}(\cdot)$ as its component, Proposition 4, P 4.1 is proved.

Next, Proposition 4, P 4.2 is proved. Suppose that $\mathbf{J}_{f_k\to j}^{(\ell)}\succ\mathbf{0}$ for all $(f_k,j)\in\widetilde{\mathcal{B}}(f_n,i)$. As $\alpha>1$, we have

$$\alpha\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\alpha\mathbf{J}_{f_k\to j}^{(\ell)}\succeq\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\alpha\mathbf{J}_{f_k\to j}^{(\ell)}\succ\mathbf{0},$$

where the equality holds when $\mathbf{W}_j^{-1}=\mathbf{0}$, which corresponds to non-informative prior for $\mathbf{x}_j$. Applying the fact that if $\mathbf{X}\succeq\mathbf{Y}\succ\mathbf{0}$, $\mathbf{Y}^{-1}\succeq\mathbf{X}^{-1}\succ\mathbf{0}$, and, according to P B.2 in Appendix B, we obtain

$$\mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\alpha\mathbf{J}_{f_k\to j}^{(\ell)}\right]^{-1}\mathbf{A}_{n,j}^T\succeq\mathbf{A}_{n,j}\left[\alpha\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\alpha\mathbf{J}_{f_k\to j}^{(\ell)}\right]^{-1}\mathbf{A}_{n,j}^T\succeq\mathbf{0}.$$

Since $\mathbf{R}_n\succ\frac{1}{\alpha}\mathbf{R}_n\succ\mathbf{0}$, we have

$$\begin{aligned}&\left[\frac{1}{\alpha}\mathbf{R}_n+\sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\alpha\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\alpha\mathbf{J}_{f_k\to j}^{(\ell)}\right]^{-1}\mathbf{A}_{n,j}^T\right]^{-1}\\&\succ\left[\mathbf{R}_n+\sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\alpha\mathbf{J}_{f_k\to j}^{(\ell)}\right]^{-1}\mathbf{A}_{n,j}^T\right]^{-1}.\end{aligned}$$

Finally, applying P B.2 in Appendix B to the above equation and taking out the common factor α, we obtain

$$\begin{aligned}&\alpha\mathbf{A}_{n,i}^T\left[\mathbf{R}_n+\sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\mathbf{J}_{f_k\to j}^{(\ell)}\right]^{-1}\mathbf{A}_{n,j}^T\right]^{-1}\mathbf{A}_{n,i}\\&\succ\mathbf{A}_{n,i}^T\left[\mathbf{R}_n+\sum_{j\in\mathcal{B}(f_n)\backslash i}\mathbf{A}_{n,j}\left[\mathbf{W}_j^{-1}+\sum_{f_k\in\mathcal{B}(j)\backslash f_n}\alpha\mathbf{J}_{f_k\to j}^{(\ell)}\right]^{-1}\mathbf{A}_{n,j}^T\right]^{-1}\mathbf{A}_{n,i}.\end{aligned}$$

Therefore, $\alpha\mathcal{F}_{n\to i}\left(\left\{\mathbf{J}_{f_k\to j}^{(\ell)}\right\}_{(f_k,j)\in\widetilde{\mathcal{B}}(f_n,i)}\right)\succ\mathcal{F}_{n\to i}\left(\left\{\alpha\mathbf{J}_{f_k\to j}^{(\ell)}\right\}_{(f_k,j)\in\widetilde{\mathcal{B}}(f_n,i)}\right)$ if $\mathbf{J}_{f_k\to j}^{(\ell)}\succ\mathbf{0}$ for all $(f_k,j)\in\widetilde{\mathcal{B}}(f_n,i)$ and $\alpha>1$. As $\mathcal{F}$ contains $\mathcal{F}_{n\to i}(\cdot)$ as its component, Proposition 4, P 4.2 is proved. In the same way, we can prove $\mathcal{F}\left(\alpha^{-1}\mathbf{J}^{(\ell)}\right)\succ\alpha^{-1}\mathcal{F}\left(\mathbf{J}^{(\ell)}\right)$ if $\mathbf{J}^{(\ell)}\succ\mathbf{0}$ and $\alpha>1$.

