Distributed Inference with Variational Message Passing in Gaussian Graphical Models: Trade-offs in Message Schedules and Convergence Conditions

Bin Li, Nan Wu, and Yik-Chung Wu

Abstract

Message passing algorithms on graphical models offer a low-complexity and distributed paradigm for performing marginalization from a high-dimensional distribution. However, the convergence behaviors of message passing algorithms can be heavily affected by the adopted message update schedule. In this paper, we focus on the variational message passing (VMP) applied to Gaussian graphical models and its convergence under different schedules is analyzed. In particular, based on the update equations of VMP under the mean-field assumption, we prove that the mean vectors obtained from VMP are the exact marginal mean vectors under any valid message passing schedule, giving the legitimacy of using VMP in Gaussian graphical models. Furthermore, three categories of valid message passing schedules, namely serial schedule, parallel schedule and randomized schedule are considered for VMP update. In the basic serial schedule, VMP unconditionally converges, but could be slow in large-scale distributed networks. To speed up the serial schedule, a group serial schedule is proposed while guaranteeing the VMP convergence. On the other hand, parallel schedule and its damped variant are applied to accelerate VMP, where the necessary and sufficient convergence conditions are derived. To allow nodes with different local computation resources to compute messages more flexibly and efficiently, a randomized schedule is proposed for VMP update, and the probabilistic necessary and sufficient convergence conditions are

Bin Li and Nan Wu are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China, and also with the Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing, Zhejiang 314000, China (e-mail: binli@bit.edu.cn; wunan@bit.edu.cn).

Yik-Chung Wu is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China (e-mail: ycwu@eee.hku.hk).

This work was supported in part by the National Natural Science Foundation of China with Grant No. 62371045.

presented. Finally, numerical results and applications are presented to illustrate the trade-offs in the ease and speed of convergence.

Index Terms

Variational message passing, distributed inference, message schedule, convergence analysis, Gaussian graphical models

I. INTRODUCTION

Computing the means of marginal distributions from a high-dimensional Gaussian distribution is an essential task in many distributed inference applications [1]–[4]. For example, in distributed peer-to-peer rating on d items (such as movies, goods and services) over social networks, the final rating $\mathbf{x}_i \in \mathbb{R}^d$ of user i after incorporating its own initial rating $\mathbf{y}_i \in \mathbb{R}^d$ and those of other users can be found by solving the optimization problem [1]

$$\min_{\{\mathbf{x}_i\}_{i=1}^n} \quad \sum_{i=1}^{n} \alpha_i \|\mathbf{x}_i - \mathbf{y}_i\|_2^2 + \sum_{k>i} \omega_{ik} \|\mathbf{x}_i - \mathbf{x}_k\|_2^2, \tag{1}$$

where $\alpha_i \geq 0$ denotes the confidence of the initial rating, and $\omega_{ik} \geq 0$ denotes the closeness between user i and user j. Since the objective function of (1) is quadratic with respect to $\{\mathbf{x}_i\}_{i=1}^n$, the solution of (1) can be obtained by computing the marginal means of the high-dimensional Gaussian distribution

$$p(\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{J}\mathbf{x} + \mathbf{h}^T\mathbf{x}\right\}, \tag{2}$$

where $\mathbf{x} \triangleq [\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_n^T]^T$, $\mathbf{J}$ is a matrix composing of sub-blocks given by $\mathbf{J}_{ii} = (\alpha_i + \sum_{k>i} \omega_{ik})\mathbf{I}$ and $\mathbf{J}_{ik} = -\omega_{ik}\mathbf{I}$ for $i \neq k$, and $\mathbf{h} = [\alpha_1\mathbf{y}_1^T, \alpha_2\mathbf{y}_2^T, \cdots, \alpha_n\mathbf{y}_n^T]^T$. Other applications having the formulations in the form of (1) include consensus propagation [2], distributed quadratic function optimization [3], and network synchronization [4].

On the other hand, many distributed minimum mean square error (MMSE) or linear MMSE estimation problems can be interpreted as finding the marginal means of a high-dimensional Gaussian distribution with the form of (2). For example, in downlink transmission with distributed beamforming [5], the transmitted signal from all base stations has the form $\mathbf{x} = K\mathbf{H}^T(\mathbf{H}\mathbf{H}^T + \beta\mathbf{I})^{-1}\mathbf{s}$, where $\mathbf{s} \triangleq [s_1, s_2, \cdots, s_N]^T$ with s_i denoting the symbol intended for user i, $\mathbf{H}$ is the downlink channel matrix from all base stations to all users, and K, β are parameters depending on the optimality criterion for transmit beamforming design. By using Woodbury matrix identity,

it can be proved that computing the symbol vector to be transmitted is equivalent to computing the mean of (2) when $\mathbf{J} = \mathbf{H}^T\mathbf{H} + \beta\mathbf{I}$ and $\mathbf{h} = K\mathbf{H}^T\mathbf{s}$ (details given in Appendix A). Similar settings also appear in distributed data detection in multi-user multi-input multi-output systems [6] and cloud radio access networks [7], or as a key subtask in many distributed nonlinear inference problems, including state estimation in electric power systems [8], and cooperative localization [9].

For a high-dimensional Gaussian distribution, the exact means of the marginal distributions can be obtained by the direct matrix inversion method. However, with the computational complexity of matrix inversion being cubic in the number of variables, the direct matrix inversion method is computationally costly in large-scale problems. Moreover, the direct matrix inversion method requires gathering the information matrix and potential vector of a high-dimensional Gaussian distribution. Thus it is not scalable for data distributed in large-scale networks, as in the above mentioned applications [1]–[9].

Due to their distributed nature and low complexity, message passing algorithms are widely used for approximate Bayesian inference on graphical models. Gaussian belief propagation (BP) [10] is a popular message passing algorithm for computing the marginal means of a high-dimensional Gaussian distribution. Other options include variational inference (VI) [11], expectation propagation (EP) [12], and approximate message passing (AMP) [13]–[15]. While there are many different types of message passing algorithms, a common challenge is that the convergence behavior is difficult to analyze and may be restrictive in the class of convergent models. In particular, the convergence analysis of Gaussian BP has been analyzed in [3], [10], [16]–[18], in which conditions such as walk-sumability or pairwise-normalizability are required. For AMP, although it is applicable to a wide range of prior distributions, its convergence typically requires the transform matrix in the linear equation containing independent and identically distributed (i.i.d.) elements [19]. Recent more advanced variants of AMP, namely orthogonal AMP [20], vector AMP [21] and memory AMP [22], [23] expand the class of transform matrix to right-unitarily-invariant matrices [22]–[24]. However, AMP-type algorithms are still far from being applicable to arbitrary transform matrix.

An apparent exception is the variational message passing (VMP) [25], [26], which is based on VI and has a simple message update rule. In general VI, instead of taking the intractable integrals involved in computing the exact marginal distributions, a variational distribution within a given family distribution is used to approximate the target distribution, where the Kullback-

Leibler (KL) divergence is used to measure the closeness of the variational distribution and the target distribution [27]. By minimizing this KL divergence under the mean-field assumption, the optimal variational marginal distributions can be obtained under a cyclic block update. Due to the sequential block update, VI is guaranteed to converge in such update schedule [28], which makes VI widely used in computing approximate marginal distributions of a complex joint distribution in many applications, including cooperative localization [29], [30], tensor decomposition [31], [32], channel estimation and data detection [33], [34].

However, if cyclic coordinate update is applied to a distributed network, this means only one node can be updated at a time. Obviously, this would lead to slow convergence and introduce long latency in distributed and large-scale networks. To speed up the convergence, parallel schedule is adopted for VMP in the context of sensors' self-localization [29] and distributed receiver design in extra-large scale MIMO systems [34]. However, the convergence of VMP under parallel schedule is no longer guaranteed. One may wonder besides the sequential update (which is slow but convergence guaranteed) and parallel update (which is fast but may diverge, making the computation useless), is there other message update schedule that is between these two extreme cases? As the first endeavor to rigorously study VMP schedules and its implication, this paper focuses on the Gaussian graphical models, which find many practical applications in its own right.

In particular, we first prove that if VMP converges in Gaussian graphical models under a valid message schedule, the mean vectors of the converged variational marginal distributions are the exact marginal mean vectors, giving the legitimacy of applying VMP in Gaussian models. In order to find the optimal variational marginal distributions using VMP, serial schedule, parallel schedule and randomized schedule are considered. In the basic serial schedule, the VMP is guaranteed to converge unconditionally. However, the basic serial schedule takes a long time in waiting for message update, which is not efficient in large-scale distributed networks. To this end, a group serial schedule with convergence guarantee is proposed for VMP update and achieves a faster convergence than the basic serial schedule.

To further accelerate the VMP convergence and allow the nodes update more frequently in large-scale distributed networks, parallel schedule is applied to VMP update and the corresponding convergence condition is derived. Since VMP may diverge in parallel schedule, damping could be applied in parallel schedule to improve convergence, and a feasible set of damping factors is derived. Recognizing that serial schedules and parallel schedule (with or without

damping) are periodic update schedules, VMP is proved to converge at a linear rate if it converges in these schedules. Furthermore, a randomized schedule with lower computational complexity and communication overhead is proposed for VMP update. With the interpretation of probabilistic damping, the necessary and sufficient convergence conditions are derived in probabilistic sense. Finally, the newly established theories on VMP convergence are corroborated by numerical results and applications, illustrating the trade-offs between convergence speed and applicability of different schedules.

The rest of the paper is organized as follows. In Section II, the general convergence properties of VMP in Gaussian graphical model are analyzed. In Section III, convergence-guaranteed models and schedules are presented. In Section IV, parallel schedule and its variants are considered with their convergence conditions analyzed. Section V studies the convergence rate and presents a summary of various message schedules. Numerical results and applications are presented in Section VI. Finally, conclusions are drawn in Section VII.

Notations: Scalars, vectors, matrices and sets are denoted by lower-case letters, bold lower-case letters, bold upper-case letters and calligraphic upper-case letters, respectively. $\|\mathbf{a}\|_2$ denotes the ℓ_2 norm of $\mathbf{a}$, and $\|\mathbf{A}\|_\infty$ denotes the ℓ_∞ induced matrix norm. The notation $\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$ and the notation $\mathbf{A} \succ \mathbf{0}$ means $\mathbf{A}$ is positive definite. $\mathbf{A}(i,j)$ denotes the element of $\mathbf{A}$ in the i-th row and j-th column while $\mathbf{A}_{ij}$ denotes the block element of $\mathbf{A}$ in the i-th row partition and j-th column partition. The notation $|\mathbf{A}|$ denotes the matrix taking element-wise absolute value of $\mathbf{A}$, while $\rho(\mathbf{A})$ denotes the largest absolute eigenvalue of $\mathbf{A}$. The notation $\text{blkdiag}(\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_n)$ denotes a block diagonal matrix with the main diagonal blocks being $\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_n$. The notation '$\mathcal{A} \setminus \mathcal{B}$' means all the elements in $\mathcal{A}$ except the elements in $\mathcal{B}$. For the multivariate Gaussian distribution of a real vector $\mathbf{x}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, we denote it as $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

II. VMP in Gaussian Graphical Model and Its Convergent Point Properties

Given a high-dimensional Gaussian distribution of a random vector $\mathbf{x} \in \mathbb{R}^N$ written in the form of (2), the task is to compute the mean vectors of the marginal distributions of $\mathbf{x}$'s subvectors in a distributed way. For a random vector $\mathbf{x}$, it can be partitioned into n non-overlapping subvectors, where $1 \leq n \leq N$. Denoting the i-th $(1 \leq i \leq n)$ subvector of $\mathbf{x}$ as $\mathbf{x}_i \in \mathbb{R}^{d_i}$, we have $\mathbf{x} \triangleq [\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_n^T]^T$ and $\sum_{i=1}^n d_i = N$. For any possible partitioning of $\mathbf{x}$, there always exists

a pairwise factorization

$$p(\mathbf{x}) \propto \prod_{i=1}^{n} \underbrace{\exp\left\{-\frac{1}{2}\mathbf{x}_i^T \mathbf{J}_{ii}\mathbf{x}_i + \mathbf{h}_i^T \mathbf{x}_i\right\}}_{f_i(\mathbf{x}_i)} \prod_{(i<k, k\in\mathcal{B}_i)} \underbrace{\exp\left\{-\mathbf{x}_k^T \mathbf{J}_{ik}\mathbf{x}_i\right\}}_{f_{ik}(\mathbf{x}_i, \mathbf{x}_k)}, \tag{3}$$

where $\mathbf{J}_{ik} \in \mathbb{R}^{d_i \times d_k}$ is the (i,k)-th block of the matrix $\mathbf{J}$, $\mathbf{h}$ has the subvectors $\{\mathbf{h}_i \in \mathbb{R}^{d_i}\}_{i=1}^n$ such that $\mathbf{h} \triangleq [\mathbf{h}_1^T, \mathbf{h}_2^T, \cdots, \mathbf{h}_n^T]^T$, and $\mathcal{B}_i \triangleq \{k \neq i | \mathbf{J}_{ik} \neq \mathbf{0}\}$. Equation (3) can be interpreted as a Gaussian graphical model, where the i-th node represents the variable $\mathbf{x}_i$ and the direct neighbors of node i are $\mathbf{x}_k$ with $k \in \mathcal{B}_i$.

Directly using the matrix inverse method, we obtain the mean vector of $p(\mathbf{x})$ being $\boldsymbol{\mu} = \mathbf{J}^{-1}\mathbf{h}$ and the covariance matrix $\boldsymbol{\Sigma} = \mathbf{J}^{-1}$. Then the $\mathbf{x}_i$'s marginal distribution $p(\mathbf{x}_i)$ can be obtained as a Gaussian distribution with the mean vector $\boldsymbol{\mu}(1 + \sum_{k=1}^{i-1} d_k : \sum_{k=1}^{i} d_k)$ and covariance matrix $\boldsymbol{\Sigma}(1 + \sum_{k=1}^{i-1} d_k : \sum_{k=1}^{i} d_k, 1 + \sum_{k=1}^{i-1} d_k : \sum_{k=1}^{i} d_k)$. However, the computational complexity of taking matrix inverse for a large-scale matrix $\mathbf{J}$ is huge and is cubic with respect to N. Moreover, if different $\{\mathbf{x}_i\}_{i=1}^n$ are distributed in disparate locations, the direct inversion method requires an additional step of gathering information to a central processing unit. Thus, it is not scalable in large-scale distributed setting.

Due to the distributed and low complexity nature, message passing algorithms are promising for computing the marginal distributions of a high-dimensional Gaussian distribution. Gaussian BP is a widely used message-passing algorithm for calculating the marginal distributions of a high-dimensional Gaussian distribution. However, Gaussian BP is not guaranteed to converge in loopy graphs [10], making it not necessarily applicable in every practical scenario. Another popular message-passing algorithm is variational message passing (VMP), which is a variant of variational inference. In this paper, we investigate VMP for Gaussian graphical model, revealing the trade-offs in different message passing schedules and the corresponding convergence properties.

The underlying principle of VMP is based on variational inference, which calculates approximate marginal distributions under the mean-field assumption and is always guaranteed to converge in a cyclic update order of the variational marginal distributions. In variational inference, a variational distribution $q(\mathbf{x})$ is used to approximate the target distribution $p(\mathbf{x})$, and the optimal $q(\mathbf{x})$ is sought by minimizing the Kullback-Leibler (KL) divergence between $q(\mathbf{x})$ and $p(\mathbf{x})$ given

by

$$D_{KL}(q(\mathbf{x})||p(\mathbf{x})) = \int q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}. \tag{4}$$

To facilitate the computation of the marginal distributions of $q(\mathbf{x})$, the mean-field approximation $q(\mathbf{x}) = \prod_{i=1}^{n} q_i(\mathbf{x}_i)$ is usually adopted, where $q_i(\mathbf{x}_i)$ is treated as an approximation to the marginal distribution $p(\mathbf{x}_i)$. With the pairwise factorization in the form of (3), the optimal $q_i^\star(\mathbf{x}_i)$ that minimizes the KL divergence in (4) with other $\{q_k(\mathbf{x}_k)\}_{k\neq i}$ held fixed can be shown to be [28, Equation 10.9]

$$q_i^\star(\mathbf{x}_i) = \frac{f_i(\mathbf{x}_i)}{Z_i} \prod_{k\in\mathcal{B}_i} \exp\left\{ \int q_k(\mathbf{x}_k) \ln f_{ik}(\mathbf{x}_i, \mathbf{x}_k) d\mathbf{x}_k \right\}, \tag{5}$$

where Z_i is a normalization constant. It can be observed that the update of $q_i^\star(\mathbf{x}_i)$ depends on $\{q_k(\mathbf{x}_k)\}_{k\in\mathcal{B}_i}$, which are the variational distributions of $\mathbf{x}_i$'s directly connected neighboring variables. Thus, $q_i^\star(\mathbf{x}_i)$ for different i can be updated alternatively, and the process resembles a message passing mechanism on the graphical model.

