Consensus Propagation

Ciamac C. Moallemi and Benjamin Van Roy, *Member, IEEE*

***Abstract*—We propose *consensus propagation*, an asynchronous distributed protocol for averaging numbers across a network. We establish convergence, characterize the convergence rate for regular graphs, and demonstrate that the protocol exhibits better scaling properties than *pairwise averaging*, an alternative that has received much recent attention. Consensus propagation can be viewed as a special case of belief propagation, and our results contribute to the belief propagation literature. In particular, beyond singly-connected graphs, there are very few classes of relevant problems for which belief propagation is known to converge.**

***Index Terms*—Belief propagation, distributed averaging, distributed consensus, distributed signal processing, Gaussian Markov random fields, max-product algorithm, message-passing algorithms, min-sum algorithm, sum-product algorithm.**

I. Introduction

CONSIDER a network of n nodes in which the ith node observes a real number $y_i \in \mathbb{R}$ and aims to compute the average $\bar{y} = \sum_{i=1}^{n} y_i/n$. The design of scalable distributed protocols for this purpose has received much recent attention and is motivated by a variety of potential needs. In both wireless sensor and peer-to-peer networks, for example, there is interest in simple protocols for computing aggregate statistics (see, e.g., [1]–[7]), and averaging enables computation of several important ones. Further, averaging serves as a primitive in the design of more sophisticated distributed information processing algorithms. For example, a maximum-likelihood estimate can be produced by an averaging protocol if each node's observations are linear in variables of interest and noise is Gaussian [8]. Reference [9] considers an averaging problem with applications to load balancing and clock synchronization. As another example, averaging protocols are central to policy-gradient-based methods for distributed optimization of network performance [10].

In this paper, we propose and analyze a new protocol—**consensus propagation**—for distributed averaging. The protocol can operate asynchronously and requires only simple iterative computations at individual nodes and communication of parsimonious messages between neighbors. There is no central hub that aggregates information. Each node only needs to be aware of its neighbors—no further information about the network topology is required. There is no need for construction of a specially structured overlay network such as a spanning tree. It is worth discussing two previously proposed and well-studied protocols that also exhibit these features.

1) **(Probabilistic Counting)** This protocol is based on ideas from [11] for counting distinct elements of a database and in [12] was adapted to produce a protocol for averaging. The outcome is random, with variance that becomes arbitrarily small as the number of nodes grows. However, for moderate numbers of nodes, say tens of thousands, high variance makes the protocol impractical. The protocol can be repeated in parallel and results combined in order to reduce variance, but this leads to onerous memory and communication requirements. Convergence time of the protocol is analyzed in [13].
2) **(Pairwise Averaging)** In this protocol, each node maintains its current estimate of the average, and each time a pair of nodes communicate, they revise their estimates to both take on the mean of their previous estimates. Convergence of this protocol in a very general model of asynchronous computation and communication was established in [14], and there has been significant follow-on work, a recent sample of which is [15]. Recent work [16], [17] has studied the convergence rate and its dependence on network topology and how pairs of nodes are sampled. Here, sampling is governed by a certain doubly stochastic matrix, and the convergence rate is characterized by its second-largest eigenvalue.

In terms of convergence rate, probabilistic counting dominates both pairwise averaging and consensus propagation in the asymptotic regime. However, consensus propagation and pairwise averaging are likely to be more effective in moderately sized networks (up to hundreds of thousands or perhaps even millions of nodes). Further, these two protocols are both naturally studied as iterative matrix algorithms. As such, pairwise averaging will serve as a baseline to which we will compare consensus propagation.

Consensus propagation is a simple algorithm with an intuitive interpretation. It can also be viewed as an asynchronous distributed version of belief propagation as applied to approximation of conditional distributions in a Gaussian Markov random field. When the network of interest is singly connected, prior results about belief propagation imply convergence of consensus propagation. However, in most cases of interest, the network is not singly connected and prior results have little to say about

Manuscript received July 6, 2005; revised March 19, 2006. This work was supported in part by the National Science Foundation under Grant IIS-0428868 and a supplement to Grant ECS-9985229 provided by the Management of Knowledge Intensive Dynamic Systems Program (MKIDS). The work of C. C. Moallemi was supported by a Benchmark Stanford Graduate Fellowship. The material in this paper was presented in part at the Neural Information Processing Systems (NIPS) Conference, Vancouver, BC, Canada, December 2005.

C. C. Moallemi is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: ciamac@stanford.edu).

B. Van Roy is with the Department of Management Science and Engineering and the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: bvr@stanford.edu).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2006.883539

convergence. In particular, Gaussian belief propagation on a graph with cycles is not guaranteed to converge, as demonstrated by numerical examples in [18].

In fact, there are very few relevant cases where belief propagation on a graph with cycles is known to converge. Some fairly general sufficient conditions have been established [19]–[22], but these conditions are abstract and it is difficult to identify interesting classes of problems that meet them. One simple case where belief propagation is guaranteed to converge is when the graph has only a single cycle and variables have finite support [23]–[25]. In its use for decoding low-density parity-check codes, though convergence guarantees have not been made, [26] establishes desirable properties of iterates, which hold with high probability. Recent work proposes the use of belief propagation to solve maximum-weight matching problems and proves convergence in that context [27]. In the Gaussian case, [18], [28] provide sufficient conditions for convergence, but these conditions are difficult to interpret and do not capture situations that correspond to consensus propagation. Since this paper was submitted for publication, a general class of results has been developed for the convergence of Gaussian belief propagation [29], [30]. These results can be viewed as a generalization of the convergence results in this paper. However, they do not address the issue of rate of convergence.

With this background, let us discuss the primary contributions of this paper.

1) We propose consensus propagation, a new asynchronous distributed protocol for averaging.
2) We prove that consensus propagation converges even when executed asynchronously. Since there are so few classes of relevant problems for which belief propagation is known to converge, even with *synchronous execution*, this is surprising.
3) We characterize the convergence time in regular graphs of the synchronous version of consensus propagation in terms of the mixing time of a certain Markov chain over edges of the graph.
4) We explain why the convergence time of consensus propagation scales more gracefully with the number of nodes than does that of pairwise averaging, and for certain classes of graphs, we quantify the improvement.

It is worth mentioning a recent and related line of research on the use of belief propagation as an asynchronous distributed protocol to arrive at consensus among nodes in a network, when each node makes a conditionally independent observation of the class of an object and would like to know the most probable class based on all observations [31]. The authors establish that belief propagation converges and provides each node with the most probable class when the network is a tree or a regular graph. They further show that for a certain class of random graphs, the result holds in an asymptotic sense as the number of nodes grows. To deal with general connected graphs, the authors offer a more complex protocol with convergence guarantees. It is interesting to note that this classification problem can be reduced to one of averaging. In particular, if each node starts out with the conditional probability of each class given its own observation and the network carries out a protocol to compute the average log probability for each class, each node obtains the conditional probabilities given *all* observations. Hence, consensus propagation also solves this classification problem.

II. Algorithm

Consider a connected undirected graph (V, E) with $V = \{1, \dots, n\}$. For each node $i \in V$, let $N(i) = \{j \mid (i,j) \in E\}$ be the set of neighbors of i. Let $\vec{E} \subseteq V \times V$ be a set consisting of two directed edges $\{i,j\}$ and $\{j,i\}$ per undirected edge $(i,j) \in E$. (In general, we will use braces for directed edges and parentheses for undirected edges.)

Each node $i \in V$ is assigned a number $y_i \in \mathbb{R}$. The goal is for each node to obtain an estimate of $\bar{y} = \sum_{i \in V} y_i / n$ through an asynchronous distributed protocol in which each node carries out simple computations and communicates parsimonious messages to its neighbors.

We propose consensus propagation as an approach to the aforementioned problem. In this protocol, if a node i communicates to a neighbor j at time t, it transmits a message consisting of two numerical values. Let $\mu_{ij}^{(t)} \in \mathbb{R}$ and $K_{ij}^{(t)} \in \mathbb{R}_+$ denote the values associated with the most recently transmitted message from i to j at or before time t. At each time t, node i has stored in memory the most recent message from each neighbor: $\{\mu_{ui}^{(t)}, K_{ui}^{(t)} \mid u \in N(i)\}$. If, at time $t+1$, node i chooses to communicate with a neighboring node $j \in N(i)$, it constructs a new message that is a function of the set of most recent messages $\{\mu_{ui}^{(t)}, K_{ui}^{(t)} \mid u \in N(i) \setminus j\}$ received from neighbors other than j. The initial values in memory before receiving any messages are arbitrary.

In order to illustrate how the parameter vectors $\mu^{(t)}$ and $K^{(t)}$ evolve, we will first describe a special case of the consensus propagation algorithm that is particularly intuitive. Then, we will describe the general algorithm and its relationship to belief propagation.

A. Intuitive Interpretation

Consider the special case of a singly connected graph. That is, a connected graph where there are no loops present (a tree). Assume, for the moment, that at every point in time, every pair of connected nodes communicates. As illustrated in Fig. 1, for any edge $\{i,j\} \in \vec{E}$, there is a set $S_{ij} \subset V$ of nodes, with $i \in S_{ij}$, that can transmit information to $S_{ji} = V \setminus S_{ij}$, with $j \in S_{ji}$, only through $\{i,j\}$. In order for nodes in S_{ji} to compute $\overline{y}$, they must at least be provided with the average μ_{ij}^* among observations at nodes in S_{ij} and the cardinality $K_{ij}^* = |S_{ij}|$. Similarly, in order for nodes in S_{ij} to compute $\overline{y}$, they must at least be provided with the average μ_{ji}^* among observations at nodes in S_{ji} and the cardinality $K_{ji}^* = |S_{ji}|$. These values must be communicated through the link $\{j,i\}$.

The messages $\mu_{ij}^{(t)}$ and $K_{ij}^{(t)}$, transmitted from node i to node j, can be viewed as iterative estimates of the quantities μ_{ij}^* and K_{ij}^*. They evolve according to

$$\mu_{ij}^{(t)} = \frac{y_i + \sum_{u \in N(i) \setminus j} K_{ui}^{(t-1)} \mu_{ui}^{(t-1)}}{1 + \sum_{u \in N(i) \setminus j} K_{ui}^{(t-1)}}, \quad \forall \{i,j\} \in \vec{E} \tag{1a}$$

$$K_{ij}^{(t)} = 1 + \sum_{u \in N(i) \setminus j} K_{ui}^{(t-1)}, \quad \forall \{i,j\} \in \vec{E}. \tag{1b}$$

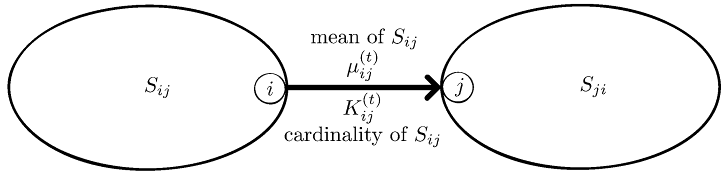

Fig. 1. Interpretation of messages in a singly connected graph with $\beta = \infty$.

