CONVERGENCE ANALYSIS OF THE INFORMATION MATRIX IN GAUSSIAN BELIEF PROPAGATION

Jian Du[†], Shaodan Ma[⋆], Yik-Chung Wu[‡], Soummya Kar[†] and José M. F. Moura[†]

Electrical and Computer Engineering, Carnegie Mellon University[†], Pittsburgh, PA
Electrical and Computer Engineering, University of Macau[⋆], Macau
Electrical and Electronic Engineering, The University of Hong Kong[‡], Hong Kong

ABSTRACT

Gaussian belief propagation (BP) has been widely used for distributed estimation in large-scale networks such as the smart grid, communication networks, and social networks, where local measurements/observations are scattered over a wide geographical area. However, the convergence of Gaussian BP is still an open issue. In this paper, we consider the convergence of Gaussian BP, focusing in particular on the convergence of the information matrix. We show analytically that the exchanged message information matrix converges for arbitrary positive semidefinite initial value, and its distance to the unique positive definite limit matrix decreases exponentially fast.

Index Terms— graphical model, belief propagation, large-scale networks, Markov random field.

1. INTRODUCTION

In large-scale linear parameter estimation with Gaussian measurements, Gaussian Belief Propagation (BP) [1] provides an efficient distributed way to compute the marginal distribution of the unknown variables, and it has been adopted in a variety of topics ranging from distributed power state estimation [2] in smart grid, distributed beamforming [3] and synchronization [4, 5] in wireless communication networks, fast solver for system of linear equations [6], distributed rate control in ad-hoc networks [7], factor analyzer network [8], sparse Bayesian learning [9], to peer-to-peer rating in social networks [10]. It has been shown that Gaussian BP computes the optimal centralized estimator if it converges [11].

Although with great empirical success, the major challenge that hinders Gaussian BP to realize its full potential is the lack of theoretical guarantees of convergence in loopy networks. Sufficient convergence conditions for Gaussian BP have been developed in [1, 12–14] when the underlying Gaussian distribution is expressed in terms of pairwise connections between scalar variables (also known as Markov random field (MRF)). These works focus on the convergence analysis

of Gaussian BP for computing the marginal distribution of a joint distribution with pairwise factors. However, the iterative equations for Gaussian BP on MRFs are different from that for distributed estimation problems such as in [2–9,15], where high order factors (non-pairwise) and vector-valued variables are involved. Therefore, these existing conditions and analysis methods are not applicable to distributed estimation problems. In this paper, we study the convergence analysis of Gaussian BP for distributed parameter estimation focusing on the convergence of message information matrix. We show analytically that, with arbitrary positive semidefinite matrix initialization, the message information matrix being exchanged among nodes converges and its distance to the unique positive definite limit matrix decreases exponentially.

Note that distributed estimation based on the consensus+ innovations philosophy proposed in [16, 17] (see also the related family of diffusion algorithms [18]) converges to the optimal centralized estimator under the assumption of global observability of the (aggregate) sensing model and connectivity 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 and the former could be arbitrary with cycles (as long as it is connected). The results in [16, 17] imply that the unknown variables **x** can be reconstructed completely at each node in the network. For large-scale networks with high dimensional **x**, it may be impractical to reconstruct **x** at every node. In [19, section 3.4], the author developed approaches to address this problem, where each node can reconstruct a set of unknown variables that should be larger than the set of variables that influence its local measurement. This paper studies a different distributed estimation problem when each node estimates only its own unknown variables under pairwise independence condition of the unknown variables; this leads to lower dimensional data exchanges between neighbors.

2. COMPUTATION MODEL

Consider a general connected network of M nodes, with $\mathcal{V} = \{1, \ldots, M\}$ denoting the set of nodes, and $\mathcal{E}_{\text{Net}} \subset \mathcal{V} \times \mathcal{V}$ as

the set of all undirect communication links in the network, i.e., if i and j are within the communication range, $(i,j) \in \mathcal{E}_{\text{Net}}$. At every node $n \in \mathcal{V}$, the local observations are in the form of $\mathbf{y}_n = \sum_{i \in n \cup \mathcal{I}(n)} \mathbf{A}_{n,i}\mathbf{x}_i + \mathbf{z}_n$, where $\mathcal{I}(n)$ denotes the set of direct neighbors of node n (i.e., all nodes 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 node i with dimension $N_i \times 1$, and with the prior distribution $p(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{x}_i|\mathbf{0}, \mathbf{W}_i)$, and $\mathbf{z}_n$ is the additive noise with distribution $\mathbf{z}_n \sim \mathcal{N}(\mathbf{z}_n|\mathbf{0}, \mathbf{R}_n)$. 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 $i \neq j$. The goal is to estimate $\mathbf{x}_i$, based on $\mathbf{y}_n$, $p(\mathbf{x}_i)$ and $p(\mathbf{z}_n)$.