Generally, VMP under an arbitrary message update schedule can be written as

$$q_i^{(t)}(\mathbf{x}_i) = \begin{cases} \frac{f_i(\mathbf{x}_i)}{Z_i} \prod_{k\in\mathcal{B}_i} \exp\left\{ \int q_k^{(\tau_k^i(t-1))}(\mathbf{x}_k) \ln f_{ik}(\mathbf{x}_i, \mathbf{x}_k) d\mathbf{x}_k \right\} & \text{if } i \in \mathcal{S}_t \\ q_i^{(t-1)}(\mathbf{x}_i) & \text{otherwise} \end{cases} \tag{6}$$

where $q_i^{(t)}(\mathbf{x}_i)$ means the variational distribution of $\mathbf{x}_i$ after the t-th iteration, $\mathcal{S}_t$ ($t \in \mathbb{N}^+$) is a set describing which $\mathbf{x}_i$ would be updated at the t-th iteration, and $0 \leq \tau_k^i(t-1) \leq t-1$ denotes which previous version of the variational distribution of $\mathbf{x}_k$ is being used for the update of the variational distribution of $\mathbf{x}_i$ at the t-th iteration. Thus specifying $\mathcal{S}_t$ and $\tau_k^i(t-1)$ for all t, k, i determine a message schedule. For any message schedule, we can have the following proposition.

Proposition 1 ***Gaussian-form Messages:*** *For any Gaussian initialization $\{q_i^{(0)}(\mathbf{x}_i)\}_{i=1}^n$, the variational distributions $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ for all $t \geq 0$ are always Gaussian under any message schedule.*

Proof: For any Gaussian initialization $q_i^{(0)}(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_i^{(0)}, \boldsymbol{\Sigma}_i^{(0)})$, according to (6), with straightforward mathematical manipulations, we obtain the result at the first iteration as

$$q_i^{(1)}(\mathbf{x}_i) = \begin{cases} \frac{f_i(\mathbf{x}_i)}{Z_i} \prod_{k \in \mathcal{B}_i} \exp\left\{ \int q_k^{(\tau_k^i(0))}(\mathbf{x}_k) \ln f_{ik}(\mathbf{x}_i, \mathbf{x}_k) d\mathbf{x}_k \right\} \text{ if } i \in \mathcal{S}_1 \\ q_i^{(0)}(\mathbf{x}_i) \qquad \text{otherwise} \end{cases}$$

$$= \begin{cases} \mathcal{N}(\mathbf{x}_i; \mathbf{J}_{ii}^{-1}\mathbf{h}_i - \sum_{k \in \mathcal{B}_i} \mathbf{J}_{ii}^{-1}\mathbf{J}_{ik}\boldsymbol{\mu}_k^{(0)}, \mathbf{J}_{ii}^{-1}) & \text{if } \ i \in \mathcal{S}_1 \\ \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_i^{(0)}, \boldsymbol{\Sigma}_i^{(0)}) & \text{otherwise.} \end{cases}$$

Therefore, $q_i^{(1)}(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_i^{(1)}, \boldsymbol{\Sigma}_i^{(1)})$, where $\boldsymbol{\mu}_i^{(1)} = \mathbf{J}_{ii}^{-1}\mathbf{h}_i - \sum_{k \in \mathcal{B}_i} \mathbf{J}_{ii}^{-1}\mathbf{J}_{ik}\boldsymbol{\mu}_k^{(0)}$ and $\boldsymbol{\Sigma}_i^{(1)} = \mathbf{J}_{ii}^{-1}$ if $i \in \mathcal{S}_1$, or $\boldsymbol{\mu}_i^{(1)} = \boldsymbol{\mu}_i^{(0)}$ and $\boldsymbol{\Sigma}_i^{(1)} = \boldsymbol{\Sigma}_i^{(0)}$ otherwise. This means that $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ are Gaussian distributions for $t = 1$. Now for $t \geq 1$, suppose that $q_i^{(\tau)}(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_i^{(\tau)}, \boldsymbol{\Sigma}_i^{(\tau)})$ for all i and $\tau \leq t$. Then at the $(t+1)$-th iteration, we obtain

$$q_i^{(t+1)}(\mathbf{x}_i) = \begin{cases} \frac{f_i(\mathbf{x}_i)}{Z_i} \prod_{k \in \mathcal{B}_i} \exp\left\{ \int q_k^{(\tau_k^i(t))}(\mathbf{x}_k) \ln f_{ik}(\mathbf{x}_i, \mathbf{x}_k) d\mathbf{x}_k \right\} \text{ if } i \in \mathcal{S}_{t+1} \\ q_i^{(t)}(\mathbf{x}_i) \qquad \text{otherwise} \end{cases}$$

$$= \begin{cases} \mathcal{N}(\mathbf{x}_i; \mathbf{J}_{ii}^{-1}\mathbf{h}_i - \sum_{k \in \mathcal{B}_i} \mathbf{J}_{ii}^{-1}\mathbf{J}_{ik}\boldsymbol{\mu}_k^{(\tau_k^i(t))}, \mathbf{J}_{ii}^{-1}) \text{ if } i \in \mathcal{S}_{t+1} \\ \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)}) \qquad \text{otherwise.} \end{cases} \tag{7}$$

This implies that $q_i^{(t+1)}(\mathbf{x}_i)$ is also a Gaussian distribution. By induction, we have proved that $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ for all $t \geq 0$ are always Gaussian for any Gaussian initialization $\{q_i^{(0)}(\mathbf{x}_i)\}_{i=1}^n$. ■

From (7), it is observed that the covariance matrices of the variational marginal distributions are fixed at $\{\mathbf{J}_{ii}^{-1}\}_{i=1}^n$. Thus, the VMP in Gaussian graphical model consists of only updating the mean vectors using

$$\boldsymbol{\mu}_i^{(t+1)} = \begin{cases} \mathbf{J}_{ii}^{-1}\mathbf{h}_i - \sum_{k \in \mathcal{B}_i} \mathbf{J}_{ii}^{-1}\mathbf{J}_{ik}\boldsymbol{\mu}_k^{(\tau_k^i(t))} & \text{if } i \in \mathcal{S}_{t+1} \\ \boldsymbol{\mu}_i^{(t)} & \text{otherwise} \end{cases} \tag{8}$$

for all i and $t \geq 0$, where $\boldsymbol{\mu}_i^{(t+1)}$ is the mean vector of $q_i^{(t+1)}(\mathbf{x}_i)$. This is different from Gaussian BP, where both mean vector and covariance matrix are updated at each iteration. It turns out that the covariance matrix $\mathbf{J}_{ii}^{-1}$ is generally not the exact covariance matrix of $p(\mathbf{x}_i)$ and overconfident. But one may wonder if the mean vector in (8) converges to the exact marginal mean vector of $p(\mathbf{x}_i)$ when $t \to \infty$.

Before we answer this question, we have to clarify that the message passing schedule under consideration should be a valid one in the sense that each node is updated infinitely often and

old messages of each node should be purged out of the network as $t \to \infty$. Mathematically, we have the following definition.

Definition 1 ***Valid Message Schedule:*** *A valid message passing schedule should satisfy* $\lim_{T\to\infty} \sum_{t=1}^{T} \mathbb{I}_{\mathcal{S}_t}(i) \to \infty$ *and* $\lim_{t\to\infty} \tau_k^i(t) \to \infty$ *for all* k, i*, where* $\mathbb{I}_{\mathcal{S}_t}(i)$ *takes the value of one if* $i \in \mathcal{S}_t$ *and zero otherwise.*

Under a valid message schedule, we formalize the correctness of the mean vectors of the variational distributions if VMP converges as follows.

Theorem 1 ***Correctness of the Converged Mean Vectors:****For the VMP updated by* (8)*, if it converges under a valid message schedule, the converged* $\boldsymbol{\mu}_i^{(t)}$ *equals to the exact marginal mean vector of* $p(\mathbf{x}_i)$ *for all* i*.*

Proof: From Proposition 1, $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ in each iteration are in Gaussian form. Thus, if the VMP algorithm converges under a valid message schedule, the converged variational marginal distributions are also in Gaussian form and we denote the converged distribution as $q_i^{(\infty)}(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i; \mathbf{m}_i^{(\infty)}, \mathbf{V}_i^{(\infty)})$. Since the VMP is derived under the mean-field assumption, the joint variational distribution of the whole vector $\mathbf{x}$ would be $q^{(\infty)}(\mathbf{x}) = \prod_{i=1}^n q_i^{(\infty)}(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}; \mathbf{m}^{(\infty)}, \mathbf{V}^{(\infty)})$ with $\mathbf{m}^{(\infty)} \triangleq [\mathbf{m}_1^{(\infty)T}, \mathbf{m}_2^{(\infty)T}, \cdots, \mathbf{m}_n^{(\infty)T}]^T$ and $\mathbf{V} \triangleq \text{blkdiag}(\mathbf{V}_1^{(\infty)}, \mathbf{V}_2^{(\infty)}, \cdots, \mathbf{V}_n^{(\infty)})$. Substituting the expressions of $q^{(\infty)}(\mathbf{x})$ and $p(\mathbf{x})$ into (4), after straightforward mathematical manipulations, we obtain

$$\begin{aligned} D_{KL}(q^{(\infty)}(\mathbf{x})||p(\mathbf{x})) = &-\frac{1}{2}\ln|\mathbf{J}\mathbf{V}^{(\infty)}| + \frac{1}{2}\text{Tr}(\mathbf{J}\mathbf{V}^{(\infty)}) \\ &+ \frac{1}{2}(\mathbf{m}^{(\infty)} - \mathbf{J}^{-1}\mathbf{h})^T\mathbf{J}(\mathbf{m}^{(\infty)} - \mathbf{J}^{-1}\mathbf{h}) + constant. \end{aligned} \tag{9}$$

Since each update of $q_i^\star(\mathbf{x}_i)$ in (5) is the global optimum when other $\{q_k(\mathbf{x}_k)\}_{k\in\mathcal{B}_i}$ are fixed, the gradient of (9) with respect to $\mathbf{m}_i^{(\infty)}$ should be zero. Furthermore, if the VMP converges, the gradient of (9) with respect to all $\{\mathbf{m}_i^{(\infty)}\}_{i=1}^n$ would be zero. That is,

$$\frac{\partial D_{KL}(q^{(\infty)}(\mathbf{x})||p(\mathbf{x}))}{\partial \mathbf{m}^{(\infty)}} = \mathbf{J}(\mathbf{m}^{(\infty)} - \mathbf{J}^{-1}\mathbf{h}) = \mathbf{0}. \tag{10}$$

Due to $\mathbf{J} \succ \mathbf{0}$, we obtain $\mathbf{m}^{(\infty)} = \mathbf{J}^{-1}\mathbf{h}$, which implies that $\{\mathbf{m}_i^{(\infty)}\}_{i=1}^n$ are the exact marginal mean vectors of $\{p(\mathbf{x}_i)\}_{i=1}^n$. ■

Theorem 1 states that the mean vectors of the variational distributions are indeed the exact marginal mean vectors that we intend to find (i.e., VMP in Gaussian graphical model achieves

Bayesian optimality if it converges). This gives the legitimacy of using VMP in Gaussian graphical models. This result is general in the sense that it holds for any message passing schedule as long as it is a valid one, and if the VMP converges.

III. Convergence Guaranteed Graphical Model Structures and Message Schedules for VMP

A. Convergence Guaranteed Graphical Models

Given that the converged mean vectors in Gaussian VMP is the correct marginal mean vectors, the next important question is when would VMP converges in Gaussian graphical models? Notice that (8) can be interpreted as an asynchronous relaxation of the vector $\boldsymbol{\mu}^{(t)} \triangleq [\boldsymbol{\mu}_1^{(t)T}, \boldsymbol{\mu}_2^{(t)T}, \cdots, \boldsymbol{\mu}_n^{(t)T}]^T$ with the update equation

$$\boldsymbol{\mu}^{(t)} = \mathbf{B}\boldsymbol{\mu}^{(t-1)} + \mathbf{d}, \tag{11}$$

where the iteration matrix $\mathbf{B}$ has the nonzero block element $\mathbf{B}_{ik} = -\mathbf{J}_{ii}^{-1}\mathbf{J}_{ik}$ for all $i = 1, 2, \cdots, n$ and $k \in \mathcal{B}_i$, and the block vector $\mathbf{d}$ has the block element $\mathbf{d}_i = \mathbf{J}_{ii}^{-1}\mathbf{h}_i$. Then, by using the Asynchronous Convergence Theorem [35, p.431], we have $\rho(|\mathbf{B}|) < 1$ is a sufficient convergence condition for (8) under any valid message passing schedule and any initialization $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$, where $|\mathbf{B}|$ denotes the matrix with the element-wise absolute value of $\mathbf{B}$. In the following, two classes of information matrix $\mathbf{J}$, namely diagonally dominant and walk-summable, are proved to guarantee $\rho(|\mathbf{B}|) < 1$. Although diagonally dominant and walk-summable conditions have been proved to be sufficient conditions for the convergence of Gaussian BP, the proofs for the sufficiency of these conditions for VMP convergence are totally different.

Proposition 2 ***Diagonal Dominance:*** *If* $\mathbf{J}$ *is a diagonally dominant matrix and* $\{\mathbf{J}_{ii}\}_{i=1}^n$ *are diagonal matrices, VMP on Gaussian graphical model converges under any valid message schedule.*

Proof: If $\mathbf{J}_{ii}$ is a diagonal matrix, $\mathbf{J}_{ii}^{-1}$ is also a diagonal matrix, whose diagonal elements $\mathbf{J}_{ii}^{-1}(k,k) = 1/\mathbf{J}(\kappa,\kappa)$ with $\kappa \triangleq \sum_{m=1}^{i-1} d_m + k$ and $k = 1, 2, \cdots, d_i$. For the κ-th row of $|\mathbf{B}|$,

we have

$$\sum_{j\neq\kappa}|\mathbf{B}(\kappa,j)| = \sum_{j\in\mathcal{B}_i}\sum_{d=1}^{d_j}|-\mathbf{J}_{ii}^{-1}(k,k)\mathbf{J}_{ij}(k,d)|$$
$$=\sum_{j\in\mathcal{B}_i}\sum_{d=1}^{d_j}\frac{1}{\mathbf{J}(\kappa,\kappa)}|\mathbf{J}(\kappa,\sum_{m=1}^{j-1}d_m+d)| = \sum_{j\neq\kappa}\frac{1}{\mathbf{J}(\kappa,\kappa)}|\mathbf{J}(\kappa,j)|. \quad (12)$$

The last equality is due to $\mathbf{J}_{ij}=\mathbf{0}$ for $j\notin\mathcal{B}_i\cup i$. If $\mathbf{J}$ is a diagonally dominant matrix, then $\sum_{j\neq\kappa}|\mathbf{J}(\kappa,j)|<\mathbf{J}(\kappa,\kappa)$, which implies $\sum_{j\neq\kappa}|\mathbf{B}(\kappa,j)|=\sum_{j\neq\kappa}1/\mathbf{J}(\kappa,\kappa)|\mathbf{J}(\kappa,j)|<1$. Therefore, we obtain $\||\mathbf{B}|\|_\infty<1$. Since $\rho(|\mathbf{B}|)$ is less than any induced matrix norm [35, Proposition A.20], we obtain $\rho(|\mathbf{B}|)<1$, which leads to convergence of VMP under any valid message schedule. ■

Proposition 3 ***Walk-Summability:*** *If* $\mathbf{J}$ *is walk-summable with* $d_i=1$ *for all* $i=1,2,\cdots,n$*, VMP on Gaussian graphical model converges under any valid message schedule.*

Proof: Since any Gaussian model can be normalized easily with $\mathbf{J}(i,i)=1$, it leads to $\mathbf{J}=\mathbf{I}-\mathbf{R}$, where $\mathbf{R}$ has zero diagonal elements and off-diagonal elements such that $\mathbf{R}(i,j)=-\mathbf{J}(i,j)$. The walk-summability implies that $\rho(|\mathbf{R}|)<1$ [16]. On the other hand, when $d_i=1$ for all $i=1,2,\cdots,n$, the iteration matrix $\mathbf{B}$ has the element $\mathbf{B}(i,k)=-\mathbf{J}(i,k)/\mathbf{J}(i,i)$. Due to $\mathbf{J}(i,i)=1$, we have $\mathbf{B}(i,k)=-\mathbf{J}(i,k)$. Thus we obtain $\mathbf{B}=\mathbf{R}$. If $\rho(|\mathbf{R}|)<1$, then $\rho(|\mathbf{B}|)<1$, which leads to convergence of VMP under any valid message schedule. ■

Propositions 2 and 3 reveal that there are some subclasses of Gaussian graphical models where VMP always converge no matter what message passing schedule we use as long as it is a valid one. Then, by Theorem 1, the correct marginal mean vectors can be obtained in these cases. But as $\mathbf{B}$ is dictated by the system setting and usually not under our control, a natural question is if $\rho(|\mathbf{B}|)<1$ is not satisfied, is there any way to construct the message passing schedules such that the VMP converges. Next two subsections would be dedicated to this question.