At last, Proposition 4, P 4.3 is proved. From Lemma 2, if we have initial message information matrix $\mathbf{J}_{f_k\to j}^{(0)}\succeq\mathbf{0}$ for all $j\in\mathcal{V}$ and $f_k\in\mathcal{B}(j)$, then we have $\mathbf{J}_{f_k\to j}^{(\ell)}\succ\mathbf{0}$ for all

$j \in \mathcal{V}$ and $f_k \in \mathcal{B}(j)$. In such case, obviously, $\mathbf{J}^{(\ell)} \succeq \mathbf{0}$. Applying $\mathcal{F}$ to both sides of this equation, and using Proposition 4, P 4.1, we have $\mathcal{F}\left(\mathbf{J}^{(\ell)}\right) \succeq \mathcal{F}(\mathbf{0})$. On the other hand, using (27), it can be easily checked that $\mathcal{F}(\mathbf{0}) = \mathbf{A}^T \left[\mathbf{\Omega} + \mathbf{H}\mathbf{\Psi}^{-1}\mathbf{H}^T\right]^{-1} \mathbf{A} \succ \mathbf{0}$, where the inequality is from Lemma 2. For proving the upper bound, we start from the fact that

$$\sum_{j \in \mathcal{B}(f_n) \backslash i} \mathbf{A}_{n,j} \left[\mathbf{W}_j^{-1} + \sum_{f_k \in \mathcal{B}(j) \backslash f_n} \mathbf{J}_{f_k \to j}^{(\ell-1)} \right]^{-1} \mathbf{A}_{n,j}^T$$

in (25), and equivalently the corresponding term

$$\mathbf{H}_{n,i} \left[\mathbf{W}_{n,i} + \mathbf{K}_{n,i} \left(\mathbf{I}_{|\mathcal{B}(f_n)|-1} \otimes \mathbf{J}^{(\ell-1)} \right) \mathbf{K}_{n,i}^T \right]^{-1} \mathbf{H}_{n,i}^T$$

in (26), are p.s.d. matrices. In (27), since

$$\mathbf{H} \left[\mathbf{\Psi} + \mathbf{K} \left(\mathbf{I}_{\sum_{n=1}^{M} |B(f_n)|(|B(f_n)|-1)} \otimes \mathbf{J}^{(\ell-1)} \right) \mathbf{K}^T \right]^{-1} \mathbf{H}^T$$

contains $\mathbf{H}_{n,i} \left[\mathbf{W}_{n,i} + \mathbf{K}_{n,i} \left(\mathbf{I}_{|\mathcal{B}(f_n)|-1} \otimes \mathbf{J}^{(\ell-1)} \right) \mathbf{K}_{n,i}^T \right]^{-1} \mathbf{H}_{n,i}^T$ as its block diagonal elements, it is also a p.s.d. matrix. With $\mathbf{\Omega} \succ \mathbf{0}$, adding to the above result gives

$$\mathbf{\Omega} + \mathbf{H} \left[\mathbf{\Psi} + \mathbf{K} \left(\mathbf{I}_{\varphi} \otimes \mathbf{J}^{(\ell)} \right) \mathbf{K}^T \right]^{-1} \mathbf{H}^T \succeq \mathbf{\Omega} \succ \mathbf{0}.$$