The Gaussian BP algorithm can be derived over the corresponding factor graph to compute the estimate of $\mathbf{x}_n$ for all $n \in \mathcal{V}$ [20]. It involves two kinds of messages: One is the message from a variable node $\mathbf{x}_j$ to its neighboring factor node f_n, defined as

$$m_{j \to f_n}^{(\ell)}(\mathbf{x}_j) = p(\mathbf{x}_j) \prod_{f_k \in \mathcal{B}(j) \setminus f_n} m_{f_k \to j}^{(\ell-1)}(\mathbf{x}_j), \quad (1)$$

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

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

$$(2)$$

where $\mathcal{B}(f_n)$ denotes the set of neighboring variable nodes of f_n. The process iterates between equations (1) and (2). At each iteration l, the approximate marginal distribution, also named belief, on $\mathbf{x}_i$ is computed locally at $\mathbf{x}_i$ as

$$b_{\text{BP}}^{(\ell)}(\mathbf{x}_i) = p(\mathbf{x}_i) \prod_{f_n \in \mathcal{B}(i)} m_{f_n \to i}^{(\ell)}(\mathbf{x}_i). \quad (3)$$

It can be shown [20] that the general expression for the message from variable node to factor node is

$$m_{j \to f_n}^{(\ell)}(\mathbf{x}_j) \propto \exp\left\{ -\frac{1}{2}\|\mathbf{x}_j - \mathbf{v}_{j \to f_n}^{(\ell)}\|_{\mathbf{C}_{j \to f_n}^{(\ell)}} \right\}, \quad (4)$$

where $\mathbf{C}_{j \to f_n}^{(\ell)}$ and $\mathbf{v}_{j \to f_n}^{(\ell)}$ are the message covariance matrix and mean vector received at variable node j at the l-th iteration, with

$$\left[\mathbf{C}_{j \to f_n}^{(\ell)}\right]^{-1} = \mathbf{W}_j^{-1} + \sum_{f_k \in \mathcal{B}(j) \setminus f_n} \left[\mathbf{C}_{f_k \to j}^{(\ell-1)}\right]^{-1}. \quad (5)$$

Furthermore, the message from factor node to variable node is given by [20]

$$m_{f_n \to i}^{(\ell)}(\mathbf{x}_i) \propto \exp\left\{ -\frac{1}{2}\|\mathbf{x}_i - \mathbf{v}_{f_n \to i}^{(\ell)}\|_{\mathbf{C}_{f_n \to i}^{(\ell)}} \right\}, \quad (6)$$

where $\mathbf{C}_{f_k \to j}^{(\ell-1)}$ and $\mathbf{v}_{f_k \to j}^{(\ell-1)}$ are the message covariance matrix and mean vector received at variable node j at the $l-1$ iteration with

$$\left[\mathbf{C}_{f_n \to i}^{(\ell)}\right]^{-1} = \mathbf{A}_{n,i}^T \left[\mathbf{R}_n + \sum_{j \in \mathcal{B}(f_n) \setminus i} \mathbf{A}_{n,j}\mathbf{C}_{j \to f_n}^{(\ell)}\mathbf{A}_{n,j}^T\right]^{-1}\mathbf{A}_{n,i}. \quad (7)$$

The following lemma shown in [20] indicates that setting the initial message covariances $[\mathbf{C}_{f_n \to i}^{(0)}]^{-1} \succeq \mathbf{0}$ for all $n, i \in \mathcal{V}$ guarantees $[\mathbf{C}_{j \to f_n}^{(\ell)}]^{-1} \succ \mathbf{0}$ for $l \geq 1$.