B. Basic Serial Schedule

The most basic form of serial schedule is to update one of $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ at each iteration, and $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ are updated cyclically. Without loss of generality, we consider $\mathcal{S}_t=\{k|k=mod(t,n)\}$ and $\tau_k^i(t-1)=t-1$ for all $t\geq 1$, where $mod(t,n)$ means the modulus of t with respect to n. It is noted that the update of $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ in the basic serial schedule can be viewed as the result of a coordinate descent algorithm in functional form for finding the minimum of

the KL divergence in (4). Thus $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ converge to a local optimum of the KL divergence [27] as $t \to \infty$, which equivalently implies that the mean vectors $\{\boldsymbol{\mu}_i^{(t)}\}_{i=1}^n$ in (8) converge in the basic serial schedule. Further with Theorem 1, the mean vector $\boldsymbol{\mu}_i^{(t)}$ converges to the exact mean vector of $p(\mathbf{x}_i)$ for all i.

In the basic serial schedule, VMP is guaranteed to converge no matter how many components the random vector $\mathbf{x}$ is partitioned. However, given a vector variable $\mathbf{x}$ of length N, the larger number of subvectors, the longer an update cycle would be. On the other hand, as the computational complexity at each update step is cubic with respect to the dimension d_i for the update of $\boldsymbol{\mu}_i^{(t)}$ using (8), there exists a trade-off between the complexity of updating the mean of each subvector and the number of update steps in an update cycle. For example, taking $n = 1$ means there is no partitioning of the vector $\mathbf{x}$ (this case actually is equivalent to direct matrix inversion method), while taking $n = N$ means the vector $\mathbf{x}$ is partitioned such that each component is a scalar. Convergence is guaranteed under the basic serial schedule for both partitions, but the number of update steps in a cycle is 1 in the former case and N in the latter case. On the other hand, each update step for the former case requires a matrix inversion of size $N \times N$, while there is no matrix inverse in the latter case.

Notice that as the indexing of different subvectors is arbitrary, the above conclusion is valid for any cyclic update schedule being a permutation of $\{1, 2, \cdots, n\}$. In fact, it allows different update orders at different update cycles, as long as each of the $\{\mathbf{x}_i\}_{i=1}^n$ is updated once in each cycle. However, no matter what update order we choose in a cycle, since only one component of $\{q_i(\mathbf{x}_i)\}_{i=1}^n$ is updated at a time in the basic serial schedule, it is not difficult to figure that the convergence of VMP could be slow. Each node would wait a long time for other nodes' update, which is not efficient in large-scale distributed networks. To speed up the convergence, we introduce a group serial schedule in the next subsection.

C. Group Serial Schedule

In certain applications, e.g., peer-to-peer rating and distributed localization, how $\mathbf{x}$ is partitioned into $\mathbf{x}_i$ is governed by the system setting, and not under our control. In this case, we wonder if there is any update schedule that is faster than the basic serial schedule but still guarantees VMP convergence. To design such schedule, we first define a new class of schedule, called group serial schedule as follows.

Definition 2 ***Group Serial Schedule:*** *All the subvectors $\{\mathbf{x}_i\}_{i=1}^n$ are divided into multiple groups, and only the variational distributions of subvectors belonging to the same group would be updated at the same time, while the variational distributions of subvectors in different groups are updated alternatively in a cyclic manner.*

Note that the basic serial schedule can be viewed as a special case of group serial schedule with n groups, where each group only contains one subvector. In group serial schedule, there exist many different choices of grouping of $\{\mathbf{x}_i\}_{i=1}^n$. However, not all groupings guarantee VMP convergence. Thus the key problem is to identify a rule to group the subvectors $\{\mathbf{x}_i\}_{i=1}^n$ such that VMP is convergent.

Recognizing that in VMP with basic serial schedule, the update of $q_i(\mathbf{x}_i)$ minimizes the KL divergence in (4) when other $\{q_k(\mathbf{x}_k)\}_{k\neq i}$ are fixed, thus we have the monotonically decreasing property: $D_{KL}(\prod_{k=1}^n q_k^{(t)}(\mathbf{x}_k)||p(\mathbf{x})) \leq D_{KL}(\prod_{k=1}^n q_k^{(t-1)}(\mathbf{x}_k)||p(\mathbf{x}))$ for all $t \geq 1$. Inspired by this underlying principle in the basic serial schedule, we propose a subclass of group serial schedule that guarantees the KL divergence in (4) monotonically decreases.

Suppose that all the $\{\mathbf{x}_i\}_{i=1}^n$ are divided into m groups with the indices of nodes belonging the j-th group collected by $\mathcal{C}_j$, where $\mathbf{x}_i$ belongs to the j-th group if and only if $i \in \mathcal{C}_j$. If the grouping is constructed such that the random variables within each group are conditionally independent when the variables in other groups are observed, we have the following conclusion.

Theorem 2 ***Convergence Guaranteed Group Serial Schedule:*** *In a group serial schedule with the random variables within a group being conditionally independent when variables in other groups are observed, $\{\boldsymbol{\mu}_i^{(t)}\}_{i=1}^n$ in* (8) *converge to the exact marginal mean vectors of $\{p(\mathbf{x}_i)\}_{i=1}^n$ for any initialization $\boldsymbol{\mu}_i^{(0)} \in \mathbb{R}^{d_i}$ when $t \to \infty$.*

Proof: Since all random variables within each group are conditionally independent, for any $i \in \mathcal{C}_j$, we have $\mathcal{B}_i \cap \mathcal{C}_j = \emptyset$. In other words, the direct neighbours of $\mathbf{x}_i$ will not be put in the set $\mathcal{C}_j$ if $i \in \mathcal{C}_j$. This also implies that the update equation of $q_i^{(t)}(\mathbf{x}_i)$ in (6) with $i \in \mathcal{C}_j$ does not depend on $q_k^{(t-1)}(\mathbf{x}_k)$ for $k \in \mathcal{C}_j$. Thus, minimizing the function $D_{KL}(\prod_{i\in\mathcal{C}_j} q_i(\mathbf{x}_i) \prod_{k\notin\mathcal{C}_j} q_k^{(t-1)}(\mathbf{x}_k)||p(\mathbf{x}))$ with respect to $\{q_i(\mathbf{x}_i)\}_{i\in\mathcal{C}_j}$ at the same time is equivalent to minimizing the function $D_{KL}(q_i(\mathbf{x}_i) \prod_{k\in\mathcal{C}_j\setminus\{i\}} q_k^{(t)}(\mathbf{x}$ with respect to $\{q_i(\mathbf{x}_i)\}_{i\in\mathcal{C}_j}$ one at a time. Since basic serial update is convergent, group serial update is also convergent. Further with Theorem 1, $\{\boldsymbol{\mu}_i^{(t)}\}_{i=1}^n$ converge to the exact marginal mean vectors when $t \to \infty$. ■

Theorem 2 gives a desired group serial schedule that guarantees the convergence of $\{\boldsymbol{\mu}_i^{(t)}\}_{i=1}^n$. The key task is to divide all the subvectors into multiple groups such that all random variables corresponding to each group are conditionally independent when variables in other groups are observed. This can be viewed as a graph coloring problem, in which each node denotes a variable of $\mathbf{x}_i$ with $i \in \{1, 2, \cdots, n\}$ and no two adjacent nodes are of the same color. In the view of graph coloring, there exists a minimum group number, which may be hard to find. While having the minimum number of groups would speed up the convergence, it is not required to find the optimal group number in the proposed group serial schedule. Thus some heuristic algorithms [38] for graph coloring can be applied to find the groups $\{\mathcal{C}_j\}_{j=1}^m$ easily.

Remark: Note that serial message schedule and group serial schedule are also convergence guaranteed for VMP in non-Gaussian models, such as those in localization application [29], [36], [37]. Especially, the group serial schedule is a modification to the basic serial schedule that can be adopted easily. This opens up many opportunities in speeding up VMP in a wide range of applications.

IV. VMP Convergence Conditions in Parallel and Randomized Message schedules

Although group serial schedule improves the convergence speed from the basic serial schedule, in each time instant, there are only a small number of nodes (compared to the total number of nodes) in the graphical model are being updated. This is especially true if the average number of direct neighbors of each node is large, leading to a substantial number of groups. If we can update all the nodes at every iteration, it is expected the convergence speed would be fast. However, in general parallel update, there is no guarantee of convergence of VMP. In this section, we derive the convergence conditions for parallel update schedule and its variants.

A. Parallel Schedule

In parallel schedule, all $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ are updated at each iteration and each update is based on the latest updated result from the last iteration, which means $\mathcal{S}_t \triangleq \{1, 2, \cdots, n\}$ and $\tau_k^i(t-1) = t-1$ for all i, k and $t \geq 1$. It can be interpreted as an extreme form of group serial schedule where there is only one group and it contains all variables. However, since parallel schedule does not satisfy the conditionally independent requirement in Theorem 2, it does not necessarily converges. To derive the convergence condition of parallel schedule, we notice that the corresponding update

equation for $\boldsymbol{\mu}^{(t)} \triangleq [\boldsymbol{\mu}_1^{(t)T}, \boldsymbol{\mu}_2^{(t)T}, \cdots, \boldsymbol{\mu}_n^{(t)T}]^T$ is given by $\boldsymbol{\mu}^{(t)} = \mathbf{B}\boldsymbol{\mu}^{(t)} + \mathbf{d}$, i.e., (11), the convergence condition is presented in the following proposition.

Proposition 4 ***Necessary and Sufficient Convergence Condition in Parallel Schedule:*** $\{\boldsymbol{\mu}_i^{(t)}\}_{i=1}^n$ *converge to the exact marginal mean vectors of* $\{p(\mathbf{x}_i)\}_{i=1}^n$ *for any initialization* $\boldsymbol{\mu}_i^{(0)} \in \mathbb{R}^{d_i}$ *in parallel schedule if and only if* $\rho(\mathbf{B}) < 1$.

Proof: For the linear update equation in (11), using [35, Proposition 2.6.1], $\boldsymbol{\mu}^{(t)}$ converges for any initialization $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$ if and only if $\rho(\mathbf{B}) < 1$. If $\boldsymbol{\mu}^{(t)}$ converges, further with Theorem 1, we obtain that $\{\boldsymbol{\mu}^{(t)}\}$ in parallel schedule would converge to the exact marginal mean vectors. ■

Proposition 4 gives the necessary and sufficient convergence condition of VMP under parallel schedule. Due to the general property that $\rho(\mathbf{B}) \leq \rho(|\mathbf{B}|)$, we have $\rho(\mathbf{B}) < 1$ if $\rho(|\mathbf{B}|) < 1$. This indicates that the convergence condition of VMP in parallel schedule is less stringent than that for any valid schedule.

B. Parallel Schedule with Damping

VMP in parallel schedule would diverge if $\rho(\mathbf{B}) \geq 1$. To improve the convergence of VMP in parallel schedule, we could introduce damping, which modifies the update equation of $\boldsymbol{\mu}^{(t)}$ in (11) as

$$\begin{aligned} \boldsymbol{\mu}^{(t)} &= r(\mathbf{B}\boldsymbol{\mu}^{(t-1)} + \mathbf{d}) + (1-r)\boldsymbol{\mu}^{(t-1)} \\ &= (r\mathbf{B} + (1-r)\mathbf{I})\boldsymbol{\mu}^{(t-1)} + r\mathbf{d}, \end{aligned} \tag{13}$$

where $r \neq 0$ is a damping factor. For the update equation of $\boldsymbol{\mu}^{(t)}$ in (13), it has a similar form as the update equation in parallel schedule. Therefore, using [35, Proposition 2.6.1], we obtain that the damped $\boldsymbol{\mu}^{(t)}$ converges for any initialization $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$ if and only if $\rho(r\mathbf{B} + (1-r)\mathbf{I}) < 1$. Moreover, the converged value $\boldsymbol{\mu}^{(\infty)}$ in damped VMP should satisfy (13), i.e.,

$$\boldsymbol{\mu}^{(\infty)} = (r\mathbf{B} + (1-r)\mathbf{I})\boldsymbol{\mu}^{(\infty)} + r\mathbf{d}, \tag{14}$$

which can be rewritten as

$$r((\mathbf{B} - \mathbf{I})\boldsymbol{\mu}^{\infty} + \mathbf{d}) = \mathbf{0}. \tag{15}$$

Since $r \neq 0$, we must have $(\mathbf{B} - \mathbf{I})\boldsymbol{\mu}^{(\infty)} + \mathbf{d} = \mathbf{0}$, which is equivalent to $\boldsymbol{\mu}^{(\infty)} = \mathbf{B}\boldsymbol{\mu}^{(\infty)} + \mathbf{d}$. This implies that the converged vector $\boldsymbol{\mu}^{(\infty)}$ in damped VMP is the same as that in parallel

schedule. Therefore, the damped $\boldsymbol{\mu}^{(t)}$ converges to the exact marginal mean vector if and only if $\rho(r\mathbf{B}+(1-r)\mathbf{I})<1$. In the following proposition, the appropriate value of the damping factor r is derived.

Proposition 5 ***Necessary and Sufficient Convergence of Damped Mean Vector:*** *The damped $\boldsymbol{\mu}^{(t)}$ converges to the exact mean vector of $p(\mathbf{x})$ for any initialization $\boldsymbol{\mu}^{(0)}\in\mathbb{R}^N$ if and only if there exists a damping factor r such that*

$$r\in\begin{cases}(\frac{2}{1-\max\limits_k\lambda_k},0) & \textit{if } \lambda_k>1 \textit{ for all } k=1,2,\cdots,N,\\ (0,\frac{2}{1-\min\limits_k\lambda_k}) & \textit{if } \lambda_k<1 \textit{ for all } k=1,2,\cdots,N,\end{cases}$$

where λ_k is the k-th eigenvalue of the matrix $\mathbf{B}$.

Proof: First, we prove that the eigenvalues $\{\lambda_k\}$ of $\mathbf{B}$ are real numbers. For notational convenience, we denote a block matrix $\mathbf{D}\triangleq[\text{blkdiag}(\mathbf{J}_{11},\mathbf{J}_{22},\cdots,\mathbf{J}_{nn})]^{-1}$. The matrix $\mathbf{B}$ can be rewritten as $\mathbf{B}=\mathbf{I}-\mathbf{DJ}$. Let $\bar{\lambda}_k$ and $\bar{\mathbf{v}}_k$ be the k-th eigenvalue and the corresponding eigenvector of $\mathbf{DJ}$, respectively. Then we have $\mathbf{DJ}\bar{\mathbf{v}}_k=\bar{\lambda}_k\bar{\mathbf{v}}_k$. If both sides of this equation are multiplied by $\mathbf{D}^{-\frac{1}{2}}$, we obtain $\mathbf{D}^{\frac{1}{2}}\mathbf{J}\bar{\mathbf{v}}_k=\bar{\lambda}_k\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{v}}_k$, which can be rewritten as $\mathbf{D}^{\frac{1}{2}}\mathbf{J}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{v}}_k=\bar{\lambda}_k\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{v}}_k$. Thus, $\bar{\lambda}_k$ and $\mathbf{D}^{-\frac{1}{2}}\bar{\mathbf{v}}_k$ can be viewed as the eigenvalue and the corresponding eigenvector of $\mathbf{D}^{\frac{1}{2}}\mathbf{J}\mathbf{D}^{\frac{1}{2}}$. With the block diagonal structure of $\mathbf{D}$ and a positive-definite matrix $\mathbf{J}$, it could be easily verified that the matrix $\mathbf{D}^{\frac{1}{2}}\mathbf{J}\mathbf{D}^{\frac{1}{2}}$ is symmetric. Therefore, the matrix $\mathbf{D}^{\frac{1}{2}}\mathbf{J}\mathbf{D}^{\frac{1}{2}}$ only has real eigenvalues, which implies that $\{\bar{\lambda}_k\}$ are real numbers. Since $\{\bar{\lambda}_k\}$ are also the eigenvalues of $\mathbf{DJ}$, we obtain that the matrix $\mathbf{DJ}$ only has real eigenvalues. With $\mathbf{B}=\mathbf{I}-\mathbf{DJ}$, we have the eigenvalues of $\mathbf{B}$ as $\lambda_k=1-\bar{\lambda}_k$ for $k=1,2,\cdots,n$. Since $\{\bar{\lambda}_k\}$ are real numbers, $\{\lambda_k\}$ are also real numbers.