Inverting both sides, we obtain $\mathbf{\Omega}^{-1} \succeq \left[\mathbf{\Omega} + \mathbf{H} \left[\mathbf{\Psi} + \mathbf{K} \left(\mathbf{I}_{\varphi} \otimes \mathbf{J}^{(\ell)} \right) \mathbf{K}^T \right]^{-1} \mathbf{H}^T \right]^{-1}$. Finally, applying P B.2 again gives

$$\mathbf{A}^T \mathbf{\Omega}^{-1} \mathbf{A} \succeq \mathbf{A}^T \left[\mathbf{\Omega} + \mathbf{H} \left[\mathbf{\Psi} + \mathbf{K} \left(\mathbf{I}_{\varphi} \otimes \mathbf{J}^{(\ell)} \right) \mathbf{K}^T \right]^{-1} \mathbf{H}^T \right]^{-1} \mathbf{A} = \mathcal{F}\left(\mathbf{J}^{(\ell)}\right).$$

Therefore, we have $\mathbf{A}^T \mathbf{\Omega}^{-1} \mathbf{A} \succeq \mathcal{F}\left(\mathbf{J}^{(\ell)}\right) \succeq \mathbf{A}^T \left[\mathbf{\Omega} + \mathbf{H}\mathbf{\Psi}^{-1}\mathbf{H}^T\right]^{-1} \mathbf{A} \succ \mathbf{0}$.

Appendix D.

Let $\mathrm{d}(\mathbf{X}_1, \mathbf{Y}_1) = \exp\{a_1\}$ and $\mathrm{d}(\mathbf{X}_2, \mathbf{Y}_2) = \exp\{a_2\}$, and $\mathrm{d}(\mathbf{X}_1 + \mathbf{X}_2, \mathbf{Y}_1 + \mathbf{Y}_2) = \exp\{a_3\}$. First, P 15.1 is proved. According to the definition of part metric in Definition 9, for arbitrary symmetric p.d matrix $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{Y}_1$, and $\mathbf{Y}_2$, we have $\mathrm{d}(\mathbf{X}_1, \mathbf{Y}_1)$, $\mathrm{d}(\mathbf{X}_2, \mathbf{Y}_2)$, and $\mathrm{d}(\mathbf{X}_1 + \mathbf{X}_2, \mathbf{Y}_1 + \mathbf{Y}_2)$ correspond to

$$a_1 \mathbf{X}_1 \succeq \mathbf{Y}_1 \succeq \frac{1}{a_1} \mathbf{X}_1, \quad a_2 \mathbf{X}_2 \succeq \mathbf{Y}_2 \succeq \frac{1}{a_2} \mathbf{X}_2, \tag{67}$$

$$a_3 (\mathbf{X}_1 + \mathbf{X}_2) \succeq \mathbf{Y}_1 + \mathbf{Y}_2 \succeq \frac{1}{a_3} (\mathbf{X}_1 + \mathbf{X}_2). \tag{68}$$

Since $\mathrm{d}(\mathbf{X}_1, \mathbf{Y}_1) > 0$ and $\mathrm{d}(\mathbf{X}_2, \mathbf{Y}_2) > 0$, we have $a_1, a_2 \geq 1$. And therefore $a_1 + a_2 > a_1$ and $a_1 + a_2 > a_2$. Then, according to (67), we have

$$(a_1 + a_2)(\mathbf{X}_1 + \mathbf{X}_2) \succeq \mathbf{Y}_1 + \mathbf{Y}_2 \succeq \frac{1}{a_1 + a_2}(\mathbf{X}_1 + \mathbf{X}_2). \tag{69}$$

Following the definition of part matric, a_3 is the smallest value satisfy the inequality in (68). Thus, by comparing (69) with (68), we obtain $a_1 + a_2 \geq a_3$ Hence, $\mathrm{d}(\mathbf{X}_1 + \mathbf{X}_2, \mathbf{Y}_1 + \mathbf{Y}_2) \leq \mathrm{d}(\mathbf{X}_1, \mathbf{Y}_1) + \mathrm{d}(\mathbf{X}_2, \mathbf{Y}_2)$.