Lemma 1. *Let the initial messages at factor node f_k be in Gaussian function forms with covariance $[\mathbf{C}_{f_k \to j}^{(0)}]^{-1} \succeq \mathbf{0}$ for all $k \in \mathcal{V}$ and $j \in \mathcal{B}(f_k)$. Then $[\mathbf{C}_{j \to f_n}^{(\ell)}]^{-1} \succ \mathbf{0}$ and $[\mathbf{C}_{f_k \to j}^{(\ell)}]^{-1} \succ \mathbf{0}$ for all $l \geq 1$ with $j \in \mathcal{V}$ and $f_n, f_k \in \mathcal{B}(j)$. Furthermore, in this case, all the messages $m_{j \to f_n}^{(\ell)}(\mathbf{x}_j)$ and $m_{f_k \to j}^{(\ell)}(\mathbf{x}_i)$ exist and are in Gaussian form.*

For this factor graph based approach, according to the message updating procedure (4) and (6), message exchange is only needed between neighboring nodes. For example, the messages transmitted from node n to its neighboring node i are $m_{f_n \to i}^{(\ell)}(\mathbf{x}_i)$ and $m_{n \to f_i}^{(\ell)}(\mathbf{x}_n)$. Thus, the message passing scheme given in (1) and (2) automatically conforms with the network topology. Furthermore, if the messages $m_{j \to f_n}^{(\ell)}(\mathbf{x}_j)$ and $m_{f_n \to i}^{(\ell)}(\mathbf{x}_i)$ exist for all l (which can be achieved using Lemma 1), 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 BP messages exist, according to the definition of belief in (3), $b_{\text{BP}}^{(\ell)}(\mathbf{x}_i)$ at iteration l is computed as [20]

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

$$(8)$$

with $\mathbf{P}_i^{(\ell)} = \left[\mathbf{W}_i^{-1} + \sum_{f_n \in \mathcal{B}(i)} \left[\mathbf{C}_{f_n \to i}^{(\ell)}\right]^{-1}\right]^{-1}$, and $\boldsymbol{\mu}_i^{(\ell)} = \mathbf{P}_i^{(\ell)} \left[\sum_{f_n \in \mathcal{B}(i)} \left[\mathbf{C}_{f_n \to i}^{(\ell)}\right]^{-1}\mathbf{v}_{f_n \to i}^{(\ell)}\right]$. The iterative computation terminates when message (4) or message (6) converges to a fixed value or the maximum number of iterations is reached.

3. CONVERGENCE OF INFORMATION MATRICES

The challenge of deploying the BP algorithm for large-scale networks is determining whether it will converge. In particular, it is generally known that if the factor graph contains cycles, the BP algorithm may diverge. Thus, determining convergence conditions for the BP algorithm is very important. Sufficient conditions for the convergence of Gaussian BP with scalar variable in loopy graphs are available in [1, 12, 14]. However, they are derived based on pairwise graphs with local functions that only involve two variables. This is in sharp

contrast to the model considered in this paper, where the f_n involves high-order interactions between vector variables, and thus the convergence results in [1, 12, 14] cannot be applied to the factor graph based vector-form Gaussian BP.

Due to the recursively updating property of $m_{j \to f_n}^{(\ell)}(\mathbf{x}_j)$ and $m_{f_n \to i}^{(\ell)}(\mathbf{x}_i)$ in (4) and (6), the message evolution can be simplified by combining these two kinds of messages into one. By substituting $[\mathbf{C}_{j \to f_n}^{(\ell)}]^{-1}$ in (5) into (7), the updating of the message covariance matrix inverse, named message information matrix in the following, can be denoted as

$$
\begin{aligned}
[\mathbf{C}_{f_n \to i}^{(\ell)}]^{-1} &= \mathbf{A}_{n,i}^T \Big[\mathbf{R}_n + \sum_{j \in \mathcal{B}(f_n) \backslash i} \mathbf{A}_{n,j} \big[\mathbf{W}_j^{-1} \\
&\quad + \sum_{f_k \in \mathcal{B}(j) \backslash f_n} [\mathbf{C}_{f_k \to j}^{(\ell-1)}]^{-1} \big]^{-1} \mathbf{A}_{n,j}^T \Big]^{-1} \mathbf{A}_{n,i} \\
&\triangleq \mathcal{F}_{n \to i} \big(\big\{ [\mathbf{C}_{f_k \to j}^{(\ell-1)}]^{-1} \big\}_{(f_k, j) \in \widetilde{\mathcal{B}}(f_n, i)} \big),
\end{aligned}
$$

where $\widetilde{\mathcal{B}}(f_n, i) = \{(f_k, j) | j \in \mathcal{B}(f_n) \backslash i, f_k \in \mathcal{B}(j) \backslash f_n\}$. Observing that $\mathbf{C}_{f_n \to i}^{(\ell)}$ in (9) is independent of $\mathbf{v}_{j \to f_n}^{(\ell)}$ and $\mathbf{v}_{f_n \to i}^{(\ell)}$ in (4) and (6), we can focus on the convergence property of $[\mathbf{C}_{f_n \to i}^{(\ell)}]^{-1}$ alone.