On the other hand, we obtain the k-th eigenvalue of $r\mathbf{B}+(1-r)\mathbf{I}$ as $r\lambda_k+(1-r)$. The condition $\rho(r\mathbf{B}+(1-r)\mathbf{I})<1$ is satisfied if and only if $|r\lambda_k+(1-r)|<1$, or equivalently $(r\lambda_k+(1-r))^2<1$, for all k. Directly solving the inequality, we obtain the following results: a) If $\lambda_k>1$, $\frac{2}{1-\lambda_k}<r<0$. b) If $\lambda_k=1$, $r=\emptyset$. c) If $\lambda_k<1$, $0<r<\frac{2}{1-\lambda_k}$. Therefore, if $\lambda_k>1$ for all $k=1,2,\cdots,n$, in order to guarantee convergence, the damping factor should satisfy $r\in(\max\limits_k\frac{2}{1-\lambda_k},0)$. Since $\frac{2}{1-\lambda_k}$ is an increasing function when $\lambda_k>1$, $\max\limits_k\frac{2}{1-\lambda_k}=\frac{2}{1-\max\limits_k\lambda_k}$. If $\lambda_k<1$ for all $k=1,2,\cdots,n$, in order to guarantee convergence, the damping factor should satisfy $r\in(0,\min\limits_k\frac{2}{1-\lambda_k})$. Since $\frac{2}{1-\lambda_k}$ is an increasing function when $\lambda_k<1$, $\min\limits_k\frac{2}{1-\lambda_k}=\frac{2}{1-\min\limits_k\lambda_k}$. Otherwise, the damping factor that guarantee convergence does not exist. ■

From Proposition 5, even when VMP in parallel schedule diverges, damped VMP will converge if the feasible set of damping factor is non-empty. In fact, Proposition 5 is a generalization of Proposition 4 since $\rho(\mathbf{B}) < 1$ implies $-1 < \lambda_k < 1$ for all k. According to Proposition 5, the allowable damping factor $r \in (0, \frac{2}{1-\min\limits_k \lambda_k})$. Due to $-1 < \lambda_k < 1$ for all $k = 1, 2, \cdots, n$, we obtain $\frac{2}{1-\min\limits_k \lambda_k} > 1$. Thus $(0, 1]$ is a subset of $(0, \frac{2}{1-\min\limits_k \lambda_k})$, and $r = 1$ is always included in the set of allowable damping factors. The above analysis also indicates that the damping factor r is not limited to the classical damping assumption $r \in (0, 1)$.

C. Randomized Schedule as a Probabilistic Damping

Note that damping could improve the VMP convergence under parallel schedule but requires all the components $\{q_i^{(t)}(\mathbf{x}_i)\}_{i=1}^n$ being updated at each iteration. This leads to frequent computation and communication for each node, which is not efficient if these nodes have different local computation and communication resources. To reduce the computation burden of some nodes or allow some nodes update more frequently than the others, we propose a randomized schedule, which allows each node to independently update or not update the message with a predefined probability. If one node does not update the message at an iteration, there is no need to retransmit old messages to neighboring nodes at this iteration, which reduces the communication overhead of the whole network as well as the computational complexity.

In particular, in randomized schedule, whether $q_i^{(t)}(\mathbf{x}_i)$ is updated at the t-th iteration is determined by a Bernoulli random variable $\psi_i^{(t)}$, where $Pr(\psi_i^{(t)} = 1) = p_i$ and $Pr(\psi_i^{(t)} = 0) = 1 - p_i$ with $p_i \in (0, 1]$. If $\psi_i^{(t)} = 1$, then $i \in \mathcal{S}_t$ and $q_i^{(t)}(\mathbf{x}_i)$ will be updated. Otherwise, $q_i^{(t)}(\mathbf{x}_i)$ will be maintained as $q_i^{(t)}(\mathbf{x}_i) = q_i^{(t-1)}(\mathbf{x}_i)$. Furthermore, if $p_i = 1$ for all $i = 1, 2, \cdots, n$, randomized schedule reduces to parallel schedule. According to the strong law of large numbers, $Pr(\lim_{T\to\infty} \sum_{t=1}^T \mathbb{I}_{\mathcal{S}_t}(i) \to \infty) = 1$. This implies that randomized schedule is a valid message passing schedule with probability 1.

Under randomized schedule, the update equation of $\boldsymbol{\mu}_i^{(t)}$ can be written as

$$\boldsymbol{\mu}_i^{(t)} = \psi_i^{(t)}(\mathbf{J}_{ii}^{-1}\mathbf{h}_i - \sum\nolimits_{k\in\mathcal{B}_i} \mathbf{J}_{ii}^{-1}\mathbf{J}_{ik}\boldsymbol{\mu}_k^{(t-1)}) + (1-\psi_i^{(t)})\boldsymbol{\mu}_i^{(t-1)}, \tag{16}$$

from which we obtain the update equation of $\boldsymbol{\mu}^{(t)}$ as

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\Psi}^{(t)}(\mathbf{B}\boldsymbol{\mu}^{(t-1)} + \mathbf{d}) + (\mathbf{I} - \boldsymbol{\Psi}^{(t)})\boldsymbol{\mu}^{(t-1)}, \tag{17}$$

where $\boldsymbol{\Psi}^{(t)} \triangleq \text{blkdiag}(\psi_1^{(t)}\mathbf{I}_1, \psi_2^{(t)}\mathbf{I}_2, \cdots, \psi_n^{(t)}\mathbf{I}_n)$ and $\mathbf{I}_k$ is a $d_k \times d_k$ identity matrix. Since the update equation of $\boldsymbol{\mu}^{(t)}$ in (17) includes random variables, $\boldsymbol{\mu}^{(t)}$ becomes a random sequence.

Therefore, the convergence needs to be defined probabilistically. In this paper, we study the expectation convergence and mean-square convergence of $\boldsymbol{\mu}^{(t)}$ as follows.

Proposition 6 ***Necessary and Sufficient Convergence Condition of Randomized Schedule in Expectation Sense:*** *In randomized schedule, the expectation* $\mathbb{E}_{\{\Psi^{(k)}\}_{k=1}^{t}}[\boldsymbol{\mu}^{(t)}]$ *converge to the exact marginal mean vector for any initialization* $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$ *if and only if* $\rho\left(\mathbf{PB}+\mathbf{I}-\mathbf{P}\right)<1$*, where* $\mathbf{P}=\text{blkdiag}(p_1\mathbf{I}_1, p_2\mathbf{I}_2, \cdots, p_n\mathbf{I}_n)$.

Proof: See Appendix B. ■

Comparing randomized schedule and parallel schedule with damping, we can see that randomized schedule can be interpreted as a probabilistic damping. However, parallel schedule with damping only uses one r to damp the update schedule while randomized schedule uses a dedicated "damping" factor p_i for $\mathbf{x}_i$. If we choose to use a single probability for all p_i in randomized schedule, the appropriate update probability to ensure expectation convergence can be computed using the second case of Proposition 5 since the message update probability has to take positive values.

Proposition 7 ***Necessary and Sufficient Convergence Condition of Randomized Schedule in Mean-Square Sense:*** *In randomized schedule,* $\boldsymbol{\mu}^{(t)}$ *converges to the exact marginal mean vector in mean-square sense for any initialization* $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$ *if and only if* $\rho(\boldsymbol{\Phi})<1$*, where* $\boldsymbol{\Phi} \triangleq \text{blkdiag}(\mathbf{I}_1\otimes\boldsymbol{\Gamma}_1, \mathbf{I}_2\otimes\boldsymbol{\Gamma}_2, \cdots, \mathbf{I}_n\otimes\boldsymbol{\Gamma}_n)((\mathbf{B}-\mathbf{I})\otimes(\mathbf{B}-\mathbf{I}))+(\mathbf{PB}-\mathbf{P})\otimes\mathbf{I}+\mathbf{I}\otimes(\mathbf{PB}-\mathbf{P})+\mathbf{I}\otimes\mathbf{I}$*. In the above expression,* $\boldsymbol{\Gamma}_k$ *is a* $N\times N$ *diagonal matrix with the nonzero elements* $\boldsymbol{\Gamma}_k(i,i)=p_k$ *for all* $i=1+\sum_{k'=1}^{k-1}d_{k'}, \cdots, \sum_{k'=1}^{k}d_{k'}$ *and otherwise* $\boldsymbol{\Gamma}_k(i,i)=p_kp_j$ *with* $i=1+\sum_{k'=1}^{j-1}d_{k'}, \cdots, \sum_{k'=1}^{j}d_{k'}$ *and* $j\neq k$.

Proof: See Appendix C. ■

V. Convergence Rate Analysis and Summary

A. Convergence Rate

Note that basic serial schedule, group serial schedule, parallel schedule, and damped parallel schedule use a periodic update of $\{\boldsymbol{\mu}_i^{(t)}\}_{i=1}^n$. In particular, for the basic serial schedule, if node i is updated using (8) at time t, the update equation can be written as

$$\boldsymbol{\mu}^{(t)}=\mathbf{A}^{(i)}\boldsymbol{\mu}^{(t-1)}+\mathbf{c}^{(i)}, \tag{18}$$

where $\mathbf{A}^{(i)}$ is a block matrix with the nonzero block elements $\mathbf{A}^{(i)}_{kk} = \mathbf{I}_{d_k}$ for $k \neq i$ and $\mathbf{A}^{(i)}_{ij} = -\mathbf{J}_{ii}^{-1}\mathbf{J}_{ij}$ for $j \in \mathcal{B}_i$, and $\mathbf{c}^{(i)}$ is a block vector with a nonzero block $\mathbf{c}^{(i)}_i = \mathbf{J}_{ii}^{-1}\mathbf{h}_i$. If we focus on $\boldsymbol{\mu}^{(t)}$ after every n updates (i.e., at $t = n(l-1)$-th iteration with $l \in \mathbb{N}^+$), the update equation of $\boldsymbol{\mu}^{(t)}$ at the $(t = nl)$-th iteration can be written as

$$\boldsymbol{\mu}^{(nl)} = \prod_{i=1}^{n} \mathbf{A}^{(n-i+1)} \boldsymbol{\mu}^{(n(l-1))} + \sum_{i=1}^{n} \prod_{k>i}^{n} \mathbf{A}^{(n-k+i+1)} \mathbf{c}^{(i)}. \tag{19}$$

Similarly, for the group serial schedule, the one step update (8) can be written as

$$\boldsymbol{\mu}^{(t)} = \mathbf{E}^{(j)} \boldsymbol{\mu}^{(t-1)} + \mathbf{p}^{(j)} \tag{20}$$

where the block matrix $\mathbf{E}^{(j)}$ has the nonzero blocks $\mathbf{E}^{(j)}_{ik} = -\mathbf{J}_{ii}^{-1}\mathbf{J}_{ik}$ for $i \in \mathcal{C}_j$, $k \in \mathcal{B}_i$ and $\mathbf{E}^{(j)}_{ii} = \mathbf{I}$ for $i \notin \mathcal{C}_j$, and the block vector $\mathbf{p}^{(j)}$ has the nonzero blocks $\mathbf{p}^{(j)}_i = \mathbf{J}_{ii}\mathbf{h}_i$ for $i \in \mathcal{C}_j$. If we focus on t being a multiple of m, i.e., $t = ml$ with $l \in \mathbb{N}^+$, we have

$$\boldsymbol{\mu}^{(ml)} = \prod_{j=1}^{m} \mathbf{E}^{(m-j+1)} \boldsymbol{\mu}^{(m(l-1))} + \sum_{j=1}^{m} \prod_{k>j}^{m} \mathbf{E}^{(m+j-k+1)} \mathbf{p}^{(j)}. \tag{21}$$

Generalizing from (19) and (21), the update equation with any periodic schedule can be written as

$$\boldsymbol{\mu}^{(Tl)} = \boldsymbol{\Xi}_T \boldsymbol{\mu}^{(T(l-1))} + \mathbf{q}_T, \tag{22}$$

for some $\boldsymbol{\Xi}_T$ and $\mathbf{q}_T$, where T denotes the period of a serial update. Note that parallel schedule can be seen as a periodic update with $T = 1$. Furthermore, the damped parallel schedule update equation (13) is also in the same form of the periodic update equation in (22). With the update equation (22), if it converges, it is shown in the following proposition that the mean vector $\boldsymbol{\mu}^{(Tl)}$ for $l \in \mathbb{N}^+$ converges at a linear rate.

Proposition 8 ***Linear Convergence Rate in Periodic Update Schedule:*** *For the VMP periodic update schedule described by* (22)*, if it converges, $\boldsymbol{\mu}^{(Tl)}$ for $l \in \mathbb{N}^+$ converge linearly to the exact marginal mean vectors when $l \to \infty$.*

Proof: If the VMP converges, according to Theorem 1, $\lim_{l\to\infty} \boldsymbol{\mu}^{(Tl)} = \lim_{t\to\infty} \boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^* = \mathbf{J}^{-1}\mathbf{h}$. Moreover, $\boldsymbol{\mu}^*$ satisfies the equality in (22), i.e.,

$$\boldsymbol{\mu}^* = \boldsymbol{\Xi}_T \boldsymbol{\mu}^* + \mathbf{q}_T. \tag{23}$$

Taking the difference between (22) and (23), we obtain

$$\boldsymbol{\mu}^{(Tl)} - \boldsymbol{\mu}^* = \boldsymbol{\Xi}_T (\boldsymbol{\mu}^{(T(l-1))} - \boldsymbol{\mu}^*). \tag{24}$$

As we assume the VMP converges, we must have $\rho(\boldsymbol{\Xi}_T) < 1$, and there exists a matrix norm $\|\cdot\|$ such that $\|\boldsymbol{\Xi}_T\| < 1$. Taking such a norm on both sides of (24), we get $\|\boldsymbol{\mu}^{(Tl)} - \boldsymbol{\mu}^*\| \leq \|\boldsymbol{\Xi}_T\|\|\boldsymbol{\mu}^{(T(l-1))} - \boldsymbol{\mu}^*\|$. Thus we obtain

$$\lim_{l\to\infty} \frac{\|\boldsymbol{\mu}^{(Tl)} - \boldsymbol{\mu}^*\|}{\|\boldsymbol{\mu}^{(T(l-1))} - \boldsymbol{\mu}^*\|} \leq \|\boldsymbol{\Xi}_T\| < 1, \tag{25}$$

which indicates that $\boldsymbol{\mu}^{(Tl)}$ for $l \in \mathbb{N}^+$ converges linearly to $\boldsymbol{\mu}^*$. ■

Using Proposition 8, we can conclude that VMP in basic serial schedule, group serial schedule, parallel schedule and damped parallel schedule all converge linearly to the true marginal mean of $p(\mathbf{x}_i)$ for all i. On the other hand, for the probabilistic schedule, since its update equation (17) after expectation would in the form of (22), Proposition 8 indicates that the expectation of VMP under the randomized schedule converges linearly.

For damped VMP, when the allowable set of r in Proposition 5 is non-empty, one may wonder is there any r that is more favorable than others? The following proposition gives an answer.

Proposition 9 ***Faster Convergence:*** *When the damping factor set is non-empty, damping factor r that leads to a smaller $\|r\mathbf{B} + (1-r)\mathbf{I}\|_2$ has a lower convergence upper bound.*

Proof: Taking the difference between (13) and (14), we have $\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(\infty)} = (r\mathbf{B} + (1-r)\mathbf{I})(\boldsymbol{\mu}^{(t-1)} - \boldsymbol{\mu}^{(\infty)})$. Thus we can obtain $\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(\infty)} = (r\mathbf{B} + (1-r)\mathbf{I})^t(\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}^{(\infty)})$. Applying ℓ_2 norm to the above equation, and using the property $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{x}\|_2$, we have $\frac{\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(\infty)}\|_2}{\|\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}^{(\infty)}\|_2} \leq \|r\mathbf{B} + (1-r)\mathbf{I}\|_2^t$. Therefore, if r is chosen such that $\|r\mathbf{B} + (1-r)\mathbf{I}\|_2$ is smaller, the upper bound of $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(\infty)})\|_2$ will converge to zero faster. ■

Proposition 9 indicates that we should choose a r that leads to the smallest $\|r\mathbf{B} + (1-r)\mathbf{I}\|_2$ to ensure the fastest decrease of the convergence bound. Furthermore, even parallel schedule is valid (i.e., $r = 1$ is inside the allowable set of r according to Proposition 5), it is worthwhile to explore if there is a r that leads to smaller $\|r\mathbf{B} + (1-r)\mathbf{I}\|_2$ such that the convergence speed is faster. Therefore, damped VMP not only has the potential for converting a non-convergent parallel schedule to a convergent one, by properly choosing the damping factor r, the convergence speed can also be accelerated. Since $\|r\mathbf{B} + (1-r)\mathbf{I}\|_2^2$ is a quadratic function with respect to r, the r that leads to the minimum $\|r\mathbf{B} + (1-r)\mathbf{I}\|_2$ can be easily found. More evidence and discussion will be provided in the simulation section.

Table I: A summary of convergence conditions for VMP in different message schedules.