At each time t, each node i computes an estimate of the global average $\overline{y}$ according to

$$x_i^{(t)} = \frac{y_i + \sum_{u \in N(i)} K_{ui}^{(t)} \mu_{ui}^{(t)}}{1 + \sum_{u \in N(i)} K_{ui}^{(t)}}.$$

Assume that the algorithm is initialized with $K^{(0)} = 0$. A simple inductive argument shows that at each time $t \geq 1$, $\mu_{ij}^{(t)}$ is the average among observations at the nodes in the set S_{ij} that are at a distance less than or equal to t from node i. Furthermore, $K_{ij}^{(t)}$ is the cardinality of this collection of nodes. Since any node in S_{ij} is at a distance from node i that it at most the diameter of the graph, if t is greater than the diameter of the graph, we have $K^{(t)} = K^*$ and $\mu^{(t)} = \mu^*$. Thus, for any $i \in V$, and t sufficiently large

$$x_i^{(t)} = \frac{y_i + \sum_{u \in N(i)} K_{ui}^* \mu_{ui}^*}{1 + \sum_{u \in N(i)} K_{ui}^*} = \overline{y}.$$

So, $x_i^{(t)}$ converges to the global average $\overline{y}$. Further, this simple algorithm converges in as short a time as is possible, since the diameter of the graph is the minimum amount of time for the two most distance nodes to communicate.

Now, suppose that the graph has cycles. For any directed edge $\{i,j\} \in \vec{E}$ that is part of a cycle, $K_{ij}^{(t)} \to \infty$. Hence, the algorithm does not converge. A heuristic fix might be to compose the iteration (1b) with one that attenuates

$$\tilde{K}_{ij}^{(t)} \leftarrow 1 + \sum_{u \in N(i) \setminus j} K_{ui}^{(t-1)}$$

$$K_{ij}^{(t)} \leftarrow \frac{\tilde{K}_{ij}^{(t)}}{1 + \tilde{K}_{ij}^{(t)}/(\beta Q_{ij})}.$$

Here, $Q_{ij} > 0$ and $\beta > 0$ are positive constants. We can view the unattenuated algorithm as setting $\beta = \infty$. In the attenuated algorithm, the message is essentially unaffected when $\tilde{K}_{ij}^{(t)}/(\beta Q_{ij})$ is small but becomes increasingly attenuated as $\tilde{K}_{ij}^{(t)}$ grows. This is exactly the kind of attenuation carried out by consensus propagation. Understanding why this kind of attenuation leads to desirable results is a subject of our analysis.

B. General Algorithm

Consensus propagation is parameterized by a scalar $\beta > 0$ and a nonnegative matrix $Q \in \mathbb{R}_+^{n \times n}$ with $Q_{ij} > 0$ if and only if $i \neq j$ and $(i,j) \in E$. For each $\{i,j\} \in \vec{E}$, it is useful to define the following three functions:

$$\mathcal{F}_{ij}(K) = \frac{1 + \sum_{u \in N(i) \setminus j} K_{ui}}{1 + \frac{1}{\beta Q_{ij}} \left(1 + \sum_{u \in N(i) \setminus j} K_{ui}\right)} \tag{2a}$$

$$\mathcal{G}_{ij}(\mu, K) = \frac{y_i + \sum_{u \in N(i) \setminus j} K_{ui} \mu_{ui}}{1 + \sum_{u \in N(i) \setminus j} K_{ui}} \tag{2b}$$

$$\mathcal{X}_i(\mu, K) = \frac{y_i + \sum_{u \in N(i)} K_{ui} \mu_{ui}}{1 + \sum_{u \in N(i)} K_{ui}}. \tag{2c}$$

For each t, denote by $U_t \subseteq \vec{E}$ the set of directed edges along which messages are transmitted at time t. Consensus propagation is presented below as Algorithm 1.

Algorithm 1 Consensus propagation.

```
for time t = 1 to ∞ do
    for all {i,j} ∈ U_t do
        K_ij^(t) ← F_ij(K^(t−1))
        μ_ij^(t) ← G_ij(μ^(t−1), K^(t−1))
    end for
    for all {i,j} ∉ U_t do
        K_ij^(t) ← K_ij^(t−1)
        μ_ij^(t) ← μ_ij^(t−1)
    end for
    x^(t) ← X(μ^(t), K^(t))
end for
```

Consensus propagation is a *distributed protocol* because computations at each node require only information that is locally available. In particular, the messages $K_{ij}^{(t)} = \mathcal{F}_{ij}(K^{(t-1)})$ and $\mu_{ij}^{(t)} = \mathcal{G}_{ij}(\mu^{(t-1)}, K^{(t-1)})$ transmitted from node i to node j depend only on $\{\mu_{ui}^{(t-1)}, K_{ui}^{(t-1)} \mid u \in N(i)\}$, which node i has stored in memory. Similarly, $x_i^{(t)}$, which serves as an estimate of $\overline{y}$, depends only on $\{\mu_{ui}^{(t)}, K_{ui}^{(t)} \mid u \in N(i)\}$.

Consensus propagation is an *asynchronous protocol* because only a subset of the potential messages are transmitted at each time. Our convergence analysis can also be extended to accommodate more general models of asynchronism that involve communication delays, as those presented in [32].

In our study of convergence *time*, we will focus on the *synchronous* version of consensus propagation. This is where $U_t = \vec{E}$ for all t. Note that synchronous consensus propagation is defined by

$$K^{(t)} = \mathcal{F}(K^{(t-1)}) \tag{3a}$$

$$\mu^{(t)} = \mathcal{G}(\mu^{(t-1)}, K^{(t-1)}) \tag{3b}$$

$$x^{(t)} = \mathcal{X}(\mu^{(t-1)}, K^{(t-1)}). \tag{3c}$$

C. Relation to Belief Propagation

Consensus propagation can also be viewed as a special case of belief propagation. In this context, belief propagation is used to approximate the marginal distributions of a vector $x \in \mathbb{R}^n$ conditioned on the observations $y \in \mathbb{R}^n$. The mode of each of the marginal distributions approximates $\overline{y}$.

Take the prior distribution over (x, y) to be the normalized product of potential functions $\{\psi_i(\cdot) \mid i \in V\}$ and compatibility functions $\{\psi_{ij}^\beta(\cdot) \mid (i,j) \in E\}$, given by

$$\psi_i(x_i) = \exp(-(x_i - y_i)^2)$$
$$\psi_{ij}^\beta(x_i, x_j) = \exp(-\beta Q_{ij}(x_i - x_j)^2)$$

where $Q_{ij} > 0$, for each edge $(i,j) \in E$, and $\beta > 0$ are constants. Note that β can be viewed as an inverse temperature parameter; as β increases, components of x associated with adjacent nodes become increasingly correlated.

Let Γ be a positive semidefinite symmetric matrix such that

$$x^\top \Gamma x = \sum_{(i,j)\in E} Q_{ij}(x_i - x_j)^2.$$

Note that when $Q_{ij} = 1$, for all edges $(i,j) \in E$, Γ is the graph Laplacian. Given the vector y of observations, the conditional density of x is

$$p^\beta(x) \propto \prod_{i\in V} \psi_i(x_i) \prod_{(i,j)\in E} \psi_{ij}^\beta(x_i, x_j) = \exp\left(-\|x - y\|_2^2 - \beta x^\top \Gamma x\right).$$

Let x^β denote the mode (maximizer) of $p^\beta(\cdot)$. Since the distribution is Gaussian, each component x_i^β is also the mode of the corresponding marginal distribution. Note that x^β it is the unique solution to the positive definite quadratic program

$$\underset{x}{\text{minimize}} \quad \|x - y\|_2^2 + \beta x^\top \Gamma x. \tag{4}$$

The following theorem relates x^β to the mean value $\overline{y}$.

Theorem 1: $\sum_i X_i^\beta / n = \overline{y}$ and $\lim_{\beta\uparrow\infty} x_i^\beta = \overline{y}$, for all $i \in V$.

Proof: The first-order conditions for optimality imply $(I + \beta\Gamma)x^\beta = y$. If we set $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$, we have $\Gamma\mathbf{1} = 0$, hence $\mathbf{1}^\top x^\beta / n = \mathbf{1}^\top y/n = \overline{y}$. Let U be an orthogonal matrix and D a diagonal matrix that form a spectral decomposition of Γ, that is $\Gamma = U^\top D U$. Then, we have $x^\beta = U^\top (I + \beta D)^{-1} U y$. It is clear that Γ has eigenvalue 0 with multiplicity 1 and corresponding normalized eigenvector $\mathbf{1}/\sqrt{n}$, and all other eigenvalues $d_2, \ldots, d_n$ of Γ are positive. Then, if $D = \mathrm{diag}(0, d_2, \ldots, d_n)$

$$\lim_{\beta\to\infty} x^\beta = \lim_{\beta\to\infty} U^\top \mathrm{diag}(1, 1/(1+\beta d_2), \ldots, 1/(1+\beta d_n))Uy = \mathbf{1}\mathbf{1}^\top y/n.$$

□

The above theorem suggests that if β is sufficiently large, then each component x_i^β can be used as an estimate of $\overline{y}$.

In belief propagation, messages are passed along edges of a Markov random field. In our case, because of the structure of the distribution $p^\beta(\cdot)$, the relevant Markov random field has the same topology as the graph (V, E). The message $M_{ij}^{(t)}(\cdot)$ passed from node i to node j at time t is a distribution on the variable x_j. Node i computes this message using incoming messages from other nodes as defined by the update equation

$$M_{ij}^{(t)}(x_j) = \kappa \int \psi_{ij}(x_i', x_j)\psi_i(x_i') \prod_{u\in N(i)\setminus j} M_{ui}^{(t-1)}(x_i')\, dx_i'. \tag{5}$$

Here, κ is a normalizing constant. Since our underlying distribution $p^\beta(\cdot)$ is Gaussian, it is natural to consider messages which are Gaussian distributions. In particular, let $(\mu_{ij}^{(t)}, K_{ij}^{(t)}) \in \mathbb{R} \times \mathbb{R}_+$ parameterize Gaussian message $M_{ij}^{(t)}(\cdot)$ according to

$$M_{ij}^{(t)}(x_j) \propto \exp\left(-K_{ij}^{(t)}(x_j - \mu_{ij}^{(t)})^2\right).$$

Then, (5) is equivalent to the synchronous consensus propagation iterations for $K^{(t)}$ and $\mu^{(t)}$.

The sequence of densities

$$p_j^{(t)}(x_j) \propto \psi_j(x_j) \prod_{i\in N(j)} M_{ij}^{(t)}(x_j) = \exp\left(-(x_j - y_j)^2 - \sum_{i\in N(j)} K_{ij}^{(t)}(x_j - \mu_{ij}^{(t)})^2\right)$$

is meant to converge to an approximation of the marginal conditional distribution of x_j. As such, an approximation to x_j^β is given by maximizing $p_j^{(t)}(\cdot)$. It is easy to show that, the maximum is attained by $x_j^{(t)} = \mathcal{X}_j(\mu^{(t)}, K^{(t)})$. With this and the aforementioned correspondences, we have shown that consensus propagation is a special case of belief propagation, and more specifically, Gaussian belief propagation.