To consider the updates of all message information matrices, we introduce the following definitions. Let $\mathbf{C}^{(\ell-1)} \triangleq \mathtt{Bdiag}(\{[\mathbf{C}_{f_n \to i}^{(\ell-1)}]^{-1}\}_{n \in \mathcal{V}, i \in \mathcal{B}(f_n)})$ be a block diagonal matrix with diagonal blocks being the message information matrices in the network at time $l-1$ with index arranged in ascending order first on n and then on i. Using the definition of $\mathbf{C}^{(\ell-1)}$, the term $\sum_{f_k \in \mathcal{B}(j) \backslash f_n} [\mathbf{C}_{f_k \to j}^{(\ell-1)}]^{-1}$ in (9) can be written as $\mathbf{\Xi}_{n,j} \mathbf{C}^{(\ell-1)} \mathbf{\Xi}_{n,j}^T$, where $\mathbf{\Xi}_{n,j}$ is for selecting appropriate components from $\mathbf{C}^{(\ell-1)}$ to form the summation. Further, define $\mathbf{H}_{n,i} = [\{\mathbf{A}_{n,j}\}_{j \in \mathcal{B}(f_n) \backslash i}]$, $\mathbf{\Psi}_{n,i} = \mathtt{Bdiag}(\{\mathbf{W}_j^{-1}\}_{j \in \mathcal{B}(f_n) \backslash i})$ and $\mathbf{K}_{n,i} = \mathtt{Bdiag}(\{\mathbf{\Xi}_{n,j}\}_{j \in \mathcal{B}(f_n) \backslash i})$, all with component blocks arranged with ascending order on j. Then (9) can be written as

$$
\begin{aligned}
[\mathbf{C}_{f_n \to i}^{(\ell)}]^{-1} &= \mathbf{A}_{n,i}^T \big\{ \mathbf{R}_n + \mathbf{H}_{n,i} [\mathbf{\Psi}_{n,i} + \mathbf{K}_{n,i} (\mathbf{I}_{|\mathcal{B}(f_n)|-1} \\
&\quad \otimes \mathbf{C}^{(\ell-1)}) \mathbf{K}_{n,i}^T]^{-1} \mathbf{H}_{n,i}^T \big\}^{-1} \mathbf{A}_{n,i}.
\end{aligned} \tag{9}
$$

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 $\mathbf{C}^{(\ell)} = \mathcal{F}(\mathbf{C}^{(\ell-1)})$. Then, by stacking $[\mathbf{C}_{f_n \to i}^{(\ell)}]^{-1}$ on the left side of (9) for all n and i as the block diagonal matrix $\mathbf{C}^{(\ell)}$, we obtain

$$
\begin{aligned}
\mathbf{C}^{(\ell)} &= \mathbf{A}^T \big\{ \mathbf{\Omega} + \mathbf{H}[\mathbf{\Psi} + \mathbf{K}(\mathbf{I}_\varphi \otimes \mathbf{C}^{(\ell-1)}) \mathbf{K}^T]^{-1} \mathbf{H}^T \big\}^{-1} \mathbf{A}, \\
&\triangleq \mathcal{F}(\mathbf{C}^{(\ell-1)}),
\end{aligned} \tag{10}
$$

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 $[\mathbf{C}_{f_n \to i}^{(\ell)}]^{-1}$ in $\mathbf{C}^{(\ell)}$). 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 properties of the updating operator $\mathcal{F}(\cdot)$, with the proof given in [20].