Next, P 15.2 is proved. Following the part metric definition of $\mathrm{d}(\mathbf{X}_1, \mathbf{Y}_1)$, $a_1\mathbf{X}_1 \succeq \mathbf{Y}_1 \succeq \frac{1}{a_1}\mathbf{X}_1$, which is equivalent to $\mathbf{Y}_1^{-1} \succeq \frac{1}{a_1}\mathbf{X}_1^{-1}$ and $a_1\mathbf{X}_1^{-1} \succeq \mathbf{Y}_1^{-1}$. Thus, $\mathrm{d}(\mathbf{X}, \mathbf{Y}) = \mathrm{d}\left(\mathbf{X}^{-1}, \mathbf{Y}^{-1}\right)$.

References

D. Bickson and D. Malkhi. A unifying framework for rating users and data items in peer-to-peer and social networks. *Peer-to-Peer Networking and Applications (PPNA) Journal*, 1(2):93–103, 2008.

E. G. Boman, D. Chen, O. Parekh, and S. Toledo. On factor width and symmetric H-matrices. *Linear algebra and its applications*, 405:239–248, 2005.

F. S. Cattivelli and A. H. Sayed. Diffusion LMS strategies for distributed estimation. *IEEE Trans. Signal Processing*, 58(3):1035–1048, 2010.

M. Chertkov and V. Y. Chernyak. Loop series for discrete statistical models on graphs. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(06):P06009, 2006.

I. Chueshov. *Monotone Random Systems Theory and Applications*. New York: Springer, 2002.

P. G. Ciarlet. *Introduction to Numerical Linear Algebra and Optimisation*. Cambridge University Press, 1989.

J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1997.

J. Du and Y. C. Wu. Network-wide distributed carrier frequency offsets estimation and compensation via belief propagation. *IEEE Trans. Signal Processing*, 61(23):5868–5877, December 2013a.

J. Du and Y. C. Wu. Distributed clock skew and offset estimation in wireless sensor networks: Asynchronous algorithm and convergence analysis. *IEEE Trans. Wireless Commun.*, 12(11):5908–5917, Nov 2013b.

J. Du, S. Kar, and J. M. F. Moura. Distributed convergence verification for Gaussian belief propagation. In *Asilomar Conference on Signals, Systems, and Computers*, 2017a.

J. Du, S. Ma, Y. C. Wu, S. Kar, and J. M. F. Moura. Convergence analysis of belief propagation for pairwise linear Gaussian models. In *IEEE Global Conference on Signal and Information Processing*, 2017b.

B. J. Frey. Local probability propagation for factor analysis. In *Neural Information Processing Systems (NIPS)*, pages 442–448, December 1999.

P.-L. Giscard, S. Thwaite, and D. Jaksch. Walk-sums, continued fractions and unique factorisation on digraphs. *arXiv preprint arXiv:1202.5523*, 2012.

P.-L. Giscard, S. Thwaite, and D. Jaksch. Evaluating matrix functions by resummations on graphs: the method of path-sums. *SIAM Journal on Matrix Analysis and Applications*, 34(2):445–469, 2013.

P.-L. Giscard, Z. Choo, S. Thwaite, and D. Jaksch. Exact inference on Gaussian graphical models of arbitrary topology using path-sums. *Journal of Machine Learning Research*, 7 (2):1–19, February 2016.

V. Gómez, J. M. Mooij, and H. J. Kappen. Truncating the loop series expansion for belief propagation. *Journal of Machine Learning Research*, 8:1987–2016, 2007.

Y. Hu, A. Kuh, T. Yang, and A. Kavcic. A belief propagation based power distribution system state estimator. *IEEE Comput. Intell. Mag.*, 2011.

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*, 6:905–936, 2005.

S. Kar. *Large Scale Networked Dynamical Systems: Distributed Inference*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, Department of Electrical and Computer Engineering, June 2010.

S. Kar and J. M. F. Moura. Consensus + innovations distributed inference over networks: cooperation and sensing in networked systems. *IEEE Signal Process. Mag.*, 30(3):99–109, 2013.