Readers familiar with belief propagation will notice that in the derivation above we have used the sum–product form of the algorithm. In this case, since the underlying distribution is Gaussian, the max–product form yields equivalent iterations.

D. Relation to Prior Results

In light of the fact that consensus propagation is a special case of Gaussian belief propagation, it is natural to ask what prior results on belief propagation—Gaussian or more broadly—have to say in this context. Results from [28], [18], [33] establish that, in the absence of degeneracy, Gaussian belief propagation has a unique fixed point and that the mode of this fixed point is unbiased. The issue of convergence, however, is largely poorly understood. As observed numerically in [18], Gaussian belief propagation can diverge, even in the absence of degeneracy. Abstract sufficient conditions for convergence that have been developed in [28], [18] are difficult to verify in the consensus propagation case.

III. Convergence

As we have discussed, Gaussian belief propagation can diverge, even when the graph has a single cycle. One might expect the same from consensus propagation. However, the following theorem establishes convergence.

Theorem 2: The following hold.

(i) There exist unique vectors (μ^β, K^β) such that

$$K^\beta = \mathcal{F}(K^\beta) \quad \text{and} \quad \mu^\beta = \mathcal{G}(\mu^\beta, K^\beta).$$

(ii) Suppose that each directed edge $\{i,j\}$ appears infinitely often in the sequence of communication sets $\{U_t\}$. Then, independent of the initial condition $(\mu^{(0)}, K^{(0)})$

$$\lim_{t\to\infty} K^{(t)} = K^\beta \quad \text{and} \quad \lim_{t\to\infty} \mu^{(t)} = \mu^\beta.$$

(iii) Given (μ^β, K^β), if $x^\beta = \mathcal{X}(\mu^\beta, K^\beta)$, then x^β is the mode of the distribution $p^\beta(\cdot)$.

Note that the condition on the communication sets in Theorem 2 (ii) corresponds to *total asynchronism* in the language of [32]. This is a weak assumption which ensures only that every component of $\mu^{(t)}$ and $K^{(t)}$ is updated infinitely often.

The proof of this theorem is deferred to the Appendix I, but it rests on two ideas. First, notice that, according to the update (2a), $K^{(t)}$ evolves independently of $\mu^{(t)}$. Hence, we analyze $K^{(t)}$ first. Following the work in [18], we prove that the functions $\{\mathcal{F}_{ij}(\cdot)\}$ are monotonic. This property is used to establish convergence to a unique fixed point. Next, we analyze $\mu^{(t)}$ assuming that $K^{(t)}$ has already converged. Given fixed K, the update equations for $\mu^{(t)}$ are linear, and we establish that they induce a contraction with respect to the maximum norm. This allows us to establish existence of a fixed point and both synchronous and asynchronous convergence.

IV. Convergence Time for Regular Graphs

In this section, we will study the convergence time of synchronous consensus propagation. For $\epsilon > 0$, we will say that an estimate $\tilde{x}$ of $\bar{y}$ is ϵ-accurate if

$$\|\tilde{x} - \bar{y}\mathbf{1}\|_{2,n} \le \epsilon. \tag{6}$$

Here, for integer m, we set $\|\cdot\|_{2,m}$ to be the norm on $\mathbb{R}^m$ defined by $\|x\|_{2,m} = \|x\|_2/\sqrt{m}$. We are interested in the number of iterations required to obtain an ϵ-accurate estimate of the mean $\bar{y}$.

Note that we are primarily interested in how the performance of consensus propagation behaves over a series of problem instances as we scale the size of the graph. Since our measure of error (6) is absolute, we require that the set of values $\{y_i\}$ lie in some bounded set. Without loss of generality, we will take $y_i \in [0,1]$, for all $i \in V$.

A. The Case of Regular Graphs

We will restrict our analysis of convergence time to cases where (V,E) is a d-regular graph, for $d \ge 2$. Extension of our analysis to broader classes of graphs remains an open issue. We will also make simplifying assumptions that $Q_{ij} = 1$, $\mu_{ij}^{(0)} = y_i$, and $K^{(0)} = [k_0]_{ij}$ for some scalar $k_0 \ge 0$.

In this restricted setting, the subspace of constant K vectors is invariant under $\mathcal{F}$. This implies that there is some scalar $k^\beta > 0$ so that $K^\beta = [k^\beta]_{ij}$. This k^β is the unique solution to the fixed-point equation

$$k^\beta = \frac{1 + (d-1)k^\beta}{1 + (1 + (d-1)k^\beta)/\beta}. \tag{7}$$

Given a uniform initial condition $K^{(0)} = [k_0]_{ij}$, we can study the sequence of iterates $\{K^{(t)}\}$ by examining the scalar sequence $\{k_t\}$, defined by

$$k_t = \frac{1 + (d-1)k_{t-1}}{1 + (1 + (d-1)k_{t-1})/\beta}. \tag{8}$$

In particular, we have $K^{(t)} = [k_t]_{ij}$, for all $t \ge 0$.

Similarly, in this setting, the equations for the evolution of $\mu^{(t)}$ take the special form

$$\mu_{ij}^{(t)} = \frac{y_i}{1 + (d-1)k_{t-1}} + \left(1 - \frac{1}{1 + (d-1)k_{t-1}}\right) \sum_{u \in N(i)\setminus j} \frac{\mu_{ui}^{(t-1)}}{d-1}.$$

Defining $\gamma_t = 1/(1 + (d-1)k_t)$, we have, in vector form

$$\mu^{(t)} = \gamma_{t-1}\hat{y} + (1 - \gamma_{t-1})\hat{P}\mu^{(t-1)} \tag{9}$$

where $\hat{y} \in \mathbb{R}^{nd}$ is a vector with $\hat{y}_{ij} = y_i$ and $\hat{P} \in \mathbb{R}_+^{nd \times nd}$ is a doubly stochastic matrix. The matrix $\hat{P}$ corresponds to a Markov chain on the set of directed edges $\vec{E}$. In this chain, a directed edge $\{i,j\}$ transitions to a directed edge $\{u,i\}$ with $u \in N(i) \setminus j$, with equal probability assigned to each such edge. As in (3), we associate each $\mu^{(t)}$ with an estimate $x^{(t)}$ of x^β according to

$$x^{(t)} = \frac{1}{1 + dk^\beta} y + \frac{dk^\beta}{1 + dk^\beta} A\mu^{(t)}$$

where $A \in \mathbb{R}_+^{n \times nd}$ is a matrix defined by

$$(A\mu)_j = \sum_{i \in N(j)} \mu_{ij}/d.$$

B. The Cesàro Mixing Time

The update (9) suggests that the convergence of $\mu^{(t)}$ is intimately tied to a notion of mixing time associated with $\hat{P}$. Let $\hat{P}^\star$ be the Cesàro limit

$$\hat{P}^\star = \lim_{t\to\infty} \sum_{\tau=0}^{t-1} \hat{P}^\tau / t.$$

Define the Cesàro mixing time $\tau^\star$ by

$$\tau^\star = \sup_{t \ge 0} \left\| \sum_{\tau=0}^{t} (\hat{P}^\tau - \hat{P}^\star) \right\|_{2,nd}.$$

Here, $\|\cdot\|_{2,nd}$ is the matrix norm induced by the corresponding vector norm $\|\cdot\|_{2,nd}$. Since $\hat{P}$ is a stochastic matrix, $\hat{P}^\star$ is well defined and $\tau^\star < \infty$. Note that, in the case where $\hat{P}$ is aperiodic, irreducible, and symmetric, $\tau^\star$ corresponds to the traditional definition of mixing time: the inverse of the spectral gap of $\hat{P}$.

C. Bounds on the Convergence Time

Let $\gamma^\beta = \lim_{t\uparrow\infty} \gamma_t = 1/(1+(d-1)k^\beta)$. With an initial condition $k_0 = k^\beta$, the update equation for $\mu^{(t)}$ becomes

$$\mu^{(t)} = \gamma^\beta \hat{y} + (1-\gamma^\beta)\hat{P}\mu^{(t-1)}.$$

Since $\gamma^\beta \in (0,1)$, this iteration is a contraction mapping, with contraction factor $1-\gamma^\beta$. It is easy to show that γ^β is monotonically decreasing in β, and as such, large values of β are likely to result in slower convergence. On the other hand, Theorem 1 suggests that large values of β are required to obtain accurate estimates of $\bar{y}$. To balance these conflicting issues, β must be appropriately chosen.

A time t^* is said to be an ϵ*-convergence time* if estimates $x^{(t)}$ are ϵ-accurate for all $t \geq t^*$. The following theorem, whose proof is deferred until the Appendix II, establishes a bound on the ϵ-convergence time of synchronous consensus propagation given appropriately chosen β, as a function of ϵ and $\tau^\star$.

Theorem 3: Suppose $k_0 \leq k^\beta$. If $d=2$ there exists a $\beta = \Theta((\tau^\star/\epsilon)^2)$ and if $d>2$ there exists a $\beta = \Theta(\tau^\star/\epsilon)$ such that some $t^* = O((\tau^\star/\epsilon)\log(\tau^\star/\epsilon))$ is an ϵ-convergence time.

In the preceding theorem, k_0 is initialized arbitrarily so long as $k_0 \leq k^\beta$. Typically, one might set $k_0 = 0$ to guarantee this. Another case of particular interest is when $k_0 = k^\beta$, so that $k_t = k^\beta$ for all $t \geq 0$. In this case, the following theorem, whose proof is deferred until the Appendix II, offers a better convergence time bound than Theorem 3.

Theorem 4: Suppose $k_0 = k^\beta$. If $d=2$ there exists a $\beta = \Theta((\tau^\star/\epsilon)^2)$ and if $d>2$ there exists a $\beta = \Theta(\tau^\star/\epsilon)$ such that some $t^* = O((\tau^\star/\epsilon)\log(1/\epsilon))$ is an ϵ-convergence time.

Theorems 3 and 4 suggest that initializing with $k_0 = k^\beta$ leads to an improvement in convergence time. However, in our computational experience, we have found that an initial condition of $k_0 = 0$ consistently results in *faster* convergence than $k_0 = k^\beta$. Hence, we suspect that a convergence time bound of $O((\tau^\star/\epsilon)\log(1/\epsilon))$ also holds for the case of $k_0 = 0$. Proving this remains an open issue.

D. Adaptive Mixing Time Search

The choice of β is critical in that it determines both convergence time and ultimate accuracy. This raises the question of how to choose β for a particular graph. The choices posited in Theorems 3 and 4 require knowledge of $\tau^\star$, which may be both difficult to compute and also requires knowledge of the graph topology. This counteracts our purpose of developing a distributed protocol.

In order to address this concern, consider Algorithm 2, which is designed for the case of $d>2$. It uses a doubling sequence of guesses $\tilde{\tau}$ for the Cesáro mixing time $\tau^\star$. Each guess leads to a choice of β and a number of iterations t^*. Note that the algorithm takes $\epsilon > 0$ as input.

Algorithm 2 Synchronous consensus propagation with adaptive mixing time search.