Property 1. *The updating operator $\mathcal{F}(\cdot)$ satisfies the following properties:*

P 1.1: $\mathcal{F}(\mathbf{C}^{(\ell)}) \succeq \mathcal{F}(\mathbf{C}^{(\ell-1)})$, if $\mathbf{C}^{(\ell)} \succeq \mathbf{C}^{(\ell-1)} \succeq \mathbf{0}$.
P 1.2: $\alpha \mathcal{F}(\mathbf{C}^{(\ell)}) \succ \mathcal{F}(\alpha \mathbf{C}^{(\ell)})$ and $\mathcal{F}(\alpha^{-1} \mathbf{C}^{(\ell)}) \succ \alpha^{-1} \mathcal{F}(\mathbf{C}^{(\ell)})$, if $\mathbf{C}^{(\ell)} \succ \mathbf{0}$ and $\alpha > 1$.
P 1.3: Define $\mathbf{U} \triangleq \mathbf{A}^T \mathbf{\Omega}^{-1} \mathbf{A}$ and $\mathbf{L} \triangleq \mathbf{A}^T [\mathbf{\Omega} + \mathbf{H}\mathbf{\Psi}^{-1}\mathbf{H}^T]^{-1} \mathbf{A}$. With arbitrary $\mathbf{C}^{(0)} \succeq \mathbf{0}$, $\mathcal{F}(\mathbf{C}^{(\ell)})$ is bounded by $\mathbf{U} \succeq \mathcal{F}(\mathbf{C}^{(\ell)}) \succeq \mathbf{L} \succ \mathbf{0}$ for $l \geq 1$.

In this paper, $\mathbf{X} \succeq \mathbf{Y}$ ($\mathbf{X} \succ \mathbf{Y}$) means that $\mathbf{X} - \mathbf{Y}$ is positive semidefinite (definite). Based on the above properties of $\mathcal{F}(\cdot)$, we can establish the convergence property for the information matrices. The following theorem establishes that there exists a unique fixed point for the mapping $\mathcal{F}(\cdot)$. The proof is omitted due to space restrictions; it is provided in [20].

Theorem 1. *With $\mathbf{C}^{(0)} \succeq \mathbf{0}$, there exists a unique positive definite fixed point for the mapping $\mathcal{F}(\cdot)$.*

Lemma 1 states that with arbitrary positive semidefinite (p.s.d.) initial message information matrices, the message information matrices will be kept as positive definite (p.d.) at every iteration. On the other hand, Theorem 1 indicates that there exists a unique fixed point for the mapping $\mathcal{F}$. Next, we will show that with arbitrary initial value $\mathbf{C}^{(0)} \succeq 0$, $\mathbf{C}^{(\ell)}$ converges to a unique p.d. matrix.

Theorem 2. *The matrix sequence $\{\mathbf{C}^{(\ell)}\}_{l=0,1,\ldots}$ defined by (10) converges to a unique positive definite matrix for any initial covariance matrix $\mathbf{C}^{(0)} \succeq \mathbf{0}$.*

Proof. With arbitrary initial value $\mathbf{C}^{(0)} \succeq \mathbf{0}$, following P 1.3, we have $\mathbf{U} \succeq \mathbf{C}^{(1)} \succeq \mathbf{L} \succ \mathbf{0}$. On the other hand, according to Theorem 1, (10) has a unique fixed point $\mathbf{C}^* \succ \mathbf{0}$. Notice that we can always choose a scalar $\alpha > 1$ such that

$$
\alpha \mathbf{C}^* \succeq \mathbf{C}^{(1)} \succeq \mathbf{L}. \tag{11}
$$

Applying $\mathcal{F}(\cdot)$ to (11) l times, and using P 1.1, we have

$$
\mathcal{F}^l(\alpha \mathbf{C}^*) \succeq \mathcal{F}^{l+1}(\mathbf{C}^{(0)}) \succeq \mathcal{F}^l(\mathbf{L}), \tag{12}
$$