Schedule	Convergence Condition	Sufficiency or Necessary
Any valid schedule	Any one of the following conditions is satisfied: (1) $\rho(\lvert\mathbf{B}\rvert) < 1$ (2) $\mathbf{J}$ is diagonally dominant with $\mathbf{J}_{ii}$ being diagonal matrices (Proposition 2) (3) $\mathbf{J}$ is walk-summable (Proposition 3)	sufficiency
Basic serial schedule	Convergence guaranteed unconditionally	-
Group serial schedule	Any group of variables being conditionally independent when variables in other groups are observed (Theorem 2)	-
Parallel schedule	$\rho(\mathbf{B}) < 1$ (Proposition 4)	both
Parallel schedule with damping	$\rho(r\mathbf{B} + (1-r)\mathbf{I}) < 1$ (Proposition 5)	both
Randomized schedule	Expectation convergence: $\rho\left(\mathbf{PB} + \mathbf{I} - \mathbf{P}\right) < 1$ (Proposition 6) Mean-square convergence: $\rho(\mathbf{\Phi}) < 1$ (Proposition 7)	both

B. Summary of the Studied Schedules

In summary, three categories of message schedules, including serial, parallel, and randomized message schedules, are described for VMP and their convergence properties are analyzed. A comparison of message schedules and their convergence conditions is further illustrated in Table I. Here are some suggestions for choosing a proper message schedule of VMP in practice. If the convergence condition of VMP in damped parallel schedule is satisfied, damped parallel schedule is preferred due to its fast convergence (further optimizing the damping factor r to accelerate the convergence is possible). If the allowable damping factors include positive values in the range $(0, 1]$, we could replace damped parallel schedule by randomized schedule. This would save communication overhead and computational complexity compared to damped VMP, but the convergence speed may not be as fast as that of damped parallel schedule if the optimal r in damped parallel schedule is a negative value. If VMP diverges in parallel schedule with damping and randomized schedule, we can at least adopt serial schedule for VMP. A group serial schedule with all random variables within each message group being conditionally independent when other groups of variables are observed guarantees the convergence of VMP, and accelerates the convergence compared to the basic serial schedule. Finally, if VMP converges in any of these schedules, it is guaranteed to converge to the exact marginal mean vectors.

The complexity of VMP depends on the mean vector update of the variational marginal

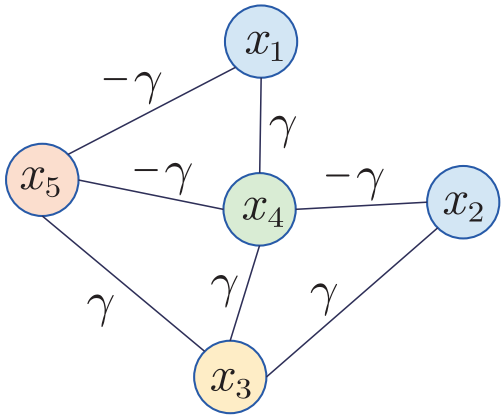

Fig. 1. An example of Gaussian Markov random field.

distributions, and requires $d_i^3 + (2|\mathcal{B}_i| + 1)d_i^2$ multiplications and $|\mathcal{B}_i|$ additions for each variable $\mathbf{x}_i \in \mathbb{R}^{d_i}$. For different schedules, the variables will be updated in different orders or probabilities. Therefore, different schedules have the same complexity for each variable update, and their complexities differ only in the number of iteration to reach convergence. Furthermore, according to the complexity of AMP-type algorithms analyzed in [23, Table II], VMP has a similar complexity as AMP and the Bayes optimal memory AMP.

VI. Numerical Results and Applications

A. Gaussian Markov Random Field

Consider a Gaussian Markov random field as shown in Fig. 1. The corresponding information matrix $\mathbf{J}$ has ones along its diagonal and the (i, j)-th position being the coefficient on the link between x_i and x_j. If there is no link between x_i and x_j, the corresponding position in $\mathbf{J}$ is zero. It can be verified that such a $\mathbf{J}$ is positive-definite if $-0.4550 \leq \gamma \leq 0.3892$. In the following, we take $\gamma = 0.2, 0.36, -0.4$ as examples. As each variable in Fig. 1 is a scalar, we have $d_k = 1$. Moreover, we consider $\mathbf{h} = [1, 1, 1, 1, 1]^T$ and the initialization $\boldsymbol{\mu}^{(0)} = [0, 0, 0, 0, 0]^T$. In group serial schedule, the nodes in Fig. 1 with the same color belong to the same group. These partition groups satisfy the condition that all variables in each group are conditionally independent when the variables in other groups are observed. According to Theorem 2, the VMP under group serial schedule would converge and the means of the converged variational distributions are the exact marginal means. It can be verified that $r\mathbf{B} + (1 - r)\mathbf{I}$ is a Hermitian matrix for any r, which leads to $\|r\mathbf{B} + (1 - r)\mathbf{I}\|_2 = \rho(r\mathbf{B} + (1 - r)\mathbf{I})$. The relative error $e(t) = \frac{\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*\|_2}{\|\boldsymbol{\mu}^*\|_2}$ from the true

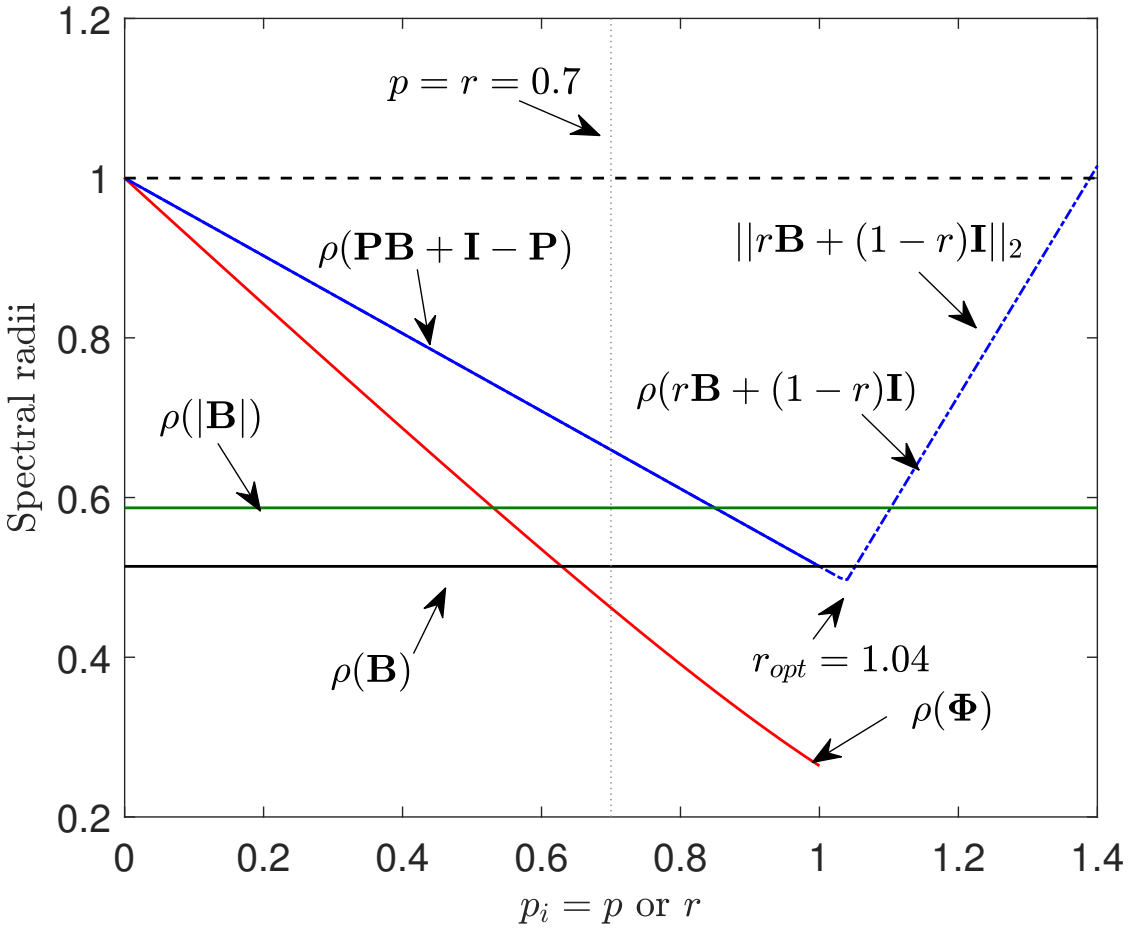

Fig. 2. Spectral radii versus damping factor r or update probability p_i when $\gamma = 0.2$.

marginal mean is used to measure the correctness and convergence of $\boldsymbol{\mu}^{(t)}$ in different schedules. Further illustrations and discussions are as follows.

1) $\gamma = 0.2$*:* Since $\mathbf{J}$ is a diagonally dominant matrix when $\gamma = 0.2$, according to Proposition 2, we must have $\rho(|\mathbf{B}|) < 1$, which is further verified in Fig. 2. Together with Theorem 1, the mean vectors obtained by (8) would converge under any valid message schedule. Therefore, VMP in basic serial schedule (denoted by Basic Serial VMP), group serial schedule (denoted by Group Serial VMP), parallel schedule (denoted by Parallel VMP) and randomized schedule (denoted by Randomized VMP) all converge. Various spectral radii required for convergence verification listed in Table I are also shown in Fig. 2, and it can be seen that all spectral radii are smaller than one.

The relative errors of $\boldsymbol{\mu}^{(t)}$ obtained from VMP under different update schedules are shown in Fig. 3. It can be seen that all considered schedules have their relative errors decrease to

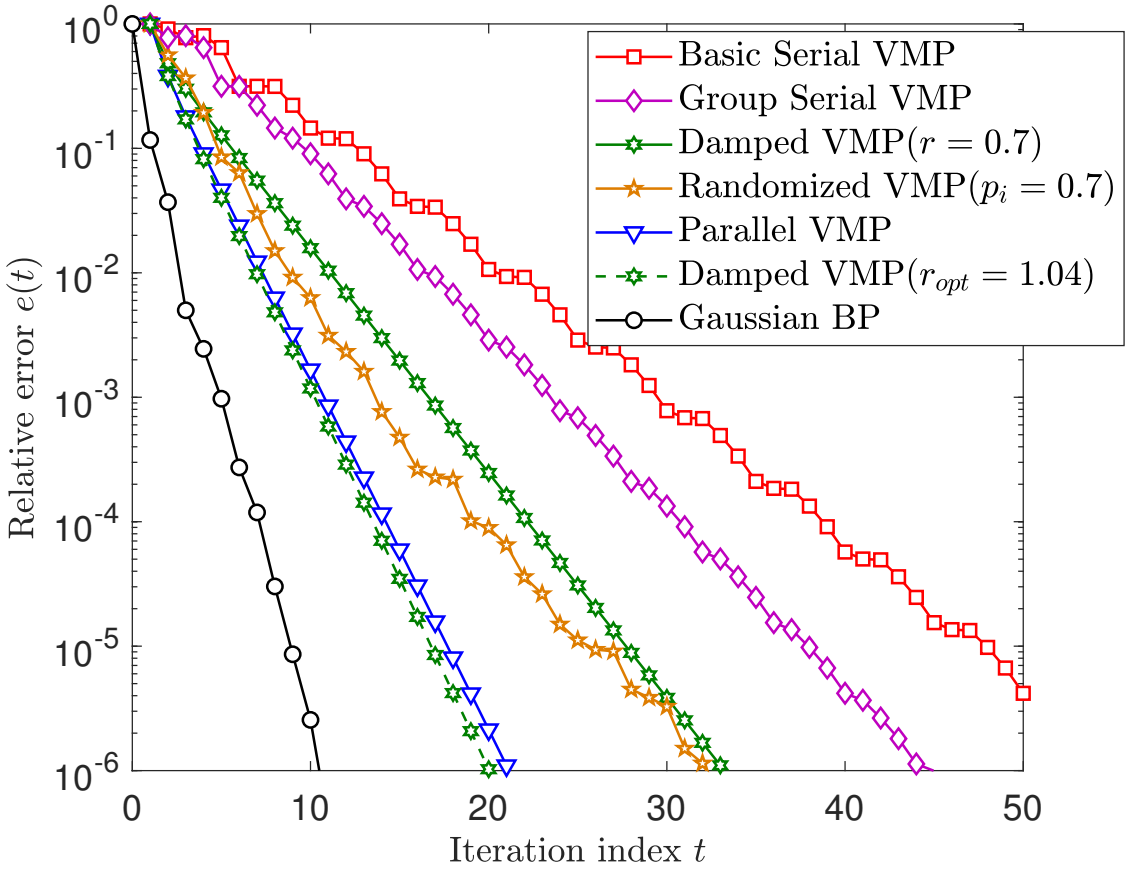

Fig. 3. Relative error $e(t)$ when $\gamma = 0.2$.

very small value when iteration number increases, showing they converge, with the basic serial schedule converges the slowest while the parallel schedule converges the fastest. In between, we have the group serial schedule improves on the basic serial schedule, but still slower than damped parallel schedule and randomized schedule. Furthermore, it is observed that the damped VMP with optimal $r = 1.04$ (identified from Fig. 2 with the smallest spectral radius) converges faster than parallel schedule without damping. Gaussian BP converges much faster than all VMP schedules because Gaussian BP updates both the mean vector and covariance in each iteration, while VMP only updates the mean vector. Note that Gaussian BP is guaranteed to converge in this setting, as $\mathbf{J}$ being diagonally dominant is a sufficient condition for Gaussian BP convergence.

2) $\gamma = 0.36$: From Fig. 4, it is observed that $\rho(|\mathbf{B}|) > 1$ in this setting, thus VMP does not necessarily converge given any valid message schedule. But fortunately, $\rho(\mathbf{B}) < 1$ in this case, so VMP would converge in parallel schedule. Furthermore, Fig. 4 shows that both the expectation and mean-square convergence conditions for randomized VMP are satisfied for any $p \in (0, 1]$.

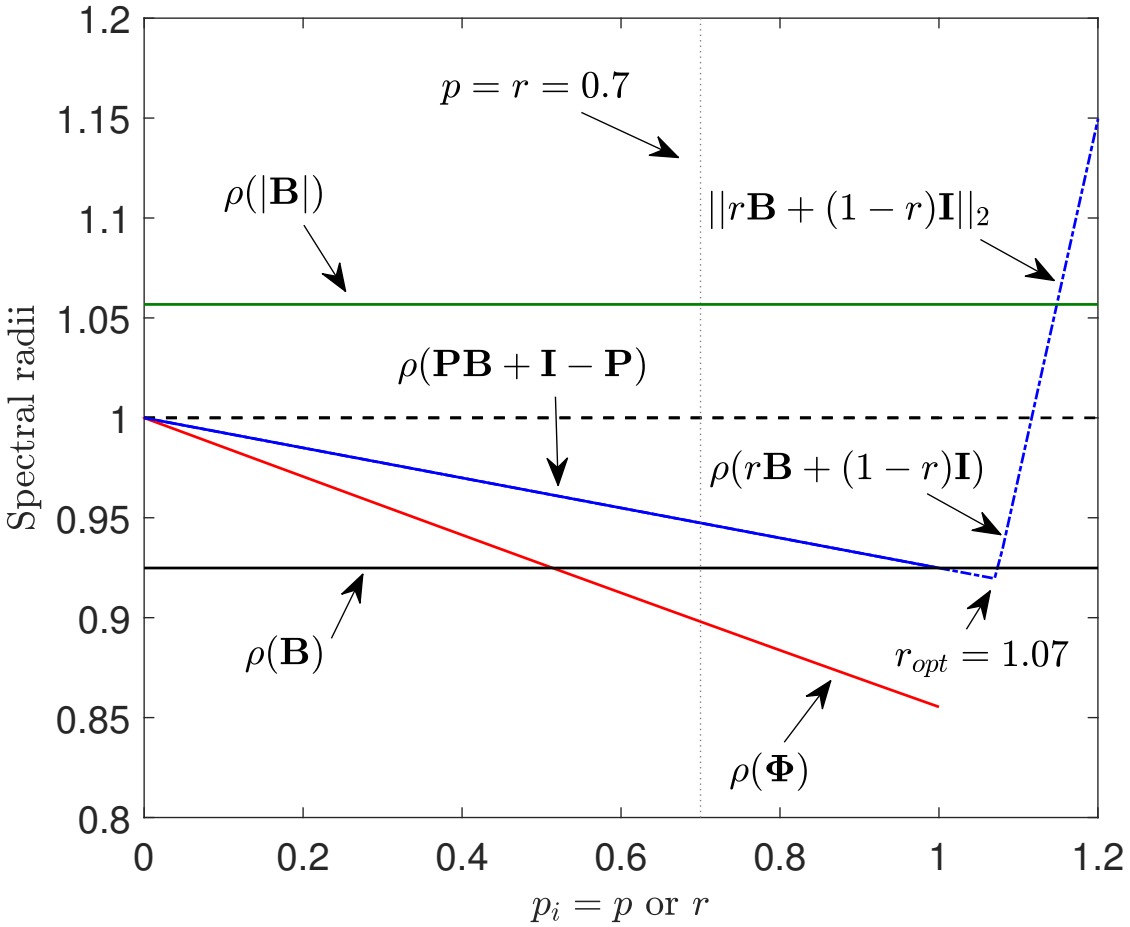

Fig. 4. Spectral radii versus damping factor r or update probability p_i when $\gamma = 0.36$.

The convergence of VMP under various schedules is numerically verified in Fig. 5. Notice that with a direct verification, the necessary and sufficient convergence condition of Gaussian BP in [18, Theorem 1] is not satisfied. This explains that Gaussian BP diverges in Fig. 5.