$K^{(0)} \leftarrow 0$, $\mu^{(0)} \leftarrow \hat{y}$, $t \leftarrow 0$
for $\ell = 0$ to ∞ **do**
  $\tilde{\tau} \leftarrow 2^\ell$
  Set β and t^* as indicated by Theorem 3, assuming $\tau^\star = \tilde{\tau}$
  for $s = 1$ to t^* **do**
    $\mu^{(t)} \leftarrow \mathcal{G}(\mu^{(t-1)}, K^{(t-1)})$, $K^{(t)} \leftarrow \mathcal{F}(K^{(t)})$
    $t \leftarrow t+1$
  end for
end for

Consider applying this procedure to a d-regular graph with fixed $d>2$ but topology otherwise unspecified. It follows from Theorem 3 that this procedure has an ϵ-convergence time of $O((\tau^\star/\epsilon)\log(\tau^\star/\epsilon))$. An entirely analogous algorithm can be designed for the case of $d=2$.

We expect that many variations of this procedure can be made effective. Asynchronous versions would involve each node adapting a local estimate of the mixing time.

V. Comparison With Pairwise Averaging

Using the results of Section IV, we can compare the performance of consensus propagation to that of pairwise averaging. Pairwise averaging is usually defined in an asynchronous setting, but there is a synchronous counterpart which works as follows. Consider a doubly stochastic symmetric matrix $P \in \mathbb{R}^{n\times n}$ such that $P_{ij} = 0$ if $i \neq j$ and $(i,j) \notin E$. Evolve estimates according to $x^{(t)} = Px^{(t-1)}$, initialized with $x^{(0)} = y$. Here, at each time t, a node i is computing a new estimate $y_i^{(t)}$ which is an average of the estimates at node i and its neighbors during the previous time step. If the matrix P is aperiodic and irreducible, then $x^{(t)} = P^t y \to \bar{y}\mathbf{1}$ as $t \uparrow \infty$.

In the case of a singly connected graph, synchronous consensus propagation converges exactly in a number of iterations equal to the diameter of the graph. Moreover, when $\beta = \infty$, this convergence is to the exact mean, as discussed in Section II-A. This is the best one can hope for under any algorithm, since the diameter is the minimum amount of time required for a message to travel between the two most distant nodes. On the other hand, for a fixed accuracy ϵ, the worst case number of iterations required by synchronous pairwise averaging on a singly connected graph scales at least quadratically in the diameter [34].

The rate of convergence of synchronous pairwise averaging is governed by the relation $\|x^{(t)} - \bar{y}\mathbf{1}\|_{2,n} \leq \lambda_2^t$, where λ_2 is the second largest eigenvalue[1] of P. Let $\tau_2 = 1/\log(1/\lambda_2)$, and call it the *mixing time* of P. In order to guarantee ϵ-accuracy (independent of y), $t > \tau_2 \log(1/\epsilon)$ suffices and $t = \Omega(\tau_2 \log(1/\epsilon))$ is required.

Consider d-regular graphs and fix a desired error tolerance ϵ. The number of iterations required by consensus propagation is $\Theta(\tau^\star \log \tau^\star)$, whereas that required by pairwise averaging is $\Theta(\tau_2)$. Both mixing times depend on the size and topology of the graph. τ_2 is the mixing time of a process on nodes that transitions along edges whereas $\tau^\star$ is the mixing time of a process on directed edges that transitions towards nodes. An important distinction is that the former process is allowed to "backtrack" where as the latter is not. By this we mean that a sequence of states (i, j, i) can be observed in the vertex process, but the sequence $(\{i, j\}, \{j, i\})$ cannot be observed in the edge process. As we will now illustrate through an example, it is this difference that makes τ_2 larger than $\tau^\star$ and, therefore, pairwise averaging less efficient than consensus propagation.

In the case of a cycle ($d = 2$) with an even number of nodes n, minimizing the mixing time over P results in $\tau_2 = \Theta(n^2)$ [35], [17], [36]. For comparison, as demonstrated in the following theorem (whose proof is deferred until the Appendix III), $\tau^\star$ is linear in n.

Theorem 5: For the cycle with n nodes, $\tau^\star \leq n/\sqrt{2}$.

Intuitively, the improvement in mixing time arises from the fact that the edge process moves around the cycle in a single direction and therefore travels distance t in order t iterations. The vertex process, on the other hand, is "diffusive" in nature. It randomly transitions back and forth among adjacent nodes, and requires order t^2 iterations to travel distance t. Nondiffusive methods have previously been suggested in the design of efficient algorithms for Markov chain sampling (see [37] and references therein).

The cycle example demonstrates a $\Theta(n/\log n)$ advantage offered by consensus propagation. Comparisons of mixing times associated with other graph topologies remains an issue for future analysis. Let us close by speculating on a uniform grid of n nodes over the m-dimensional unit torus. Here, $n^{1/m}$ is an integer, and each vertex has $2m$ neighbors, each a distance $n^{-1/m}$ away. With P optimized, it can be shown that $\tau_2 = \Theta(n^{2/m})$ [38]. We put forth a conjecture on $\tau^\star$.

Conjecture 1: For the m-dimensional torus with n nodes, $\tau^\star = \Theta(n^{(2m-1)/m^2})$.

Appendix I
Proof of Theorem 2

Given initial vectors $(\mu^{(0)}, K^{(0)})$, and a sequence of communication sets $\{U_1, U_2, \ldots\}$, the consensus propagation algorithm evolves parameter values over time according to

$$K_{ij}^{(t)} = \begin{cases} \mathcal{F}_{ij}(K^{(t-1)}), & \text{if } \{i,j\} \in U_t \\ K_{ij}^{(t-1)}, & \text{otherwise} \end{cases} \tag{10}$$

$$\mu_{ij}^{(t)} = \begin{cases} \mathcal{G}_{ij}(\mu^{(t-1)}, K^{(t-1)}), & \text{if } \{i,j\} \in U_t \\ \mu_{ij}^{(t-1)}, & \text{otherwise} \end{cases} \tag{11}$$

for times $t > 0$.

In order to establish Theorem 2, we will first study convergence of the inverse variance parameters $K^{(t)}$, and subsequently the mean parameters $\mu^{(t)}$.

A. Convergence of Inverse Variance Updates

Our analysis of the convergence of the inverse variance parameters follows the work in [18]. We begin with a fundamental lemma.

Lemma 1: For each $\{i, j\} \in \vec{E}$, the following facts hold.

(i) The function $\mathcal{F}_{ij}(\cdot)$ is continuous.

(ii) The function $\mathcal{F}_{ij}(\cdot)$ is monotonic. That is, if $K \leq K'$, where the inequality is interpreted component-wise, then $\mathcal{F}_{ij}(K) \leq \mathcal{F}_{ij}(K')$.

(iii) If $K'_{ij} = \mathcal{F}_{ij}(K)$, then $0 < K'_{ij} < \beta Q_{ij}$.

(iv) If $\alpha > 1$, then $\alpha \mathcal{F}_{ij}(K) > \mathcal{F}_{ij}(\alpha K)$.

Proof: Define the function $f : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$f(x) = \frac{1}{\gamma + \frac{1}{1+x}}$$

where $\gamma > 0$. Fact (i) follows from the fact that f is continuous. Fact (ii) follows from the fact that $f(x)$ is strictly increasing. Fact (iii) follows from the fact that $f(x) \in (0, 1/\gamma)$ for all $x \geq 0$. Fact (iv) follows from the fact that $\alpha f(x) \geq f(\alpha x)$. □

Now we consider the sequence of iterates $\{K^{(0)}, K^{(1)}, \ldots\}$ which evolve according to (10).

Lemma 2: Let $K^{(0)}$ be such that $\mathcal{F}_{ij}(K^{(0)}) \geq K^{(0)}$ for all $\{i, j\} \in \vec{E}$ (for example, $K^{(0)} = 0$). Then $K^{(t)}$ converges to a vector K^β such that $K^\beta = \mathcal{F}(K^\beta)$.

Proof: Convergence follows from the fact that the iterates are component-wise bounded and monotonic. The limit point must be a fixed point by continuity. □

Given the existence of a single fixed point, we can establish that the fixed point must be unique.

Lemma 3: The $\mathcal{F}$ operator has a unique fixed point K^β.

Proof: Denote K^β to be the fixed point obtained by iterating with initial condition $K^{(0)} = 0$, and let K' be some other fixed point. It is clear that $K^{(0)} < K'$, thus, by monotonicity, we must have $K^\beta \leq K'$. Define

$$\gamma = \inf\left\{\alpha \in [1, \infty) : K' \leq \alpha K^\beta\right\}.$$

It is clear that γ is well defined since $0 < \{K_{ij}^\beta, K'_{ij}\} < \beta Q_{ij}$. Also, we must have $\gamma > 1$, since $K^\beta \neq K'$. Then

$$K'_{ij} = \mathcal{F}_{ij}(K') \leq \mathcal{F}_{ij}(\gamma K^\beta) < \gamma \mathcal{F}_{ij}(K^\beta) = \gamma K_{ij}^\beta.$$

This contradicts the definition of γ. Hence, there is a unique fixed point. □

Lemma 4: Given an arbitrary initial condition $K^{(0)} \in \mathbb{R}_+^{|\vec{E}|}$

$$\lim_{t \to \infty} K^{(t)} = K^\beta.$$

[1]Here, we take the standard approach of ignoring the smallest eigenvalue of P. We will assume that this eigenvalue is smaller than λ_2 in magnitude. Note that a constant probability can be added to each self-loop of any particular matrix P so that this is true.

Proof: If $0 \leq K^{(0)} \leq K^{\beta}$, the result holds by monotonicity. Assume that $K^{\beta} \leq K^{(0)}$

$$\gamma = \inf\left\{\alpha \in [1,\infty) : K^{(0)} \leq \alpha K^{\beta}\right\}.$$

Then

$$K_{ij}^{\beta} \leq \mathcal{F}_{ij}(K^{(0)}) \leq \mathcal{F}_{ij}(\gamma K^{\beta}) \leq \gamma \mathcal{F}_{ij}(K^{\beta}) = \gamma K_{ij}^{\beta}.$$

Define a sequence $\{\tilde{K}^{(t)}\}$ by

$$\tilde{K}^{(0)} = \gamma K^{\beta}$$

and, for all $\{i,j\} \in \vec{E}, t > 0$

$$\tilde{K}_{ij}^{(t)} = \begin{cases} \mathcal{F}_{ij}(\tilde{K}^{(t-1)}), & \text{if } \{i,j\} \in U_t \\ K_{ij}^{(t-1)}, & \text{otherwise.} \end{cases}$$

Since $\mathcal{F}_{ij}(\tilde{K}^{(0)}) \leq \gamma \mathcal{F}_{ij}(K^{\beta}) = \tilde{K}_{ij}^{(0)}$, the sequence $\{\tilde{K}^{(t)}\}$ is monotonically decreasing and must have a limit which is a fixed point. Since the fixed point is unique, we have $\tilde{K}^{(t)} \to K^{\beta}$. But, $K^{\beta} \leq K^{(0)} \leq \tilde{K}^{(0)}$. By monotonicity, we also have $\tilde{K}^{(t)} \to K^{\beta}$.