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

We start from the left inequality in (12). Following the fixed point definition, $\alpha \mathbf{C}^* = \alpha \mathcal{F}(\mathbf{C}^*)$. Then, according to P 1.2, $\alpha \mathbf{C}^* \succ \mathcal{F}(\alpha \mathbf{C}^*)$. Applying $\mathcal{F}$ again gives $\mathcal{F}(\alpha \mathbf{C}^*) \succ \mathcal{F}^2(\alpha \mathbf{C}^*)$. Applying $\mathcal{F}(\cdot)$ repeatedly, we can obtain $\mathcal{F}^2(\alpha \mathbf{C}^*) \succ \mathcal{F}^3(\alpha \mathbf{C}^*) \succ \mathcal{F}^4(\alpha \mathbf{C}^*)$, etc. Thus $\mathcal{F}^l(\alpha \mathbf{C}^*)$ is a decreasing 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 $\mathbf{L}$, $\mathcal{F}^l(\alpha\mathbf{C}^*)$ is convergent. Finally, since there exists only one fixed point for $\mathcal{F}(\cdot)$, $\lim_{l\to\infty}\mathcal{F}^l(\alpha\mathbf{C}^*) = \mathbf{C}^*$. On the other hand, for the right hand side of (12), as $\mathcal{F}(\cdot) \succeq \mathbf{L}$, we have $\mathcal{F}(\mathbf{L}) \succeq \mathbf{L}$. Applying $\mathcal{F}$ repeatedly gives $\mathcal{F}^2(\mathbf{L}) \succeq \mathcal{F}(\mathbf{L})$, $\mathcal{F}^3(\mathbf{L}) \succeq \mathcal{F}^2(\mathbf{L})$, etc. So, $\mathcal{F}^l(\mathbf{L})$ is an increasing sequence (with respect to the partial order induced by the cone of p.s.d. matrices). Since $\mathcal{F}(\cdot)$ is upper bounded by $\mathbf{U}$, $\mathcal{F}^l(\mathbf{L})$ is a convergent sequence. Again due to the unique fixed point, we have $\lim_{l\to\infty}\mathcal{F}^l(\mathbf{L}) = \mathbf{C}^*$. Finally, taking the limit with respect to l on (12) we have $\lim_{l\to\infty}\mathcal{F}^l(\mathbf{C}^{(0)}) = \mathbf{C}^*$, for arbitrary initial $\mathbf{C}^{(0)} \succeq \mathbf{0}$. $\qquad\square$

According to Theorem 2, the covariance matrix $\mathbf{C}_{f_n\to i}^{(\ell)}$ converges if all initial information matrices are p.s.d., i.e., $\left[\mathbf{C}_{f_n\to i}^{(0)}\right]^{-1} \succeq \mathbf{0}$ for all $i \in \mathcal{V}$ and $f_n \in \mathcal{B}(i)$. Notice that, for the pairwise model, the information matrix does not necessarily converge for all initial non-negative value (in the scalar variable case) as shown in [12,13]. Moreover, due to the computation of $[\mathbf{C}_{f_n\to i}^{(\ell)}]^{-1}$ being independent of the local observations $\mathbf{y}_n$, as long as the network topology does not change, the converged value $[\mathbf{C}_{f_n\to i}^*]^{-1}$ can be precomputed offline and stored at each node, and there is no need to re-compute $[\mathbf{C}_{f_n\to i}^*]^{-1}$ even if $\mathbf{y}_n$ 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 the part metric [21], in the following, we start by introducing the part metric.

Definition 1. *Part (Birkhoff) Metric [21]: For arbitrary matrices X and Y with the same dimension, if there exists $\alpha \geq 1$ such that $\alpha X \succeq Y \succeq \alpha^{-1}X$, X and Y are called the parts, and $d(X,Y) \triangleq \inf\{\log\alpha : \alpha X \succeq Y \succeq \alpha^{-1}X, \alpha \geq 1\}$ defines a metric called the part metric.*

Next, we will show that $\{\mathbf{C}^{(\ell)}\}_{l=1,\dots}$ converges at a geometric rate with respect to the part metric in $\mathcal{C}$, which is constructed as

$$\mathcal{C} = \{\mathbf{C}^{(\ell)}|\mathbf{U} \succeq \mathbf{C}^{(\ell)} \succeq \mathbf{C}^*+\epsilon\mathbf{I}\}\cup\{\mathbf{C}^{(\ell)}|\mathbf{C}^*-\epsilon\mathbf{I} \succeq \mathbf{C}^{(\ell)} \succeq \mathbf{L}\},$$

where $\epsilon > 0$ is a scalar and can be arbitrarily small.