S. Kar, J. M. F. Moura, and H.V. Poor. Distributed linear parameter estimation: asymptotically efficient adaptive strategies. *SIAM Journal on Control and Optimization*, 51(3): 2200–2229, 2013.

U. Krause and R. Nussbaum. A limit set trichotomy for self-mappings of normal cones in Banach spaces. *Nonlinear Analysis, Theory, Methods & Applications*, 20(7):855–870, 1993.

F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Information Theory*, 47(2):498–519, February 2001.

F. Lehmann. Iterative mitigation of intercell interference in cellular networks based on Gaussian belief propagation. *IEEE Trans. Veh. Technol.*, 61(6):2544–2558, July 2012.

D. M. Malioutov, J. K. Johnson, and A. S. Willsky. Walk-sums and belief propagation in Gaussian graphical models. *Journal of Machine Learning Research*, 7(2):2031–2064, February 2006.

C. C. Moallemi and B. Van Roy. Convergence of min-sum message passing for quadratic optimization. *IEEE Trans. Information Theory*, 55(5):2413–2423, 2009a.

C. C. Moallemi and B. Van Roy. Convergence of min-sum message passing for quadratic optimization. *IEEE Transactions on Information Theory*, 55(5):2413–2423, 2009b.

J. M. Mooij and H. J. Kappen. Sufficient conditions for convergence of loopy belief propagation. In F. Bacchus and T. Jaakkola, editors, *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 396–403, Corvallis, Oregon, 2005. AUAI Press.

J. M. Mooij and H. J. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, December 2007.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: an empirical study. In *15th Conf. Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden*, pages 467–475, July 1999.

National Research Council. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press, 2013.

B. L. Ng, J. S. Evans, S. V. Hanly, and D. Aktas. Distributed downlink beamforming with cooperative base stations. *IEEE Trans. Information Theory*, 54(12):5491–5499, December 2008.

N. Noorshams and M. J. Wainwright. Belief propagation for continuous state spaces: Stochastic message-passing with quantitative guarantees. *Journal of Machine Learning Research*, 14:2799–2835, 2013.

S. Ravanbakhsh and R. Greiner. Perturbed message passing for constraint satisfaction problem. *Journal of Machine Learning Research*, 16:1249–1274, 2015.

O. Shental, P. H. Siegel, J. K. Wolf, D. Bickson, and D. Dolev. Gaussian belief propagation solver for systems of linear equations. In *2008 IEEE International Symposium on Information Theory (ISIT 2008)*, pages 1863–1867, July 2008a.

O. Shental, P. H. Siegel, J. K. Wolf, D. Bickson, and D. Dolev. Gaussian belief propagation solver for systems of linear equations. In *2008 IEEE International Symposium on Information Theory*, pages 1863–1867, 2008b.

Q. Su and Y. C. Wu. On convergence conditions of Gaussian belief propagation. *IEEE Trans. Signal Processing*, 63(5):1144–1155, March 2015.

X. Tan and J. Li. Computationally efficient sparse Bayesian learning via belief propagation. *IEEE Trans. Signal Processing*, 58(4):2010–2021, April 2010.

Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation,*, 12:1–41, 2000.

Y. Weiss and W. T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10):2173–2200, March 2001a.

Y. Weiss and W. T. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Trans. Information Theory*, 47(2): 736–744, February 2001b.

R. Xiong, W. Ding, S. Ma, and W. Gao. A practical algorithm for tanner graph based image interpolation. In *2010 IEEE International Conference on Image Processing (ICIP 2010)*, pages 1989–1992, 2010.

E. Zeidler. *Nonlinear Analysis and its Applications IV-Applications to Mathematical Physics*. Springer-Verlag New York Inc., 1985.

G. Zhang, W. Xu, and Y. Wang. Fast distributed rate control algorithm with QoS support in ad-hoc networks. In *2010 IEEE Global Telecommunications Conference (GLOBECOM 2010)*, pages 1–5, December 2010.