3) $\gamma = -0.4$: In Fig. 6, since $\rho(|\mathbf{B}|) > 1$, VMP may not be guaranteed to converge in any valid message schedule. Moreover, since $\rho(\mathbf{B}) > 1$ in this case, according to Proposition 4, parallel VMP will diverge. Fortunately, from Fig. 6, we can see that there is a wide range of p_i or r that would make the corresponding spectral radii smaller than one. This means that parallel schedule with damping or randomized schedule would converge if we choose the damping factor or the update probability properly. In particular, according to Proposition 5, the range of r that make the damped VMP convergent is $r \in (0, 0.9864)$. For randomized VMP, Fig. 6 indicates that the expectation convergence requires $p \in (0, 0.9864)$ and the mean-square convergence requires $p \in (0, 0.9815)$. The convergence behaviors of VMP under various schedules are further

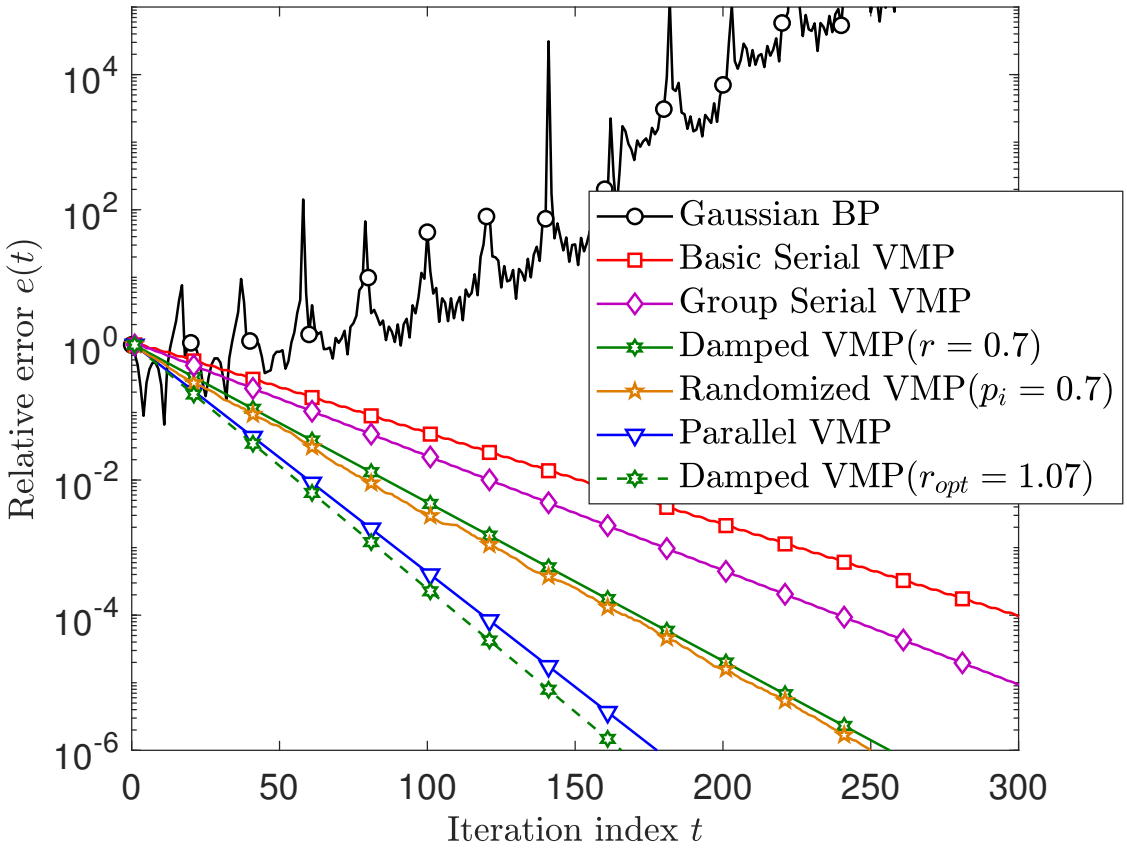

Fig. 5. Relative error $e(t)$ when $\gamma = 0.36$.

demonstrated in Fig. 7, which is consistent with that predicted by the theories. Notice that through numerical calculation, the necessary and sufficient convergence condition of Gaussian BP in [18, Theorem 1] is not satisfied. Thus Gaussian BP does not converge in Fig. 7.

B. Distributed Peer-to-Peer Rating

In social networks, the ratings on product items are likely to affect users' choices. We may share our ratings on certain items while at the same time affected by the ratings posted by others. This corresponds to the peer-to-peer rating problem in (1), whose solution can be obtained by computing the marginal means of $p(\mathbf{x})$ in (2). Thus the peer-to-peer rating problem can be solved by VMP.

To illustrate this application, we generate a network with $n = 100$, $d_i = 2$, ω_{ik} being uniformly distributed from 0 to 1 if there exists a relation between user i and user k and equaling to zero otherwise, and 5% of users do not have initial ratings with $\alpha_i = 0$ while $\alpha_i = 1$ for other users.

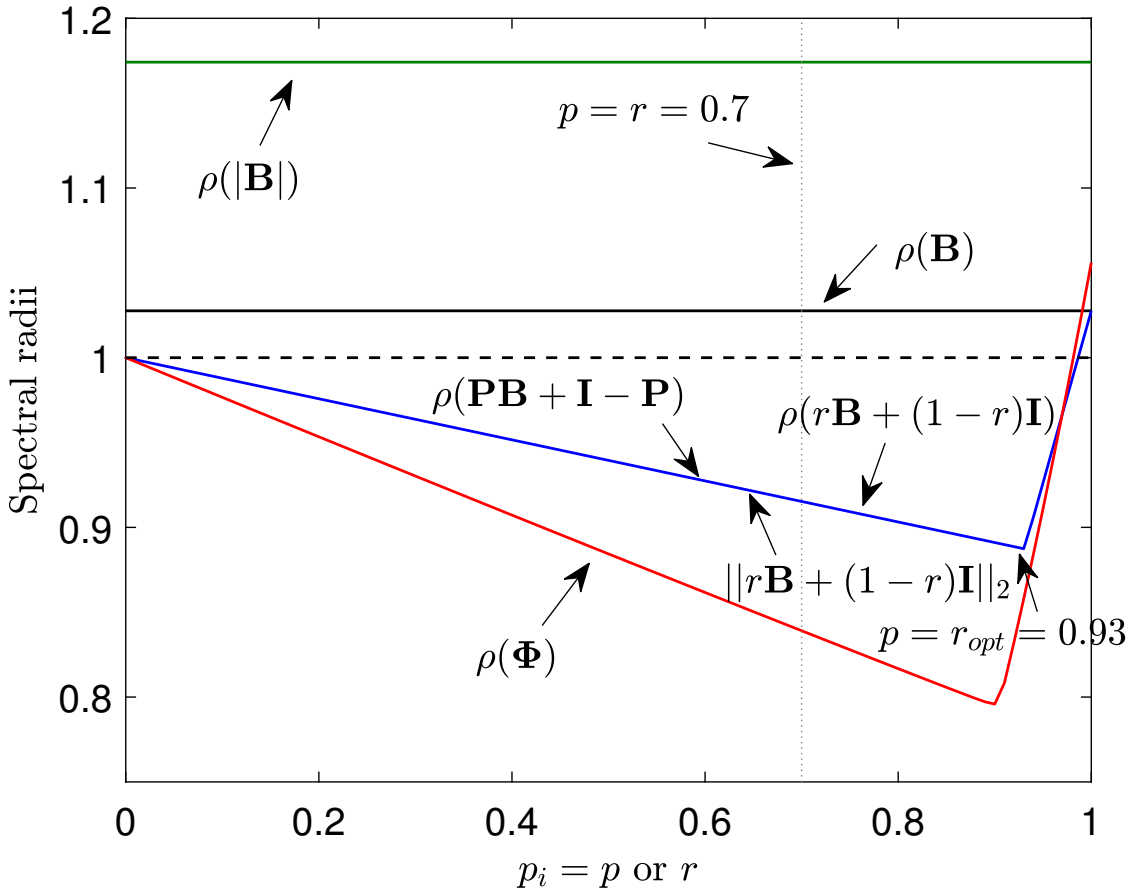

Fig. 6. Spectral radii versus damping factor r or update probability p_i when $\gamma = -0.4$.

Fig. 8 shows an example network of peer-to-peer rating. In Fig. 8, different colors are used to denote the node grouping in group serial schedule, where the nodes having the same color belong to the same group. The grouping is obtained by using a graph coloring algorithm [38].

It can be directly verified that $\mathbf{J}$ for the network in Fig. 8 is a diagonally dominant matrix with diagonal block matrices $\{\mathbf{J}_{ii}\}_{i=1}^{n}$. According to Proposition 2, we obtain $\rho(|\mathbf{B}|) < 1$, which implies that VMP converge in any valid message schedule. This is numerically verified in Fig. 9, which shows the relative error of the estimated marginal mean vectors of VMP in different schedules. In particular, since the network in Fig. 8 is large, basic serial schedule can be very slow. Group serial schedule improves the convergence speed significantly, while the parallel schedule without damping or with damping factor $r = 1.03$ converges faster. On the other hand, for the damped parallel schedules, through direct computation, we obtain the allowable damping factor $r \in (0, 1.2294)$ according to Proposition 5, and the r that minimizes $\|r\mathbf{B} + (1-r)\mathbf{I}\|_2$

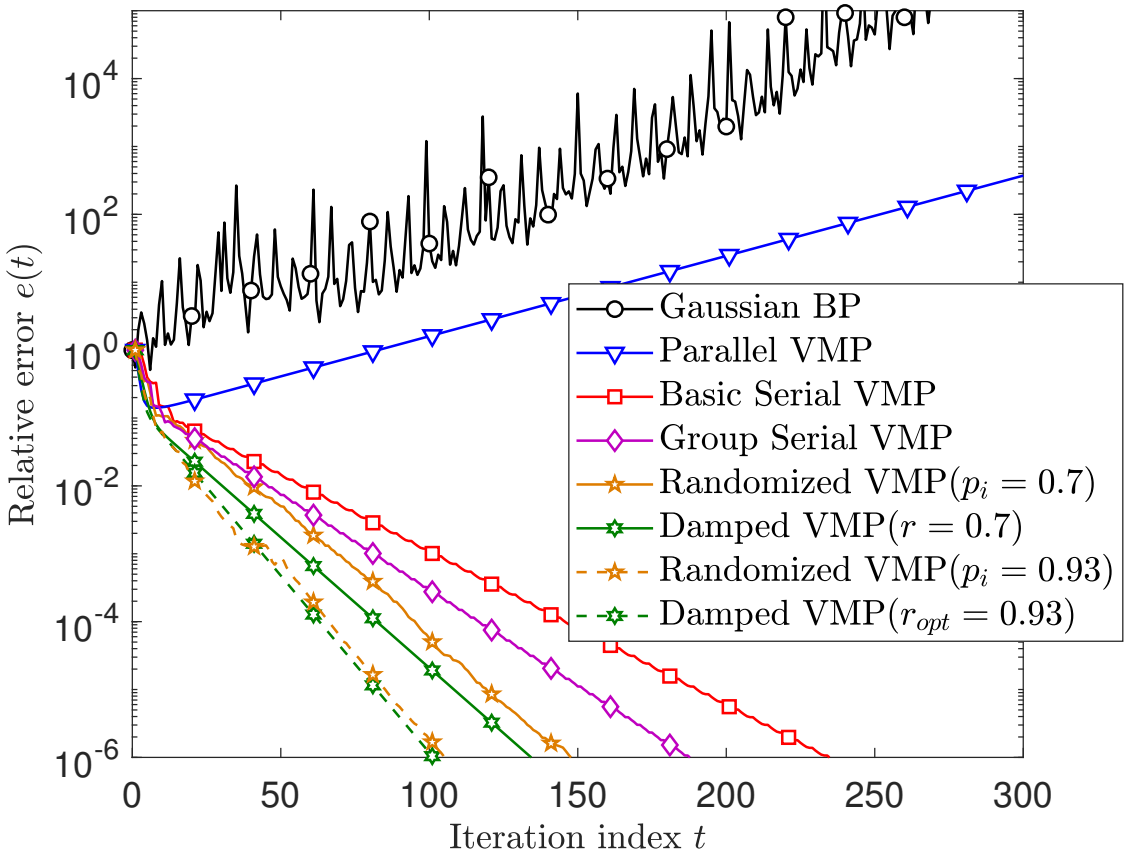

Fig. 7. Relative error $e(t)$ when $\gamma = -0.4$.

is $r = 1.03$. From Fig. 9, it can be seen that the setting $r = 0.75$ converges slower than that of $r = 1.03$, which verifies the prediction from Proposition 9. For the randomized schedule, we considered two variations. The first one uses $p_i = 0.75$ for all i, while the second one draws each p_i from uniform distribution between $(0.5, 1)$. It is observed that randomized VMP schedule with $p_i = 0.75$ for all nodes converges almost at the same speed as that of the damped VMP. Furthermore, it is found that randomized schedule with the same message update probability for all nodes converges slightly faster than diverse message update probabilities among nodes.

C. Distributed Downlink Beamforming

Consider a cellular network of N cells, where users within a cell do not interfere with each other and each particular channel is assigned to only one user. However, user interference may come from other cells. It is assumed that intercell interference only comes from immediate

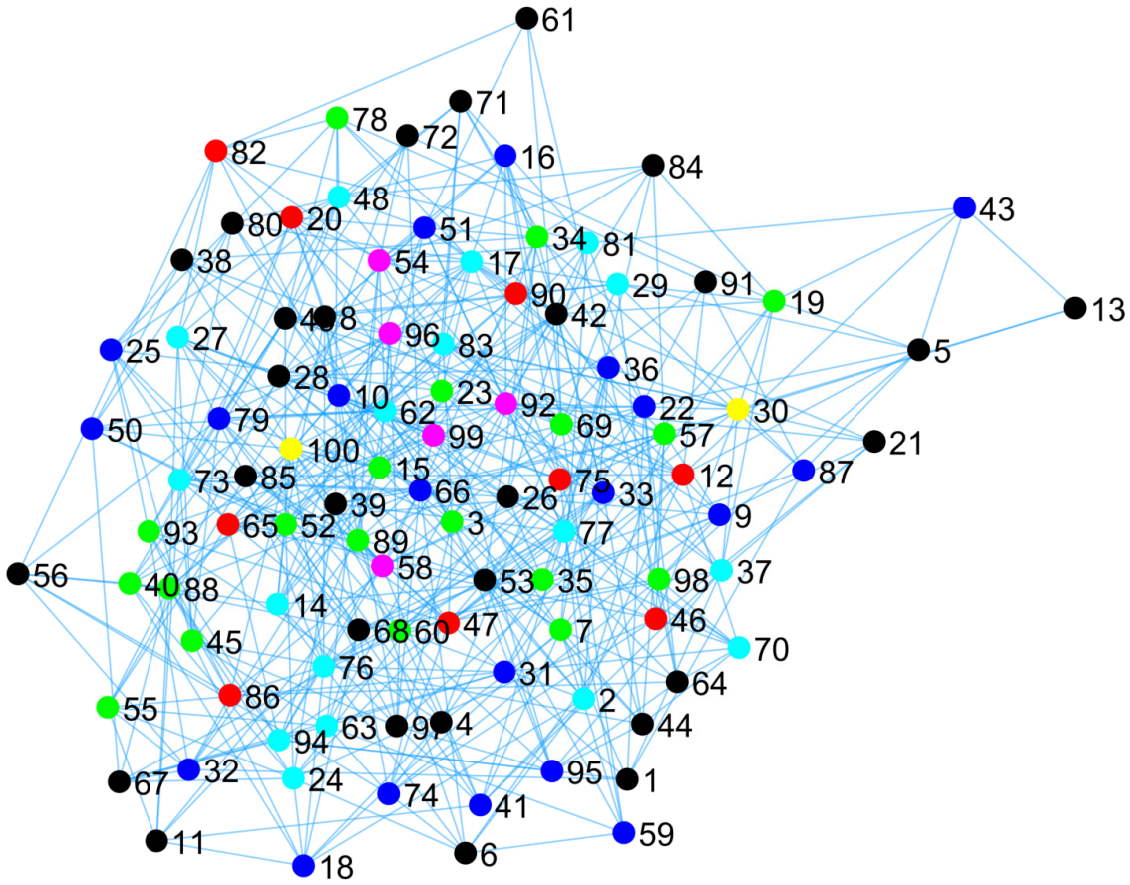

Fig. 8. Peer-to-peer network and group partition.

neighboring cells. In distributed downlink beamforming, multiple base stations (BSs) cooperate to transmit the information to multiple users.