Theorem 3. *With the initial covariance matrix set to be an arbitrary p.s.d. matrix, i.e., $[\mathbf{C}_{f_n\to i}^{(0)}]^{-1} \succeq \mathbf{0}$, the sequence $\{\mathbf{C}^{(\ell)}\}_{l=0,1,\dots}$ converges at a geometric rate with respect to the part metric in $\mathcal{C}$.*

Proof. Consider two matrices $\mathbf{C}^{(\ell)} \in \mathcal{C}$, and $\mathbf{C}^* \notin \mathcal{C}$, according to Definition 1, we have $d(\mathbf{C}^{(\ell)}, \mathbf{C}^*) \triangleq \inf\{\log\alpha : \alpha\mathbf{C}^{(\ell)} \succeq \mathbf{C}^* \succeq \alpha^{-1}\mathbf{C}^{(\ell)}\}$. Since $d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)$ is the smallest number satisfying $\alpha\mathbf{C}^{(\ell)} \succeq \mathbf{C}^* \succeq \alpha^{-1}\mathbf{C}^{(\ell)}$, this is equivalent to

$$\exp\{d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\}\mathbf{C}^{(\ell)} \succeq \mathbf{C}^* \succeq \exp\{-d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\}\mathbf{C}^{(\ell)}. \tag{13}$$

Applying P 1.1 to (13), we have $\exp\{d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\}\mathcal{F}(\mathbf{C}^{(\ell)}) \succeq \mathcal{F}(\mathbf{C}^*) \succeq \exp\{-d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\}\mathcal{F}(\mathbf{C}^{(\ell)})$. Then applying P 1.2 and considering that $\exp\{d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\} > 1$ and $\exp\{-d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\} < 1$, we obtain

$$\begin{aligned}&\exp\{d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\}\mathcal{F}(\mathbf{C}^{(\ell)})\\&\quad\succ \mathcal{F}(\mathbf{C}^*) \succ \exp\{-d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\}\mathcal{F}(\mathbf{C}^{(\ell)}).\end{aligned} \tag{14}$$

Notice that, for arbitrary p.d. matrices $\mathbf{X}$ and $\mathbf{Y}$, if $\mathbf{X}-k\mathbf{Y} \succ \mathbf{0}$ then, by definition that, we have $\mathbf{x}^T\mathbf{X}\mathbf{x} - k\mathbf{x}^T\mathbf{Y}\mathbf{x} > 0$. 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 $\triangle d > 0$ such that

$$\begin{aligned}&\exp\{-\triangle d + d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\}\mathcal{F}(\mathbf{C}^{(\ell)})\\&\quad\succ \mathcal{F}(\mathbf{C}^*) \succ \exp\{\triangle d - d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\}\mathcal{F}(\mathbf{C}^{(\ell)}).\end{aligned} \tag{15}$$

Now, using the definition of part metric, (15) is equivalent to

$$-\triangle d + d(\mathbf{C}^{(\ell)}, \mathbf{C}^*) \geq d(\mathcal{F}(\mathbf{C}^{(\ell)}), \mathcal{F}(\mathbf{C}^*)). \tag{16}$$

Hence, we obtain $d(\mathcal{F}(\mathbf{C}^{(\ell)}), \mathcal{F}(\mathbf{C}^*)) < d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)$. This result holds for any $\mathbf{C}^{(\ell)} \in \mathcal{C}$, $d(\mathcal{F}(\mathbf{C}^{(\ell)}), \mathcal{F}(\mathbf{C}^*)) < cd(\mathbf{C}^{(\ell)}, \mathbf{C}^*)$, where $c = \sup_{\mathbf{C}^l\in\mathcal{C}} \frac{d(\mathcal{F}(\mathbf{C}^l), \mathcal{F}(\mathbf{C}^*))}{d(\mathbf{C}^l, \mathbf{C}^*)} < 1$. Consequently, we have $d(\mathbf{C}^{(\ell)}, \mathbf{C}^*) < c^l d(\mathbf{C}^{(0)}, \mathbf{C}^*)$. Thus the sequence $\{\mathbf{C}^{(\ell)}\}_{l=1,\dots}$ converges at a geometric rate with respect to the part metric. $\qquad\square$

It is useful to have an estimate of the convergence rate of $\mathbf{C}^{(\ell)}$ in terms of the more standard induced matrix norms. According to [22, Lemma 2.3], the convergence rate of $||\mathbf{C}^{(0)} - \mathbf{C}^*||$ is dominated by that of $d(\mathbf{C}^{(0)}, \mathbf{C}^*)$, where $||\cdot||$ is a monotone norm defined on the p.s.d. cone, with $||\cdot||_2$ and $||\cdot||_F$ being examples of such matrix norms [23, 2.2-10]. More specifically,

$$\begin{aligned}&(2\exp\{d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\} - \exp\{-d(\mathbf{C}^{(\ell)}, \mathbf{C}^*)\} - 1)\\&\quad\times\min\{||\mathbf{C}^{(\ell)}||, ||\mathbf{C}^*||\} \geq ||\mathbf{C}^{(\ell)} - \mathbf{C}^*||.\end{aligned} \tag{17}$$