Now, consider the case of general $K^{(0)}$. Define $\underline{K}$ and $\overline{K}$ such that $\underline{K} \leq K^{(0)} \leq \overline{K}$ and $\underline{K} \leq K^{\beta} \leq \overline{K}$. By the previous two cases and monotonicity, we again have$K^{(t)} \to K^{\beta}$. □

B. Convergence of Mean Updates

In this subsection, we will consider certain properties of the updates for the mean parameters. Define the operator $\mathcal{G}(\cdot, K)$ to be the synchronous update of all components of the mean vector according to

$$\mathcal{G}(\mu, K)_{ij} = \mathcal{G}_{ij}(\mu, K), \quad \forall \{i,j\} \in \vec{E}.$$

Lemma 5: There exists $\alpha \in (0,1)$ so that

(i) for all $\mu, \mu' \in \mathbb{R}^{\vec{E}}$

$$\|\mathcal{G}(\mu, K^{\beta}) - \mathcal{G}(\mu', K^{\beta})\|_{\infty} < \alpha \|\mu - \mu'\|_{\infty};$$

(ii) if t is sufficiently large, for all $\mu, \mu' \in \mathbb{R}^{\vec{E}}$

$$\|\mathcal{G}(\mu, K^{(t)}) - \mathcal{G}(\mu', K^{(t)})\|_{\infty} < \alpha \|\mu - \mu'\|_{\infty}.$$

Proof: Set

$$\bar{\alpha} = \max_{\substack{u \in N(i)\setminus j \\ \{i,j\} \in \vec{E}}} \frac{K_{ui}^{\beta}}{1 + \sum_{u' \in N(i)\setminus j} K_{u'i}^{\beta}}.$$

Observing that $\bar{\alpha} < 1$, Part (i) follows.

Define

$$\bar{\alpha}_t = \max_{\substack{u \in N(i)\setminus j \\ \{i,j\} \in \vec{E}}} \frac{K_{ui}^{(t)}}{1 + \sum_{u' \in N(i)\setminus j} K_{u'i}^{(t)}}.$$

Since $K^{(t)} \to K^{\beta}$, by continuity $\bar{\alpha}_t \to \bar{\alpha} < 1$. Then, Part (ii) follows. □

Lemma 5 states that $\mathcal{G}(\cdot, K^{\beta})$ is a maximum norm contraction. This leads to the following lemma.

Lemma 6: The following hold.

(i) There is unique fixed point μ^{β} such that

$$\mu^{\beta} = \mathcal{G}(\mu^{\beta}, K^{\beta}).$$

(ii) There exists T_1 such if $t \geq T_1$, the operator $\mathcal{G}(\cdot, K^{(t)})$ has a unique fixed point $\nu^{(t)}$. That is,

$$\nu^{(t)} = \mathcal{G}(\nu^{(t)}, K^{(t)}).$$

(iii) For any $\epsilon > 0$, there exists $T_2 \geq T_1$ so that if $t \geq T_2$

$$\|\nu^{(t)} - \mu^{\beta}\|_{\infty} < \epsilon.$$

Proof: For Part (i), since $\mathcal{G}(\cdot, K^{\beta})$ is a maximum norm contraction, existence of a unique fixed point μ^{β} follows from, for example, Proposition 3.1.1 in [32]. Part (ii) is established similarly.

For Part (iii), note for t sufficiently large, the linear system of equations

$$\nu = \mathcal{G}(\nu, K^{(t)})$$

over $\nu \in \mathbb{R}^{\vec{E}}$ is nonsingular, by Part (ii). Since $K^{(t)} \to K^{\beta}$, the coefficients of this system of equations continuously converge to those of

$$\nu = \mathcal{G}(\nu, K^{\beta}).$$

Then, we must have $\nu^{(t)} \to \mu^{\beta}$. □

C. Overall Convergence

We are now ready to prove Theorem 2.

Theorem 2: Assume that the communication sets $\{U_t\}$ have the property that every directed edge $\{i,j\} \in \vec{E}$ appears in U_t for infinitely many t. The following hold.

(i) There are unique vectors (μ^{β}, K^{β}) such that

$$K^{\beta} = \mathcal{F}(K^{\beta}) \quad \text{and} \quad \mu^{\beta} = \mathcal{G}(\mu^{\beta}, K^{\beta}).$$

(ii) Independent of the initial condition $(\mu^{(0)}, K^{(0)})$

$$\lim_{t\to\infty} K^{(t)} = K^{\beta} \quad \text{and} \quad \lim_{t\to\infty} \mu^{(t)} = \mu^{\beta}.$$

(iii) Given (μ^{β}, K^{β}), if $x^{\beta} = \mathcal{X}(\mu^{\beta}, K^{\beta})$, then x^{β} is the mode of the distribution $p^{\beta}(\cdot)$.

Proof: Existence and uniqueness of the fixed point K^{β} and convergence of the vector $K^{(t)}$ to K^{β} follow from Lemmas 3 and 4, respectively. Existence and uniqueness of the fixed point μ^{β} follows from Lemma 6.

To establish the balance of Part (ii), we need to show that $\mu^{(t)} \to \mu^{\beta}$. We will use a variant of the "box condition" argument of Proposition 6.2.1 in [32].

Fix any $\epsilon > 0$. By Lemma 6, pick T_2 so that if $t \geq T_2$, then $\nu^{(t)}$ exists with $\nu^{(t)} = \mathcal{G}(\nu^{(t)}, K^{(t)})$ and $\|\nu^{(t)} - \mu^\beta\|_\infty < \epsilon$. For $t > T_2$, if $\{i,j\} \in U_t$

$$\begin{aligned}|\mu_{ij}^{(t)} - \mu_{ij}^\beta| &\leq |\mu_{ij}^{(t)} - \nu_{ij}^{(t-1)}| + |\nu_{ij}^{(t-1)} - \mu_{ij}^\beta| \\ &= |\mathcal{G}_{ij}(\mu^{(t-1)}, K^{(t-1)}) - \mathcal{G}_{ij}(\nu^{(t-1)}, K^{(t-1)})| \\ &\quad + \|\nu^{(t-1)} - \mu^\beta\|_\infty \\ &< \alpha\|\mu^{(t-1)} - \nu^{(t-1)}\|_\infty + \|\nu^{(t-1)} - \mu^\beta\|_\infty \\ &\leq \alpha\|\mu^{(t-1)} - \mu^\beta\|_\infty + (1+\alpha)\|\nu^{(t-1)} - \mu^\beta\|_\infty \\ &\leq \alpha\|\mu^{(t-1)} - \mu^\beta\|_\infty + (1+\alpha)\epsilon. \end{aligned} \tag{12}$$

For $k \geq 0$, define $\mathcal{A}_k$ to be the set of vectors $\mu \in \mathbb{R}^{|\vec{E}|}$ such that

$$\|\mu - \mu^\beta\|_\infty \leq \alpha^k \|\mu^{(T_2)} - \mu^\beta\|_\infty + (1+\alpha)\epsilon/(1-\alpha).$$

We would like to show that for every $k \geq 0$, there is a time t_k such that $\mu^{(t)} \in \mathcal{A}_k$, for all $t \geq t_k$. We proceed by induction.

When $k = 0$, set $t_k = T_2$. Clearly, $\mu^{(T_2)} \in \mathcal{A}_0$. Assume that $\mu^{(t-1)} \in \mathcal{A}_0$, for some $t > T_2$. Then, if $\{i,j\} \in U_t$, from (12)

$$\begin{aligned}|\mu_{ij}^{(t)} - \mu_{ij}^\beta| &< \alpha\|\mu^{(t-1)} - \mu^\beta\|_\infty + (1+\alpha)\epsilon \\ &< \alpha\|\mu^{(T_2)} - \mu^\beta\|_\infty + \frac{1+\alpha}{1-\alpha}\alpha\epsilon + (1+\alpha)\epsilon \\ &< \|\mu^{(T_2)} - \mu^\beta\|_\infty + \frac{1+\alpha}{1-\alpha}\epsilon.\end{aligned}$$

If $\{i,j\} \notin U_t$

$$|\mu_{ij}^{(t)} - \mu_{ij}^\beta| = |\mu_{ij}^{(t-1)} - \mu_{ij}^\beta| < \|\mu^{(T_2)} - \mu^\beta\|_\infty + \frac{1+\alpha}{1-\alpha}\epsilon.$$

Thus, $\mu^{(t)} \in \mathcal{A}_0$. By induction, $\mu^{(t)} \in \mathcal{A}_0$ for all $t \geq T_2$.

Now, assume that t_{k-1} exists, for some $k-1 \geq 0$. Let $t > t_{k-1}$ be some time such that $\{i,j\} \in U_t$. Then, by (12) and the fact that $\mu^{(t-1)} \in \mathcal{A}_{k-1}$

$$\begin{aligned}|\mu_{ij}^{(t)} - \mu_{ij}^\beta| &< \alpha\|\mu^{(t-1)} - \mu^\beta\|_\infty + (1+\alpha)\epsilon \\ &< \alpha^k\|\mu^{(T_2)} - \mu^\beta\|_\infty + \frac{1+\alpha}{1-\alpha}\epsilon.\end{aligned}$$

For each $\{i,j\} \in \vec{E}$, let $\tau_{ij}^k > t_{k-1}$ be the earliest time after t_{k-1} that $\{i,j\} \in U_{\tau_{ij}^k}$. If we set t_k to be the largest of these times, we have $\mu^{(t)} \in \mathcal{A}_k$, for all $t \geq t_k$.

We have established that

$$\limsup_{t\to\infty} \|\mu^{(t)} - \mu^\beta\|_\infty \leq \alpha^k \|\mu^{(T_2)} - \mu^\beta\|_\infty + \frac{1+\alpha}{1-\alpha}\epsilon$$

for all $k \geq 0$. Taking a limit as $k \to \infty$, we have

$$\limsup_{t\to\infty} \|\mu^{(t)} - \mu^\beta\|_\infty \leq \frac{1+\alpha}{1-\alpha}\epsilon.$$

Since ϵ was arbitrary, we have the convergence $\mu^{(t)} \to \mu^\beta$.

Part (iii) follows from the fact that Gaussian belief propagation, when it converges, computes exact means [28], [18], [33]. □

APPENDIX II
PROOFS OF THEOREMS 3 AND 4

In this appendix, we will prove Theorems 3 and 4. We will start with some preliminary lemmas.

A. Preliminary Lemmas

The following lemma provides bounds on k^β and γ^β in terms of β.