The received signals of users can be written in a vector form

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{26}$$

where $\mathbf{x} \triangleq [x_1, x_2, \cdots, x_N]^T \in \mathbb{C}^N$ with x_n being the transmitted signal at BS n, $\mathbf{y} \triangleq [y_1, y_2, \cdots, y_N]^T \in \mathbb{C}^N$ with y_i being the recevied signal at user i, and $\mathbf{n} \triangleq [n_1, n_2, \cdots, n_N]^T \in \mathbb{C}^N$ with n_i being the complex Gaussian noise with zero mean and variance σ^2. The matrix $\mathbf{H} \in \mathbb{C}^{N\times N}$ denotes the downlink channel matrix with $\mathbf{H}(i,j)$ being the channel coefficient from BS i to user j. For the complex-valued equation in (26), it is equivalent to

$$\underbrace{\begin{bmatrix} \Re\{\mathbf{y}\} \\ \Im\{\mathbf{y}\} \end{bmatrix}}_{\tilde{\mathbf{y}}} = \underbrace{\begin{bmatrix} \Re\{\mathbf{H}\} & -\Im\{\mathbf{H}\} \\ \Im\{\mathbf{H}\} & \Re\{\mathbf{H}\} \end{bmatrix}}_{\tilde{\mathbf{H}}} \underbrace{\begin{bmatrix} \Re\{\mathbf{x}\} \\ \Im\{\mathbf{x}\} \end{bmatrix}}_{\tilde{\mathbf{x}}} + \underbrace{\begin{bmatrix} \Re\{\mathbf{n}\} \\ \Im\{\mathbf{n}\} \end{bmatrix}}_{\tilde{\mathbf{n}}}, \tag{27}$$

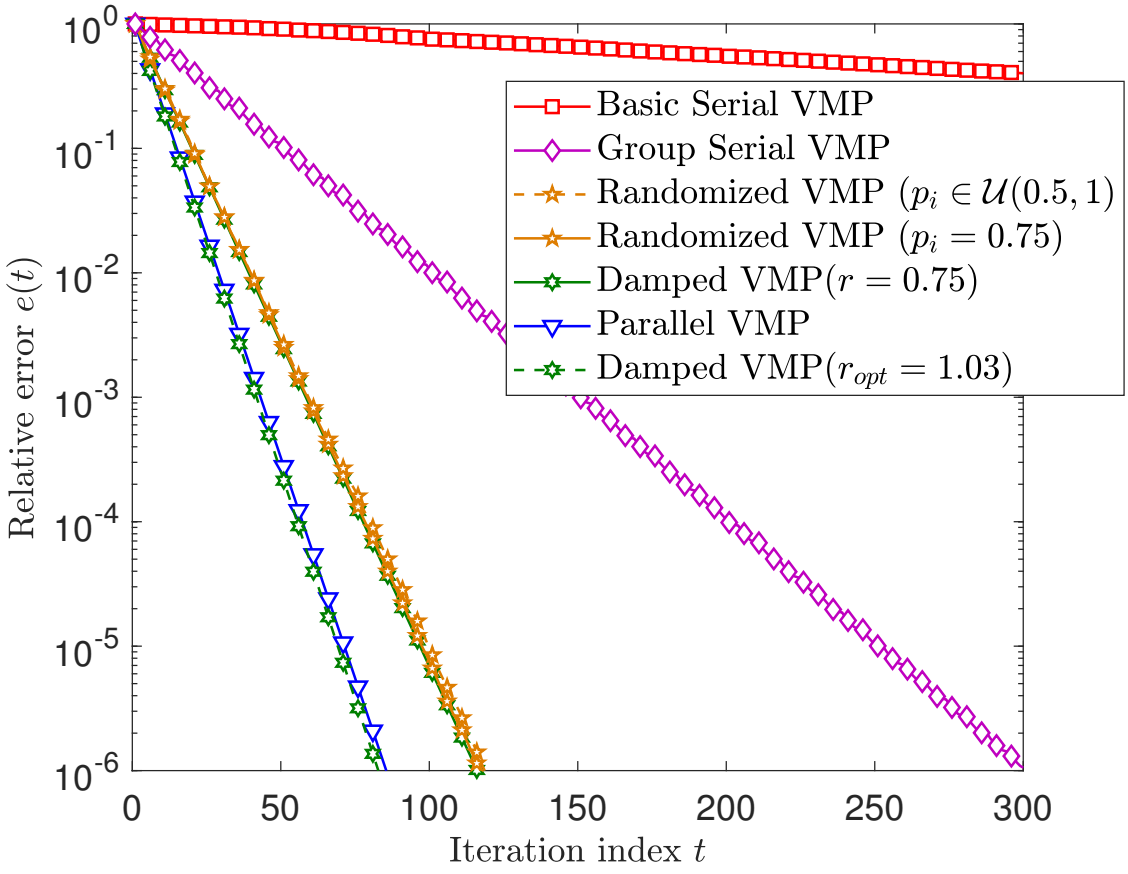

Fig. 9. Relative error $e(t)$ of peer-to-peer rating application.

where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denotes the real part and imaginary part, respectively.

In the simulations, we consider a hexagonal network of $N = 49$ cells and the transmit beamformer with the form $K\tilde{\mathbf{H}}^T(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T + \beta\mathbf{I})^{-1}$ [5]. Thus the transmitted signal $\tilde{\mathbf{x}} = K\tilde{\mathbf{H}}^T(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T + \beta\mathbf{I})^{-1}\tilde{\mathbf{s}}$, where $\tilde{\mathbf{s}} = [\Re\{\mathbf{s}\}^T \ \Im\{\mathbf{s}\}^T]^T$ denotes the intended signal to users. Following [5], it is assumed that $\{s_i\}_{i=1}^N$ are i.i.d. complex Gaussian variables with zero mean and unit variance. Moreover, in the simulations, we set $K = 1$ and $\beta = N/SNR = 7$, where $SNR = \frac{P_t}{\sigma^2} = 7$ with P_t denotes the power constraint imposed on the transmit beamformer. Furthermore, $\mathbf{H}(i, j)$ is a complex Gaussian variable with zero mean and unit variance if there exists a link between BS i and user j. The VMP under different schedules will stop when the relative error is less than 10^{-6} or the maximum iteration 10000 is achieved. The results in this section are obtained through 10000 independent random channel realizations.

To illustrate the convergence probabilities and convergence speeds under different schedules, Fig. 10 shows the cumulative distribution function (CDF) of minimum iteration number that

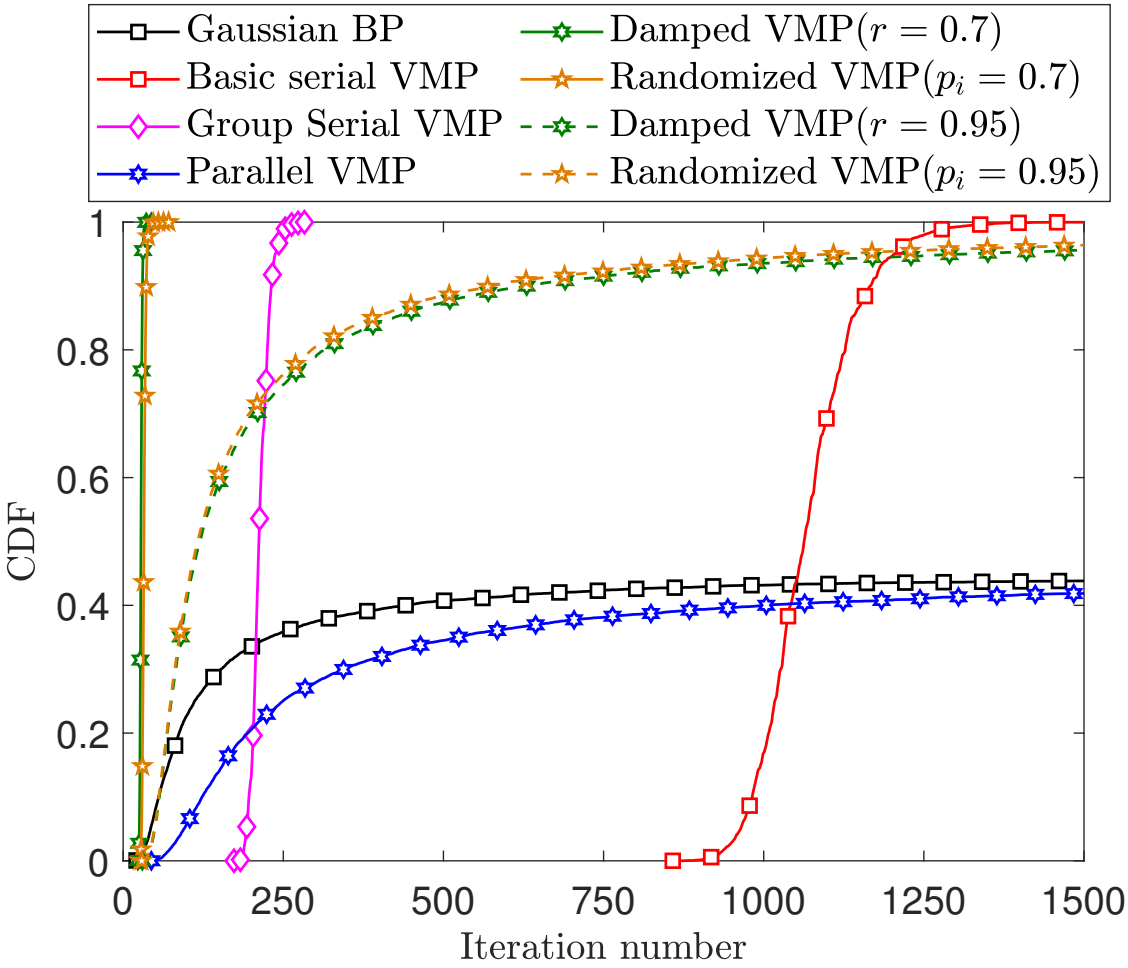

Fig. 10. The CDF of iteration number such that $e(t) < 10^{-6}$.

guarantees $e(t) < 10^{-6}$. As expected, VMP under basic serial schedule and group serial schedule converge with 100% probability. Although VMP in parallel schedule and Gaussian BP could converge faster than VMP in serial schedules, they only converge in less than 50% of the trials. However, for the damped VMP with $r = 0.95$ or randomized VMP with $p_i = 0.95$, they increase the convergence probabilities significantly, albeit cannot reach 100%. On the other hand, if damped VMP with $r = 0.7$ or randomized VMP with $p_i = 0.7$ is used, it is observed that both damped VMP and randomized VMP not only reach 100% convergence probability, but also achieve much faster convergence speeds than parallel VMP. This is due to the proper selection of the damping factor r that accelerates VMP convergence as discussed in Proposition 9 and randomized VMP can be treated as probabilistic damped VMP. The simulation results clearly show the trade-off between chance and speed of convergence in VMP under different message

passing schedules

VII. Conclusions

In this paper, VMP was adopted to compute the marginal means in Gaussian graphical models under different message update schedules. To establish the legitimacy of using VMP, it was proved that if VMP converges, the means of the converged VMP are the exact marginal means. In order to find the optimal variational marginal distributions, serial schedule, parallel schedule and randomized schedule were applied to VMP. The VMP in basic serial schedule is always guaranteed to converge but could be slow in convergence. To accelerate the VMP with convergence guarantee, group serial schedule with all variables in each group being conditionally independent when other groups of variables are observed was proposed. On the other hand, parallel schedule updates each node's message at each iteration, offering an efficient update in distributed and large-scale networks. However, parallel schedule might not converge in a particular Gaussian graphical model. To enlarge the class of models where parallel schedule would converge, damping was applied, and a set of feasible damping factors was derived. Furthermore, randomized schedule, where only a random subset of the variational marginal distributions were updated in each iteration, was introduced to reduce the computation and communication overheads. With the interpretation of probabilistic damping, the necessary and sufficient convergence conditions of random schedule was established in expectation sense and mean square sense. Numerical results and applications were presented to corroborate the convergence properties of various schedules, illustrating the trade-offs between the speed and ease of convergence.

Appendix

A. Proof of the transmitted symbol vector equivalent to the mean vector of a Gaussian model

The mean of a high-dimensional Gaussian distribution with $\mathbf{J} = \mathbf{H}^T\mathbf{H} + \beta\mathbf{I}$ and $\mathbf{h} = K\mathbf{H}^T\mathbf{s}$ can be computed by

$$\mathbf{m} = \mathbf{J}^{-1}\mathbf{h} = (\mathbf{H}^T\mathbf{H} + \beta\mathbf{I})^{-1}K\mathbf{H}^T\mathbf{s} \tag{28}$$

$$= K\beta^{-1}(\beta^{-1}\mathbf{H}^T\mathbf{H} + \mathbf{I})^{-1}\mathbf{H}^T\mathbf{s}. \tag{29}$$

According to Woodbury matrix identity $(\mathbf{I}+\mathbf{UV})^{-1}\mathbf{U}=\mathbf{U}(\mathbf{I}+\mathbf{VU})^{-1}$, by setting $\mathbf{U}=\mathbf{H}^T$ and $\mathbf{V}=\beta^{-1}\mathbf{H}$, we have

$$(\beta^{-1}\mathbf{H}^T\mathbf{H}+\mathbf{I})^{-1}\mathbf{H}^T=\mathbf{H}^T(\beta^{-1}\mathbf{H}\mathbf{H}^T+\mathbf{I})^{-1}. \tag{30}$$

By substituting (30) into (29), we obtain

$$\mathbf{m}=K\beta^{-1}\mathbf{H}^T(\beta^{-1}\mathbf{H}\mathbf{H}^T+\mathbf{I})^{-1}\mathbf{s} \tag{31}$$

$$=K\mathbf{H}^T(\mathbf{H}\mathbf{H}^T+\beta\mathbf{I})^{-1}\mathbf{s}, \tag{32}$$

which is the downlink transmitted symbol vector in distributed beamforming in [5].

B. Proof of Proposition 6

From (11), we obtain the converged mean vector $\boldsymbol{\mu}^*=(\mathbf{I}-\mathbf{B})^{-1}\mathbf{d}$. It can easily verified that $\boldsymbol{\mu}^*$ satisfies (17) as

$$\boldsymbol{\mu}^*=\boldsymbol{\Psi}^{(t)}(\mathbf{B}\boldsymbol{\mu}^*+\mathbf{d})+(\mathbf{I}-\boldsymbol{\Psi}^{(t)})\boldsymbol{\mu}^*. \tag{33}$$

Taking the difference between (17) and (33), we obtain

$$\boldsymbol{\mu}^{(t)}-\boldsymbol{\mu}^*=\left(\boldsymbol{\Psi}^{(t)}\mathbf{B}+\mathbf{I}-\boldsymbol{\Psi}^{(t)}\right)\left(\boldsymbol{\mu}^{(t-1)}-\boldsymbol{\mu}^*\right) \tag{34}$$

$$=\prod_{k=0}^{t-1}\left(\boldsymbol{\Psi}^{(t-k)}\mathbf{B}+\mathbf{I}-\boldsymbol{\Psi}^{(t-k)}\right)\left(\boldsymbol{\mu}^{(0)}-\boldsymbol{\mu}^*\right). \tag{35}$$

Taking the expectation on both sides of (35), we get

$$\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}\left[\boldsymbol{\mu}^{(t)}-\boldsymbol{\mu}^*\right]=\prod_{k=0}^{t-1}\mathbb{E}_{\boldsymbol{\Psi}^{(t-k)}}\left[\boldsymbol{\Psi}^{(t-k)}\mathbf{B}+\mathbf{I}-\boldsymbol{\Psi}^{(t-k)}\right]\left(\boldsymbol{\mu}^{(0)}-\boldsymbol{\mu}^*\right). \tag{36}$$

This is due to the i.i.d. property of $\{\psi_i^{(t)}\}_{i=1}^n$ for all t. Since $\mathbb{E}_{\boldsymbol{\Psi}^{(t-k)}}[\boldsymbol{\Psi}^{(t-k)}\mathbf{B}+\mathbf{I}-\boldsymbol{\Psi}^{(t-k)}]=\mathbb{E}_{\boldsymbol{\Psi}^{(t-k)}}[\boldsymbol{\Psi}^{(t-k)}]\mathbf{B}+\mathbf{I}-\mathbb{E}_{\boldsymbol{\Psi}^{(t-k)}}[\boldsymbol{\Psi}^{(t-k)}]$ and $\mathbb{E}_{\boldsymbol{\Psi}^{(t-k)}}[\boldsymbol{\Psi}^{(t-k)}]=\mathbf{P}$ with $\mathbf{P}=\text{blkdiag}(p_1\mathbf{I}_1,p_2\mathbf{I}_2,\cdots,p_n\mathbf{I}_n)$, we obtain $\mathbb{E}_{\boldsymbol{\Psi}^{(t-k)}}[\boldsymbol{\Psi}^{(t-k)}\mathbf{B}+\mathbf{I}-\boldsymbol{\Psi}^{(t-k)}]=\mathbf{P}\mathbf{B}+\mathbf{I}-\mathbf{P}$. Further taking the limit on both sides of (36), $\lim_{t\to\infty}\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}\left[\boldsymbol{\mu}^{(t)}\right]-\boldsymbol{\mu}^*=\lim_{t\to\infty}(\mathbf{P}\mathbf{B}+\mathbf{I}-\mathbf{P})^t\left(\boldsymbol{\mu}^{(0)}-\boldsymbol{\mu}^*\right)$. Therefore, $\lim_{t\to\infty}\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}\left[\boldsymbol{\mu}^{(t)}\right]$ converges to $\boldsymbol{\mu}^*$ for any initialization $\boldsymbol{\mu}^{(0)}\in\mathbb{R}^N$ if and only if $\lim_{t\to\infty}(\mathbf{P}\mathbf{B}+\mathbf{I}-\mathbf{P})^t=\mathbf{0}$, which is satisfied if and only if $\rho(\mathbf{P}\mathbf{B}+\mathbf{I}-\mathbf{P})<1$.