The physical meaning of Theorem 3 is that the sequence $\{\mathbf{C}^{(\ell)}\}_{l=1,\dots}$ converges at a geometric rate (the distance between $\mathbf{C}^{(\ell)}$ and $\mathbf{C}^*$ decreases exponentially) before $\mathbf{C}^{(\ell)}$ enters $\mathbf{C}^*$'s neighborhood, which can be chosen arbitrarily small.

4. CONCLUSION

This paper has established the convergence of the exchanged message information matrix of Gaussian belief propagation (BP) for distributed estimation. We have shown analytically that, with arbitrary positive semidefinite initial value, the information matrix converges to a unique positive definite matrix at geometric rate. The convergence guaranteed property and fast convergence rate of the message information matrix pave the way for the convergence analysis of the Gaussian BP message mean vector.

5. REFERENCES

[1] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *Neural Computation*, vol. 13, no. 10, pp. 2173–2200, Mar. 2001.

[2] Y. Hu, A. Kuh, T. Yang, and A. Kavcic, "A belief propagation based power distribution system state estimator," *IEEE Comput. Intell. Mag.*, vol. 6, no. 3, pp. 36–46, Aug 2011.

[3] B. L. Ng, J. Evans, S. Hanly, and D. Aktas, "Distributed downlink beamforming with cooperative base stations," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5491–5499, Dec 2008.

[4] 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.*, vol. 12, no. 11, pp. 5908–5917, Nov 2013.

[5] ——, "Network-wide distributed carrier frequency offsets estimation and compensation via belief propagation," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5868–5877, 2013.

[6] O. Shental, P. Siegel, J. 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)*, July 2008, pp. 1863–1867.

[7] G. Zhang, W. Xu, and Y. Wang, "Fast distributed rate control algorithm with QoS support in ad-hoc," in *2010 IEEE Global Telecommunications Conference (GLOBECOM 2010)*.

[8] B. J. Frey, "Local probability propagation for factor analysis," in *Neural Information Processing Systems (NIPS)*, Dec 1999, pp. 442–448.

[9] X. Tan and J. Li, "Computationally efficient sparse Bayesian learning via belief propagation," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2010–2021, April 2010.

[10] 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*, vol. 1, no. 2, pp. 93–103, 2008.

[11] Y. Weiss and W. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 736–744, Feb. 2001.

[12] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Walk-sums and belief propagation in Gaussian graphical models," *Journal of Machine Learning Research*, vol. 7, no. 2, pp. 2031–2064, Feb. 2006.

[13] C. C. Moallemi and B. V. Roy, "Convergence of min-sum message passing for quadratic optimization," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2413–2423, 2009.

[14] Q. Su and Y.-C. Wu, "On convergence conditions of Gaussian belief propagation," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1144–1155, March 2015.

[15] F. Lehmann, "Iterative mitigation of intercell interference in cellular networks based on Gaussian belief propagation," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2544–2558, July 2012.

[16] S. Kar and J. M. F. Moura, "Consensus+innovations distributed inference over networks: cooperation and sensing in networked systems," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 99–109, 2013.

[17] S. Kar, J. M. F. Moura, and H. Poor, "Distributed linear parameter estimation: asymptotically efficient adaptive strategies," *SIAM Journal on Control and Optimization*, vol. 51, no. 3, pp. 2200–2229, 2013.

[18] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, 2010.

[19] S. Kar, "Large scale networked dynamical systems: Distributed inference," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, Department of Electrical and Computer Engineering, June 2010.

[20] J. Du, S. Ma, Y.-C. Wu, S. Kar, and J. M. F. Moura, "Convergence analysis of distributed inference with vector-valued Gaussian belief propagation," submitted to *Journal of Machine Learning Research* [Preprint Available]: https://users.ece.cmu.edu/~soummyak/GBP_convergence.

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

[22] U. Krause and R. Nussbaum, "A limit set trichotomy for self-mappings of normal cones in Banach spaces," *Nonlinear Analysis, Theory, Methods&Applications*, vol. 20, no. 7, pp. 855–870, 1993.

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