Lemma 7: If $d = 2$

$$2\sqrt{\beta} - 1/2 < k^\beta < 2\sqrt{\beta}$$
$$\frac{1}{2\sqrt{\beta}+1} < \gamma^\beta < \frac{1}{2\sqrt{\beta}+1/2}.$$

If $d > 2$

$$\left(1 - \frac{1}{d-1}\right)\beta - \frac{1}{d-1} < k^\beta < \beta$$
$$\frac{1}{1+(d-1)\beta} < \gamma^\beta < \frac{1}{(d-2)\beta}.$$

Proof: Starting with the fixed point (7), some algebra leads to

$$\frac{d-1}{\beta}(k^\beta)^2 + (2 + \frac{1}{\beta} - d)k^\beta - 1 = 0.$$

The quadratic formula gives us

$$k^\beta = \frac{\beta}{2} - \frac{\beta+1}{2(d-1)} + \sqrt{\left(\frac{\beta}{2} - \frac{\beta+1}{2(d-1)}\right)^2 + 4\frac{\beta}{d-1}}$$

from which it is easy to derived the desired bounds. □

The following lemma offers useful expressions for the fixed point μ^β and the mode x^β.

Lemma 8:

$$\mu^\beta = \sum_{\tau=0}^\infty \gamma^\beta(1-\gamma^\beta)^\tau \hat{P}^\tau \hat{y} \tag{13a}$$

$$x^\beta = \frac{y}{1+dk^\beta} + \frac{dk^\beta}{1+dk^\beta}\sum_{\tau=0}^\infty \gamma^\beta(1-\gamma^\beta)^\tau A\hat{P}^\tau \hat{y}. \tag{13b}$$

Proof: If we consider the algorithm when $k_0 = k^\beta$, then $k_t = k^\beta$ and $\gamma_t = \gamma^\beta$ for all $t \geq 0$. Then, using (9) and induction, we have

$$\mu^{(t)} = \sum_{\tau=0}^t \gamma^\beta(1-\gamma^\beta)^\tau \hat{P}^\tau \hat{y}$$

$$x^{(t)} = \frac{y}{1+dk^\beta} + \frac{dk^\beta}{1+dk^\beta}\sum_{\tau=0}^t \gamma^\beta(1-\gamma^\beta)^\tau A\hat{P}^\tau \hat{y}.$$

The result follows from the fact that as $t \to \infty$, $\mu^{(t)} \to \mu^\beta$, and $x^{(t)} \to x^\beta$ (Theorem 2). □

The following lemma provides an estimate of the distance between fixed points μ^β and $\mu^{\beta'}$ in terms of $|\gamma^\beta - \gamma^{\beta'}|$.

Lemma 9: Given $0 \leq \beta' < \beta$, we have

$$\|\mu^{\beta} - \mu^{\beta'}\|_{2,nd} \leq \tau^{\star}(\gamma^{\beta'} - \gamma^{\beta})(1 + 4/\gamma^{\beta}).$$

Proof: Using (13)

$$\begin{aligned}\|\mu^{\beta} - \mu^{\beta'}\|_{2,nd} &= \left\|\sum_{\tau=0}^{\infty}\gamma^{\beta}(1-\gamma^{\beta})^{\tau}\hat{P}^{\tau}\hat{y} - \sum_{\tau=0}^{\infty}\gamma^{\beta'}(1-\gamma^{\beta'})^{\tau}\hat{P}^{\tau}\hat{y}\right\|_{2,nd}\\ &\leq \left\|\sum_{\tau=0}^{\infty}\left(\gamma^{\beta}(1-\gamma^{\beta})^{\tau} - \gamma^{\beta'}(1-\gamma^{\beta'})^{\tau}\right)\hat{P}^{\tau}\right\|_{2,nd}.\end{aligned}$$

Since

$$\sum_{\tau=0}^{\infty}\gamma^{\beta}(1-\gamma^{\beta})^{\tau} - \gamma^{\beta'}(1-\gamma^{\beta'})^{\tau} = 0$$

we have

$$\begin{aligned}\|\mu^{\beta} - \mu^{\beta'}\|_{2,nd} &\leq \left\|\sum_{\tau=0}^{\infty}\left(\gamma^{\beta}(1-\gamma^{\beta})^{\tau} - \gamma^{\beta'}(1-\gamma^{\beta'})^{\tau}\right)(\hat{P}^{\tau} - \hat{P}^{\star})\right\|_{2,nd}\\ &= \left\|\sum_{\tau=0}^{\infty}\sum_{s=\tau}^{\infty}\left((\gamma^{\beta})^2(1-\gamma^{\beta})^{s} - (\gamma^{\beta'})^2(1-\gamma^{\beta'})^{s}\right)(\hat{P}^{\tau} - \hat{P}^{\star})\right\|_{2,nd}\\ &= \left\|\sum_{s=0}^{\infty}\left((\gamma^{\beta})^2(1-\gamma^{\beta})^{s} - (\gamma^{\beta'})^2(1-\gamma^{\beta'})^{s}\right)\sum_{\tau=0}^{s}(\hat{P}^{\tau} - \hat{P}^{\star})\right\|_{2,nd}\\ &\leq \tau^{\star}\sum_{s=0}^{\infty}|(\gamma^{\beta})^2(1-\gamma^{\beta})^{s} - (\gamma^{\beta'})^2(1-\gamma^{\beta'})^{s}|.\end{aligned}$$

Hence, we wish to bound the sum

$$\Delta = \sum_{s=0}^{\infty}|(\gamma^{\beta})^2(1-\gamma^{\beta})^{s} - (\gamma^{\beta'})^2(1-\gamma^{\beta'})^{s}|.$$

Set

$$T = \left\lfloor 2\frac{\log\gamma^{\beta'} - \log\gamma^{\beta}}{\log(1-\gamma^{\beta}) - \log(1-\gamma^{\beta'})}\right\rfloor.$$

Note that

$$\begin{aligned}(\gamma^{\beta})^2(1-\gamma^{\beta})^{s} &\leq (\gamma^{\beta'})^2(1-\gamma^{\beta'})^{s}, \quad \text{if } s \leq T\\ (\gamma^{\beta})^2(1-\gamma^{\beta})^{s} &\geq (\gamma^{\beta'})^2(1-\gamma^{\beta'})^{s}, \quad \text{if } s > T.\end{aligned}$$

Holding γ^{β} fixed, it is easy to verify that T is nondecreasing as $\gamma^{\beta'} \downarrow \gamma^{\beta}$. Hence,

$$\begin{aligned}T &\leq 2\frac{\log\gamma^{\beta'} - \log\gamma^{\beta}}{\log(1-\gamma^{\beta}) - \log(1-\gamma^{\beta'})}\\ &\leq \lim_{\gamma^{\beta'}\downarrow\gamma^{\beta}} 2\frac{\log\gamma^{\beta'} - \log\gamma^{\beta}}{\log(1-\gamma^{\beta}) - \log(1-\gamma^{\beta'})}\\ &= 2(1-\gamma^{\beta})/\gamma^{\beta}.\end{aligned} \tag{14}$$

Using the preceding results

$$\begin{aligned}\Delta &= \sum_{s=0}^{T}\left((\gamma^{\beta'})^2(1-\gamma^{\beta'})^{s} - (\gamma^{\beta})^2(1-\gamma^{\beta})^{s}\right)\\ &\quad + \sum_{s=T+1}^{\infty}\left((\gamma^{\beta})^2(1-\gamma^{\beta})^{s} - (\gamma^{\beta'})^2(1-\gamma^{\beta'})^{s}\right)\\ &= \gamma^{\beta'} - \gamma^{\beta} - 2\gamma^{\beta'}(1-\gamma^{\beta'})^{T+1} + 2\gamma^{\beta}(1-\gamma^{\beta})^{T+1}\\ &\leq \gamma^{\beta'} - \gamma^{\beta} + 2\gamma^{\beta'}\left((1-\gamma^{\beta})^{T+1} - (1-\gamma^{\beta'})^{T+1}\right).\end{aligned}$$

Now, note that if $0 < a \leq b \leq 1$, for integer $\ell > 0$

$$\begin{aligned}b^{\ell} - a^{\ell} &= b^{\ell}(1 - (a/b)^{\ell})\\ &= b^{\ell}(1 - a/b)\sum_{i=0}^{\ell-1}(a/b)^{i}\\ &\leq \ell b^{\ell-1}(b-a)\\ &\leq \ell(b-a).\end{aligned}$$

Applying this inequality and using (14), we have

$$\begin{aligned}\Delta &\leq (\gamma^{\beta'} - \gamma^{\beta})\left(1 + 2(T+1)\gamma^{\beta'}\right)\\ &\leq (\gamma^{\beta'} - \gamma^{\beta})\left(1 + 2\gamma^{\beta'}(2/\gamma^{\beta} - 1)\right)\\ &\leq (\gamma^{\beta'} - \gamma^{\beta})(1 + 4\gamma^{\beta'}/\gamma^{\beta})\\ &\leq (\gamma^{\beta'} - \gamma^{\beta})(1 + 4/\gamma^{\beta}),\end{aligned}$$

which completes the proof. □

The following lemma characterizes the rate at which $\gamma_t \downarrow \gamma^{\beta}$.

Lemma 10: Assume that $\gamma^{\beta} \leq \gamma_0 \leq 1$. Then, $\{\gamma_t\}$ is a nonincreasing sequence and

$$|\gamma_t - \gamma^{\beta}| \leq \frac{(d-1)^t}{(1/\beta + \gamma^{\beta} + d - 1)^{2t}}.$$

Proof: Define the function

$$f(\gamma) = \frac{1}{1 + \frac{d-1}{1/\beta + \gamma}}.$$

Note that, from the definition of γ_t and (8), $\gamma_t = f(\gamma_{t-1})$. Further, from the definition of γ^{β} and (7), it is clear that $\gamma^{\beta} = f(\gamma^{\beta})$. Since $k_0 \leq k^{\beta}$, then $\gamma_0 \geq \gamma^{\beta}$, and since $k_t \uparrow k^{\beta}$ (from Lemma 1 (ii)), $\gamma_t \downarrow \gamma^{\beta}$. Also, if $\gamma \in [\gamma^{\beta}, 1]$

$$f'(\gamma) = \frac{d-1}{(1/\beta + \gamma + d - 1)^2} \leq \frac{d-1}{(1/\beta + \gamma^{\beta} + d - 1)^2}.$$

Then, by the Mean Value Theorem

$$\begin{aligned}|\gamma_t - \gamma^{\beta}| &= |f(\gamma_{t-1}) - f(\gamma^{\beta})|\\ &\leq \max_{\gamma\in[\gamma^{\beta},1]}|f'(\gamma)||\gamma_{t-1} - \gamma^{\beta}|\\ &\leq \frac{d-1}{(1/\beta + \gamma^{\beta} + d - 1)^2}|\gamma_{t-1} - \gamma^{\beta}|\\ &\leq \frac{(d-1)^t}{(1/\beta + \gamma^{\beta} + d - 1)^{2t}}|\gamma_0 - \gamma^{\beta}|.\end{aligned}$$

□

The following lemma establishes a bound on the distance between $x^{(t)}$ and $\bar{y}\mathbf{1}$ in terms of the distance between $\mu^{(t)}$ and μ^β.