C. Proof of Proposition 7

First, taking the difference between (17) and (33), we obtain

$$\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^* = \left(\boldsymbol{\Psi}^{(t)}\mathbf{B} + \mathbf{I} - \boldsymbol{\Psi}^{(t)}\right)\left(\boldsymbol{\mu}^{(t-1)} - \boldsymbol{\mu}^*\right). \tag{37}$$

Denoting $\boldsymbol{\Xi} \triangleq \boldsymbol{\Psi}^{(t)}\mathbf{B} + \mathbf{I} - \boldsymbol{\Psi}^{(t)}$ and further taking the Kronecker product on both sides of (37), we get

$$(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*) \otimes (\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*) = (\boldsymbol{\Xi} \otimes \boldsymbol{\Xi})\left((\boldsymbol{\mu}^{(t-1)} - \boldsymbol{\mu}^*) \otimes (\boldsymbol{\mu}^{(t-1)} - \boldsymbol{\mu}^*)\right). \tag{38}$$

Taking the expectation on both sides of (38) with respect to $\{\boldsymbol{\Psi}^{(k)}\}_{k=0}^{t}$, it leads to

$$\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}\left[\left(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*\right) \otimes \left(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*\right)\right] = \mathbb{E}_{\boldsymbol{\Psi}^{(t)}}[\boldsymbol{\Xi}\otimes\boldsymbol{\Xi}]\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t-1}}\left[(\boldsymbol{\mu}^{(t-1)}-\boldsymbol{\mu}^*)\otimes(\boldsymbol{\mu}^{(t-1)}-\boldsymbol{\mu}^*)\right], \tag{39}$$

where $\boldsymbol{\Xi}$ only depends on $\boldsymbol{\Psi}^{(t)}$ and $\boldsymbol{\mu}^{(t-1)}-\boldsymbol{\mu}^*$ depends on $\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t-1}$. Since the iterative equation of $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\boldsymbol{\mu}^{(t)}-\boldsymbol{\mu}^*)\otimes(\boldsymbol{\mu}^{(t)}-\boldsymbol{\mu}^*)]$ in (39) is a linear equation, $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\boldsymbol{\mu}^{(t)}-\boldsymbol{\mu}^*)\otimes(\boldsymbol{\mu}^{(t)}-\boldsymbol{\mu}^*)]$ converges to zero for any initialization $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$ if and only if $\rho(\mathbb{E}_{\boldsymbol{\Psi}^{(t)}}[\boldsymbol{\Xi}\otimes\boldsymbol{\Xi}]) < 1$ [35, Proposition 2.6.1].

Next, we prove that $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*\|_2^2]$ converges to zero for any initialization $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$ if and only if $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*) \otimes (\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*)]$ converges to zero. On one hand, if $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*) \otimes (\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*)] = \mathbf{0}$, then $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\mu_i^{(t)} - \mu_i^*)(\mu_j^{(t)} - \mu_j^*)] = 0$ for all $i, j = 1, 2, \cdots, N$. By taking $i = j$, we obtain $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\mu_i^{(t)}-\mu_i^*)^2] = 0$ for all $i = 1, 2, \cdots, N$, which implies $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[\|\boldsymbol{\mu}^{(t)}-\boldsymbol{\mu}^*\|^2] = 0$. On the other hand, if $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[\|\boldsymbol{\mu}^{(t)}-\boldsymbol{\mu}^*\|^2] = 0$, it implies $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\mu_i^{(t)}-\mu_i^*)^2] = 0$ for all $i = 1, 2, \cdots, N$. Using the Cauchy–Schwarz inequality, we have $|\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\mu_i^{(t)}-\mu_i^*)(\mu_j^{(t)}-\mu_j^*)|^2 \le \mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\mu_i^{(t)}-\mu_i^*)^2]\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\mu_j^{(t)}-\mu_j^*)^2] = 0$, which implies $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\mu_i^{(t)}-\mu_i^*)(\mu_j^{(t)}-\mu_j^*)] = 0$. Therfore, we have $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*) \otimes (\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*)] = \mathbf{0}$.

With the results from the above two paragraphs, it can be concluded that $\mathbb{E}_{\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{t}}[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^*\|_2^2]$ converges to zero for any initialization $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$ if and only if $\rho(\mathbb{E}_{\boldsymbol{\Psi}^{(t)}}[\boldsymbol{\Xi}\otimes\boldsymbol{\Xi}]) < 1$, i.e.,

$$\rho\left(\mathbb{E}_{\boldsymbol{\Psi}^{(t)}}\left[(\boldsymbol{\Psi}^{(t)}\mathbf{B}+\mathbf{I}-\boldsymbol{\Psi}^{(t)})\otimes(\boldsymbol{\Psi}^{(t)}\mathbf{B}+\mathbf{I}-\boldsymbol{\Psi}^{(t)})\right]\right) < 1, \tag{40}$$

where $\mathbb{E}_{\boldsymbol{\Psi}^{(t)}}[(\boldsymbol{\Psi}^{(t)}\mathbf{B} + \mathbf{I} - \boldsymbol{\Psi}^{(t)}) \otimes (\boldsymbol{\Psi}^{(t)}\mathbf{B} + \mathbf{I} - \boldsymbol{\Psi}^{(t)})] = \mathbb{E}_{\boldsymbol{\Psi}^{(t)}}[\boldsymbol{\Psi}^{(t)} \otimes \boldsymbol{\Psi}^{(t)}]((\mathbf{B} - \mathbf{I}) \otimes (\mathbf{B} - \mathbf{I})) + (\mathbb{E}_{\boldsymbol{\Psi}^{(t)}}[\boldsymbol{\Psi}^{(t)}](\mathbf{B}-\mathbf{I})) \otimes \mathbf{I} + \mathbf{I} \otimes (\mathbb{E}_{\boldsymbol{\Psi}^{(t)}}[\boldsymbol{\Psi}^{(t)}](\mathbf{B}-\mathbf{I})) + \mathbf{I} \otimes \mathbf{I}$. Further, we obtain $\mathbb{E}_{\boldsymbol{\Psi}^{(t)}}[\boldsymbol{\Psi}^{(t)} \otimes \boldsymbol{\Psi}^{(t)}] = \text{blkdiag}(\mathbf{I}_1 \otimes \boldsymbol{\Gamma}_1, \mathbf{I}_2 \otimes \boldsymbol{\Gamma}_2, \cdots, \mathbf{I}_n \otimes \boldsymbol{\Gamma}_n)$, where $\boldsymbol{\Gamma}_k$ is a $N \times N$ diagonal matrix with

the nonzero elements $\mathbf{\Gamma}_k(i,i) = p_k$ for all $i = 1 + \sum_{k'=1}^{k-1} d_{k'}, \cdots, \sum_{k'=1}^{k} d_{k'}$ and otherwise $\mathbf{\Gamma}_k(i,i) = p_k p_j$ with $i = 1 + \sum_{k'=1}^{j-1} d_{k'}, \cdots, \sum_{k'=1}^{j} d_{k'}$ and $j \neq k$. Further with $\mathbb{E}_{\mathbf{\Psi}^{(t)}}[\mathbf{\Psi}^{(t)}] = \mathbf{P}$, we get $\mathbb{E}_{\mathbf{\Psi}^{(t)}}[(\mathbf{\Psi}^{(t)}\mathbf{B}+\mathbf{I}-\mathbf{\Psi}^{(t)})\otimes(\mathbf{\Psi}^{(t)}\mathbf{B}+\mathbf{I}-\mathbf{\Psi}^{(t)})] = \text{blkdiag}(\mathbf{I}_1\otimes\mathbf{\Gamma}_1, \mathbf{I}_2\otimes\mathbf{\Gamma}_2, \cdots, \mathbf{I}_n\otimes\mathbf{\Gamma}_n)((\mathbf{B}-\mathbf{I})\otimes(\mathbf{B}-\mathbf{I}))+(\mathbf{PB}-\mathbf{P})\otimes\mathbf{I}+\mathbf{I}\otimes(\mathbf{PB}-\mathbf{P})+\mathbf{I}\otimes\mathbf{I}$. Therefore, $\mathbb{E}_{\{\mathbf{\Psi}^{(k)}\}_{k=1}^{t}}[\|\boldsymbol{\mu}^{(l)}-\boldsymbol{\mu}^{*}\|_2^2]$ converges to zero for any initialization $\boldsymbol{\mu}^{(0)} \in \mathbb{R}^N$ if and only if $\rho(\text{blkdiag}(\mathbf{I}_1\otimes\mathbf{\Gamma}_1, \mathbf{I}_2\otimes\mathbf{\Gamma}_2, \cdots, \mathbf{I}_n\otimes\mathbf{\Gamma}_n)((\mathbf{B}-\mathbf{I})\otimes(\mathbf{B}-\mathbf{I}))+(\mathbf{PB}-\mathbf{P})\otimes\mathbf{I}+\mathbf{I}\otimes(\mathbf{PB}-\mathbf{P})+\mathbf{I}\otimes\mathbf{I}) < 1$.

REFERENCES

[1] D. Bickson, D. Malkhi, and L. Zhou, "Peer-to-peer rating," in *Proc. 7th IEEE Int. Conf. Peer-to-Peer Comput.*, 2007, pp. 1-8.

[2] C. C. Moallemi and B. Van Roy, "Consensus propagation," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4753-4766, Nov. 2006.

[3] 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, May 2009.

[4] J. Du and Y. -C. Wu, "Network-wide distributed carrier frequency offsets estimation and compensation via belief propagation," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5868-5877, Dec.1, 2013.

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

[6] L. Liu, C. Yuen, Y. L. Guan, Y. Li and Y. Su, "Convergence analysis and assurance for Gaussian message passing iterative detector in massive MU-MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6487-6501, Sept. 2016.

[7] C. Fan, X. Yuan and Y. J. Zhang, "Scalable uplink signal detection in C-RANs via randomized Gaussian message passing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5187-5200, Aug. 2017.

[8] M. Cosovic and D. Vukobratovic, "Distributed Gauss–Newton method for state estimation using belief propagation," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 648-658, Jan. 2019.

[9] B. Li, N. Wu, Y. C. Wu, and Y. Li, "Convergence-guaranteed parametric Bayesian distributed cooperative localization," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8179-8192, Oct. 2022.

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

[11] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, K. S. Lawrence, "An introduction to variational methods for graphical models," *Mach. learn.*, no. 37, pp. 183-233, Nov. 1999.

[12] T. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 17th Conf. Uncertain. Artif. Intell.*, 2001, pp. 362-369.

[13] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[14] M. Luo, Q. Guo, M. Jin, Y. C. Eldar, D. Huang and X. Meng, "Unitary approximate message passing for sparse Bayesian learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 6023-6039, Sept. 2021.

[15] F. Tian, L. Liu and X. Chen, "Generalized memory approximate message passing for generalized linear model," *IEEE Trans. Signal Process.*, vol. 70, pp. 6404-6418, 2022.

[16] D. M. Malioutov, J. Johnson, and A. Willsky, "Walk-sums and belief propagation in Gaussian graphical models," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 2031-2064, Oct. 2006.

[17] J. Du, S. D. Ma, Y. C. Wu, S. Kar, and J. M. F. Moura, "Convergence analysis of distributed inference with vector-valued Gaussian belief propagation," *J. Mach. Learn. Res.*, vol. 18, no. 172, pp. 1-38, Apr. 2018.

[18] B. Li and Y. C. Wu, "Convergence of Gaussian belief propagation under general pairwise factorization: connecting Gaussian MRF with pairwise linear Gaussian model," *J. Mach. Learn. Res.*, vol. 20, no. 144, pp. 1-30, Oct. 2019.

[19] S. Rangan, P. Schniter, A. K. Fletcher and S. Sarkar, "On the convergence of approximate message passing with arbitrary matrices," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5339-5351, Sept. 2019.

[20] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020-2033, 2017.

[21] S. Rangan, P. Schniter and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664-6684, Oct. 2019.

[22] L. Liu, S. Huang, and B. M. Kurkoski, "Sufficient statistic memory approximate message passing, " in *Proc. IEEE Int. Symp. Inf. Theory*, 2021, pp. 1378-1383.

[23] L. Liu, S. Huang, and B. M. Kurkoski, "Memory AMP, "*IEEE Trans. Inf. Theory*, vol. 68, no. 12, pp. 8015-8039, Dec. 2022.

[24] K. Takeuchi, "On the convergence of orthogonal/vector AMP: long-memory message passing strategy," *IEEE Trans. Inf. Theory*, vol. 68, no. 12, pp. 8121-8138, Dec. 2022.

[25] J. Winn, and C. M. Bishop, "Variational message passing," *J. Mach. Learn. Res.*, vol. 6, no. 23, pp. 661-694, Apr. 2005.

[26] J. Dauwels, "On variational message passing on factor graphs," In *Proc. IEEE Int. Symp. Inf. Theory*, 2007, pp. 2546-2550.

[27] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: a review for statisticians," *J. Am. Stat. Assoc.*, vol. 112, no. 518, pp. 859-877, Jul. 2017.

[28] C. M. Bishop, Pattern recognition and machine learning. Springer, 2006.

[29] C. Pedersen, T. Pedersen, B. H. Fleury, "A variational message passing algorithm for sensor self-localization in wireless networks," in *Proc. IEEE Int. Symp. Inf. Theory*, 2011, pp. 2158-2162.

[30] A. Liu, L. Lian, V. Lau, G. Liu and M. -J. Zhao, "Cloud-assisted cooperative localization for vehicle platoons: a turbo approach," *IEEE Trans. Signal Process.*, vol. 68, pp. 605-620, 2020.

[31] L. Cheng, Y. C. Wu, and H. V. Poor, "Scaling probabilistic tensor canonical polyadic decomposition to massive data," *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5534-5548, Nov. 2018.

[32] L. Yang, J. Fang, H. Duan, H. Li and B. Zeng, "Fast low-rank Bayesian matrix completion with hierarchical gaussian prior models," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2804-2817, Jun. 2018.

[33] S. S. Thoota and C. R. Murthy, "Massive MIMO-OFDM systems with low resolution ADCs: Cramér-Rao bound, sparse channel estimation, and soft symbol decoding," *IEEE Trans. Signal Process.*, vol. 70, pp. 4835-4850, 2022.

[34] A. Amiri, S. Rezaie, C. N. Manchón and E. de Carvalho, "Distributed receiver processing for extra-large MIMO arrays: a message passing approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2654-2667, Apr. 2022.

[35] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[36] Y. Li, S. Ma, G. Yang and K. Wong, "Robust localization for mixed LOS/NLOS environments with anchor uncertainties," *IEEE Trans. Commun.*, vol. 68, no.7, pp. 4507-4521, Jul. 2020.

[37] Y. Li, S. Ma, G. Yang and K. Wong, "Secure localization and velocity estimation in mobile IoT networks with malicious attacks," *IEEE Internet of Things Journal*, vol. 8, no.8, pp. 6878-6892, Apr. 2021.

[38] J. Postigo, J. Soto-Begazo, V. R. Fiorela, G. M. Picha, R. Flores-Quispe and Y. Velazco-Paredes, "Comparative analysis of the main graph coloring algorithms," in *Proc. IEEE Colombian Conf. Commun. Comput.*, 2021, pp.1-6.

Bin Li (Member, IEEE) received the B.S. degree in information engineering and the M.S. degree in information and communication engineering from Beijing Institute of Technology, Beijing, China, in 2012 and 2015, respectively, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong, Hong Kong, in 2019. Currently, he is an Associate Professor with the School of Information and Electronics, Beijing Institute of Technology. His research interests include signal processing, wireless communications, and machine learning. He served as an Editorial Board Member of IEICE Transactions on Communications and KSII Transactions on Internet and Information Systems.

Nan Wu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Beijing Institute of Technology (BIT), Beijing, China, in 2003, 2005, and 2011, respectively. From 2008 to 2009, he was a visiting Ph.D. student with the Department of Electrical Engineering, The Pennsylvania State University, USA. He is currently a Professor with the School of Information and Electronics, BIT. His research interest includes signal processing in wireless communication networks. He serves as an Editorial Board Member of IEEE Wireless Communications Letters, IEEE Access, International Journal of Electronics and Communications, and KSII Transactions on Internet and Information Systems. He was a recipient of the National Excellent Doctoral Dissertation Award by MOE of China in 2013.

Yik-Chung Wu (Senior Member, IEEE) received the B.Eng. (EEE) and M.Phil. degrees from The University of Hong Kong (HKU) in 1998 and 2001, respectively, and the Ph.D. degree from Texas A&M University, College Station, in 2005. From 2005 to 2006, he was with Thomson Corporate Research, Princeton, NJ, USA, as a Member of Technical Staff. Since 2006, he has been with HKU, where he is currently as an Associate Professor. He was a Visiting Scholar at Princeton University in 2015 and 2017. His research interests include signal processing, machine learning, and communication systems. He served as an Editor for IEEE COMMUNICATIONS LETTERS and IEEE TRANSACTIONS ON COMMUNICATIONS. He is currently a Senior Area Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING, an Associate Editor for IEEE WIRELESS COMMUNICATIONS LETTERS, and Journal of Communications and Networks.