Lemma 11:

$$\|x^{(t)} - \bar{y}\mathbf{1}\|_{2,n} \leq \gamma_t + \gamma^\beta\tau^\star + \|\mu^{(t)} - \mu^\beta\|_{2,nd}.$$

Proof: First, note that, using (13)

$$\begin{aligned}\|\mu^\beta - \hat{P}^\star\hat{y}\|_{2,nd} &= \left\|\sum_{\tau=0}^{\infty}\gamma^\beta(1-\gamma^\beta)^\tau\hat{P}^\tau - \hat{P}^\star\right\|_{2,nd}\\ &= \left\|\sum_{\tau=0}^{\infty}\gamma^\beta(1-\gamma^\beta)^\tau(\hat{P}^\tau - \hat{P}^\star)\right\|_{2,nd}\\ &= \left\|\sum_{\tau=0}^{\infty}(\gamma^\beta)^2\sum_{s=\tau}^{\infty}(1-\gamma^\beta)^s(\hat{P}^\tau - \hat{P}^\star)\right\|_{2,nd}\\ &\leq (\gamma^\beta)^2\sum_{s=0}^{\infty}(1-\gamma^\beta)^s\left\|\sum_{\tau=0}^{s}(\hat{P}^\tau - \hat{P}^\star)\right\|_{2,nd}\\ &\leq \gamma^\beta\tau^\star. \end{aligned} \tag{15}$$

Next, using Theorem 1, Lemma 7, and (15), we have

$$\begin{aligned}\bar{y}\mathbf{1} &= \lim_{\beta\to\infty} x^\beta\\ &= \lim_{\beta\to\infty}\frac{y}{1+dk^\beta} + \frac{dk^\beta}{1+dk^\beta}A\mu^\beta\\ &= \lim_{\beta\to\infty} A\mu^\beta\\ &= A\hat{P}^\star\hat{y}.\end{aligned}$$

Now

$$\begin{aligned}\|x^{(t)} - \bar{y}\mathbf{1}\|_{2,n} &\leq \frac{1}{1+dk_t}\|y - \bar{y}\mathbf{1}\|_{2,n}\\ &\quad + \frac{dk_t}{1+dk_t}\|A\mu^{(t)} - \bar{y}\mathbf{1}\|_{2,n}\\ &\leq \gamma_t + \|A\mu^{(t)} - \bar{y}\mathbf{1}\|_{2,n}\\ &\leq \gamma_t + \|A\mu^{(t)} - A\hat{P}^\star\hat{y}\|_{2,n}.\end{aligned}$$

By examining the structure of A, it follows from the Cauchy–Schwartz inequality that

$$\|A(\mu^{(t)} - \hat{P}^\star\hat{y})\|_{2,n} \leq \|\mu^{(t)} - \hat{P}^\star\hat{y}\|_{2,nd}.$$

Thus, using (15)

$$\begin{aligned}\|x^{(t)} - \bar{y}\mathbf{1}\|_{2,n} &\leq \gamma_t + \|\mu^{(t)} - \hat{P}^\star\hat{y}\|_{2,nd}\\ &\leq \gamma_t + \|\mu^\beta - \hat{P}^\star\hat{y}\|_{2,nd} + \|\mu^{(t)} - \mu^\beta\|_{2,nd}\\ &\leq \gamma_t + \gamma^\beta\tau^\star + \|\mu^{(t)} - \mu^\beta\|_{2,nd}.\end{aligned}$$ □

Proof of Theorem 3: Theorem 3 follows immediately from the following lemma.

Lemma 12: Fix $\epsilon > 0$, and pick β so that

$$\begin{aligned}&\beta \geq \max\left\{(2(1+\tau^\star)/\epsilon - 1/2)^2/4, 9/16\right\}, &&\text{if } d = 2\\ &\beta \geq \max\left\{2(1+\tau^\star)/(\epsilon(d-2)), 3/(d-2)\right\}, &&\text{if } d > 2.\end{aligned}$$

Assume that $k_0 \leq k^\beta$. Define

$$t^* = \left(1+2\sqrt{\beta}\right)\log\left(\frac{2+9\tau^\star\left(5+8\sqrt{\beta}\right)\left(1/2+\sqrt{\beta}\right)}{\epsilon/2}\right)$$

if $d = 2$, and

$$t^* = (1+(d-1)\beta)\log\left(\frac{2+4\tau^\star\left(5+4(d-1)\beta\right)}{\epsilon/2}\right)$$

if $d > 2$. Then, t^* is an ϵ-convergence time.

Proof: Let β_t be the value of β implied by k_t, that is, the unique value such that $k_t = k^{\beta_t}$. Define

$$\Delta_t = \|\mu^{(t)} - \mu^{\beta_t}\|_{2,nd}.$$

Note that the matrix $\hat{P}$ is doubly stochastic and hence nonexpansive under the $\|\cdot\|_{2,nd}$ norm. Then, from (9) and the fact that μ^{β_t} is a fixed point

$$\begin{aligned}\Delta_t &= \|\gamma_t\hat{y} + (1-\gamma_t)\hat{P}\mu^{(t-1)} - \gamma_t\hat{y} - (1-\gamma_t)\hat{P}\mu^{\beta_t}\|_{2,nd}\\ &= \|(1-\gamma_t)\hat{P}(\mu^{(t-1)} - \mu^{\beta_t})\|_{2,nd}\\ &\leq (1-\gamma_t)\|\mu^{(t-1)} - \mu^{\beta_t}\|_{2,nd}\\ &\leq (1-\gamma^\beta)\|\mu^{(t-1)} - \mu^{\beta_t}\|_{2,nd}\\ &\leq (1-\gamma^\beta)\left(\Delta_{t-1} + \|\mu^{\beta_{t-1}} - \mu^{\beta_t}\|_{2,nd}\right).\end{aligned}$$

Now, using Lemmas 9 and 10

$$\begin{aligned}\Delta_t &\leq (1-\gamma^\beta)\left(\Delta_{t-1} + \tau^*(\gamma_{t-1} - \gamma_t)(1+4/\gamma_t)\right)\\ &\leq (1-\gamma^\beta)\left(\Delta_{t-1} + \tau^*\left(\gamma_{t-1} - \gamma^\beta\right)\left(1+4/\gamma^\beta\right)\right)\\ &\leq (1-\gamma^\beta)\left(\Delta_{t-1} + \tau^*\alpha^{t-1}\left(1+4/\gamma^\beta\right)\right).\end{aligned} \tag{16}$$

Here, we define

$$\alpha = \begin{cases}1/(\gamma^\beta+1)^2, & \text{if } d = 2,\\ 1/(d-1), & \text{if } d > 2.\end{cases}$$

We would like to ensure that $\alpha < 1-\gamma^\beta$. For $d = 2$, some algebra reveals that this is true when $0 < \gamma^\beta < (\sqrt{5}-1)/2$. By the fact that $\beta \geq 9/16$ and Lemma 7, we have

$$0 < \gamma^\beta < \frac{1}{2\sqrt{\beta}+1} \leq 2/5 < \frac{\sqrt{5}-1}{2}.$$

For $d > 2$, using the fact that $\beta \geq 3/(d-2)$ and Lemma 7

$$\begin{aligned}0 < \frac{\alpha}{1-\gamma^\beta} &< \frac{(d-2)\beta}{(d-1)((d-2)\beta-1)}\\ &< \frac{3}{2(d-1)} \leq 3/4 < 1.\end{aligned} \tag{17}$$

By induction using (16), we have

$$\begin{aligned}\Delta_t &\leq (1-\gamma^\beta)^t + \tau^*\left(1+4/\gamma^\beta\right)\sum_{s=0}^{t-1}(1-\gamma^\beta)^{t-s}\alpha^s\\ &\leq (1-\gamma^\beta)^t\left(1+\tau^*\frac{1+4/\gamma^\beta}{1-\alpha/(1-\gamma^\beta)}\right).\end{aligned}$$

Now, notice that using the above results and Lemmas 9–11

$$\begin{aligned}\|x^{(t)} - \bar{y}\mathbf{1}\|_{2,n} &\\ \le \gamma_t + \gamma^\beta\tau^\star + \|\mu^{(t)} - \mu^\beta\|_{2,nd} &\\ \le \gamma_t + \gamma^\beta\tau^\star + \Delta_t + \|\mu^{\beta_t} - \mu^\beta\|_{2,nd} &\\ \le \gamma^\beta(1+\tau^\star) + (\gamma_t - \gamma^\beta) + \Delta_t + \tau^\star(\gamma_t - \gamma^\beta)(1+4/\gamma^\beta) &\\ \le \gamma^\beta(1+\tau^\star) + \alpha^t + (1-\gamma^\beta)^t\left(1+\tau^*\frac{1+4/\gamma^\beta}{1-\alpha/(1-\gamma^\beta)}\right) &\\ \quad + \tau^\star\alpha^t(1+4/\gamma^\beta) &\\ \le (1-\gamma^\beta)^t\left(2+\tau^*\left(1+4/\gamma^\beta\right)\left(1+\frac{1}{1-\alpha/(1-\gamma^\beta)}\right)\right) &\\ \quad + \gamma^\beta(1+\tau^\star). &\end{aligned}$$

When $d = 2$, using Lemma 7 and the fact that $\beta \ge (2(1+\tau^\star)/\epsilon - 1/2)^2/4$, we have

$$(1+\tau^\star)\gamma^\beta < \frac{1+\tau^\star}{2\sqrt{\beta}+1/2} \le \epsilon/2.$$

Similarly, when $d > 2$, since $\beta \ge 2(1+\tau^\star)/(\epsilon(d-2))$

$$(1+\tau^\star)\gamma^\beta < \frac{1+\tau^\star}{(d-2)\beta} \le \epsilon/2.$$

Thus, we will have $\|x^{(t)} - \bar{y}\mathbf{1}\|_{2,n} \le \epsilon$ if

$$(1-\gamma^\beta)^t\left(2+\tau^*\left(1+4/\gamma^\beta\right)\left(1+\frac{1}{1-\alpha/(1-\gamma^\beta)}\right)\right) \le \epsilon/2. \tag{18}$$

This will be true when

$$t \ge \frac{1}{\gamma^\beta}\log\left(\frac{2+\tau^*\left(1+4/\gamma^\beta\right)\left(1+\frac{1}{1-\alpha/(1-\gamma^\beta)}\right)}{\epsilon/2}\right). \tag{19}$$

(We have used the fact that $\log(1-\gamma^\beta) \le -\gamma^\beta$.) To complete the theorem, it suffices to show that t^* is an upper bound to the right-hand side of (19).

Consider the $d = 2$ case. From Lemma 7, it follows that

$$\begin{aligned}1/\gamma^\beta &< 1 + 2\sqrt{\beta}\\ 1 + 4/\gamma^\beta &< 5 + 8\sqrt{\beta}.\end{aligned}$$

Finally

$$\begin{aligned}\frac{1}{1-\alpha/(1-\gamma^\beta)} &= \frac{1}{1-\frac{1}{(1+\gamma^\beta)^2(1-\gamma^\beta)}}\\ &= \frac{1}{\gamma^\beta}\frac{(1+\gamma^\beta)^2(1-\gamma^\beta)}{1-\gamma^\beta-(\gamma^\beta)^2}\\ &= \frac{h(\gamma^\beta)}{\gamma^\beta}.\end{aligned}$$

Since $\beta \ge 9/16$, from Lemma 7, $\gamma^\beta \in (0, 1/2)$. It is easy to verify that for such γ^β, the rational function $h(\gamma^\beta)$ satisfies $h(\gamma^\beta) < h(1/2) = 9/2$. Thus,

$$\frac{1}{1-\alpha/(1-\gamma^\beta)} < \frac{9}{2\gamma^\beta} < 9/2 + 9\sqrt{\beta}.$$

For the $d > 2$ case, from Lemma 7, it follows that

$$\begin{aligned}1/\gamma^\beta &< 1 + (d-1)\beta\\ 1 + 4/\gamma^\beta &\le 5 + 4(d-1)\beta.\end{aligned}$$

Finally, using (17)

$$\frac{1}{1-\alpha/(1-\gamma^\beta)} < \frac{1}{1-3/4} = 4. \qquad \square$$

Proof of Theorem 4: Theorem 4 follows immediately from the following lemma.

Lemma 13: Fix $\epsilon > 0$, and pick β so that

$$\begin{aligned}\beta &\ge (2(1+\tau^\star)/\epsilon - 1/2)^2/4, && \text{if } d = 2\\ \beta &\ge 2(1+\tau^\star)/(\epsilon(d-2)), && \text{if } d > 2.\end{aligned}$$

Assume that $k_0 = k^\beta$, and define

$$t^* = \begin{cases}\left(1+2\sqrt{\beta}\right)\log(2/\epsilon), & \text{if } d = 2\\ \left(1+(d-1)\beta\right)\log(2/\epsilon), & \text{if } d > 2.\end{cases}$$

Then, t^* is an ϵ-convergence time.

Proof: Note that in this case, we have $k_t = k^\beta$ and $\gamma_t = \gamma^\beta$, for all $t \ge 0$. We will follow the same strategy as the proof of Lemma 12. Define

$$\Delta_t = \|\mu^{(t)} - \mu^\beta\|_{2,nd}.$$

Note that the matrix $\hat{P}$ is doubly stochastic and hence nonexpansive under the $\|\cdot\|_{2,nd}$ norm. Then, from (9) and the fact that μ^{β_t} is a fixed point

$$\begin{aligned}\Delta_t &= \|\gamma^\beta\hat{y} + (1-\gamma^\beta)\hat{P}\mu^{(t-1)} - \gamma^\beta\hat{y} - (1-\gamma^\beta)\hat{P}\mu^\beta\|_{2,nd}\\ &= \|(1-\gamma^\beta)\hat{P}(\mu^{(t-1)} - \mu^\beta)\|_{2,nd}\\ &\le (1-\gamma^\beta)\|\mu^{(t-1)} - \mu^\beta\|_{2,nd}\\ &= (1-\gamma^\beta)\Delta_{t-1}\\ &\le (1-\gamma^\beta)^t\end{aligned}$$

where the last step follows by induction.

Now, notice that, using the result and Lemmas 11

$$\begin{aligned}\|x^{(t)} - \bar{y}\mathbf{1}\|_{2,n} &\le \gamma^\beta(1+\tau^\star) + \Delta_t\\ &\le \gamma^\beta(1+\tau^\star) + (1-\gamma^\beta)^t.\end{aligned}$$

When $d = 2$, using Lemma 7 and the fact that $\beta \ge (2(1+\tau^\star)/\epsilon - 1/2)^2/4$, we have

$$(1+\tau^\star)\gamma^\beta < \frac{1+\tau^\star}{2\sqrt{\beta}+1/2} \le \epsilon/2.$$

Similarly, when $d > 2$, since $\beta \ge 2(1+\tau^\star)/(\epsilon(d-2))$

$$(1+\tau^\star)\gamma^\beta < \frac{1+\tau^\star}{(d-2)\beta} \le \epsilon/2.$$

Thus, we will have $\|x^{(t)} - \bar{y}\mathbf{1}\|_{2,n} \le \epsilon$ if

$$(1-\gamma^\beta)^t \le \epsilon/2.$$

This will be true when

$$t \ge \frac{1}{\gamma^\beta}\log(2/\epsilon). \tag{20}$$

(We have used the fact that $\log(1-\gamma^\beta) \leq -\gamma^\beta$.) To complete the theorem, it suffices to show that t^* is an upper bound to the right-hand side of (20).

Consider the $d=2$ case. From Lemma 7, it follows that

$$1/\gamma^\beta < 1 + 2\sqrt{\beta}.$$

For the $d>2$ case, from Lemma 7, it follows that

$$1/\gamma^\beta < 1 + (d-1)\beta. \qquad \square$$

Appendix III
Proof of Theorem 5

Theorem 5: For the cycle with n nodes, $\tau^\star \leq n/\sqrt{2}$.

Proof: Let $e^{ij} \in \mathbb{R}^{2n}$ be the vector with $\{i,j\}$th component equal to 1 and each other component equal to 0. It is easy to see that for any $\{i,j\} \in \vec{E}$

$$\sup_t \left\| \sum_{\tau=0}^{t} (\hat{P}^\tau - \hat{P}^\star) e^{ij} \right\|_{2,2n}^2 = \left\| \sum_{\tau=0}^{\lfloor n/2 \rfloor} (\hat{P}^\tau - \hat{P}^\star) e^{ij} \right\|_{2,2n}^2 \leq \frac{1}{2\sqrt{2}}.$$

We then have

$$\begin{aligned}
\tau^\star &= \sup_{t,\mu} \frac{\left\|\sum_{\tau=0}^{t}(\hat{P}^\tau - \hat{P}^\star)\mu\right\|_{2,2n}}{\|\mu\|_{2,2n}} \\
&= \sup_{t,\mu} \frac{\left\|\sum_{\tau=0}^{t}(\hat{P}^\tau - \hat{P}^\star)\sum_{\{i,j\}}\mu_{ij}e^{ij}\right\|_{2,2n}}{\left\|\sum_{\{i,j\}}\mu_{ij}e^{ij}\right\|_{2,2n}} \\
&\leq \sup_{t,\mu} \frac{\sum_{\{i,j\}}\mu_{ij}\left\|\sum_{\tau=0}^{t}(\hat{P}^\tau - \hat{P}^\star)e^{ij}\right\|_{2,2n}}{\left\|\sum_{\{i,j\}}\mu_{ij}e^{ij}\right\|_{2,2n}} \\
&\leq \sup_{\mu} \frac{\sum_{\{i,j\}}\mu_{ij}}{2\sqrt{2}\|\sum_{\{i,j\}}\mu_{ij}e^{ij}\|_{2,2n}} \\
&= \sup_{\mu} \frac{\sum_{\{i,j\}}\mu_{ij}}{2\sqrt{2}\sqrt{\sum_{\{i,j\}}\mu_{ij}^2/2n}} \\
&\leq \sup_{\mu} \frac{\sum_{\{i,j\}}\mu_{ij}}{2\sqrt{2}\sum_{\{i,j\}}|\mu_{ij}|/2n} \\
&\leq \frac{n}{\sqrt{2}}. \qquad \square
\end{aligned}$$

Acknowledgment

The authors wish to thank Balaji Prabhakar and Ashish Goel for their insights and comments.

References

[1] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks," in *Proc. ACM/IEEE In. Confe. Mobile Computing and Networking*, Boston, MA, Aug. 2000, pp. 56–67.

[2] S. R. Madden, M. J. Franklin, J. Hellerstein, and W. Hong, "Tag: A tiny aggregation service for ad hoc sensor networks," in *Proc. USENIX Symp. Operating Systems Design and Implementation*, Boston, MA, Dec. 2002, pp. 131–146.

[3] S. R. Madden, R. Szewczyk, M. J. Franklin, and D. Culler, "Supporting aggregate queries over ad-hoc wireless sensor networks," in *Proc. Workshop on Mobile Computing Systems and Applications*, Callicoon, NY, Jun. 2002, pp. 49–58.

[4] J. Zhao, R. Govindan, and D. Estrin, "Computing aggregates for monitoring wireless sensor networks," in *Proc. Int. Workshop on Sensor Net Protocols and Applications*, Anchorage, AK, May 2003, pp. 139–148.

[5] M. Bawa, H. Garcia-Molina, A. Gionis, and R. Motwani, Estimating Aggregates on a Peer-to-Peer Network, Stanford Univ. Database Group, Stanford, CA, 2003, Tech. Rep.

[6] M. Jelasity and A. Montresor, "Epidemic-style proactive aggregation in large overlay networks," in *Proc. 24th Int. Conf. Distributed Computing*, Tokyo, Japan, Mar. 2004, pp. 102–109.

[7] A. Montresor, M. Jelasity, and O. Babaoglu, "Robust aggregation protocols for large-scale overlay networks," in *Proc. Int. Conf. Dependable Systems and Networks*, Florence, Italy, Jun. 2004, pp. 19–28.

[8] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Proc. Int. Conf. Information Processing in Sensor Networks*, Los Angeles, CA, Apr. 2005, pp. 63–70.

[9] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed Average Consensus with Least-Mean-Square Deviation," May 2005, preprint.

[10] C. C. Moallemi and B. Van Roy, "Distributed optimization in adaptive networks," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Achölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[11] P. Flajolet and G. N. Martin, "Probabilistic counting algorithms for data base applications," *J. Comput. Syst. Sci.*, vol. 31, no. 2, pp. 182–209, 1985.

[12] J. Considine, F. Li, G. Kollios, and J. Byers, "Approximate aggregation techniques for sensor databases," in *Proc. Int. Conf. Data Engineering*, Boston, MA, Mar. 2004, p. 449.

[13] D. Mosk-Aoyama and D. Shah, Information Dissemination via Gossip: Applications to Averaging and Coding, 2005, preprint.

[14] J. N. Tsitsiklis, "Problems in decentralized decision-making and computation," Ph.D. dissertation, MIT, Cambridge, MA, 1984.

[15] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, "Convergence in multiagent coordination, consensus, and flocking," in *Proc. Joint 44th IEEE Conf. Decision and Control and European Control Conference*, Seveille, Spain, Dec. 2005, pp. 2996–3000.

[16] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *Foundations of Computer Science, 2003. Proc. 44th Annu, IEEE Symp. Theory of Computing*, Cambridge, MA, Oct. 2003, pp. 482–491.

[17] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: Design, analysis and applications," in *Proc. IEEE Infocom*, Miami, FL, Mar. 2005, pp. 1653–1664.

[18] P. Rusmevichientong and B. Van Roy, "An analysis of belief propagation on the turbo decoding graph with Gaussian densities," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 745–765, Feb. 2001.

[19] S. Tatikonda and M. I. Jordan, "Loopy belief propagation and Gibbs measures," in *Proc. Conf. Uncertainty in Artificial Intelligence*, Edmonton, AB, Canada, Aug. 2002, vol. 18, pp. 493–500.

[20] T. Heskes, "On the uniqueness of loopy belief propagation fixed points," *Neural Comput.*, vol. 16, no. 11, pp. 2379–2413, 2004.

[21] A. T. Ihler, J. W. Fisher, III, and A. S. Willsky, "Message errors in belief propagation," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 609–616.

[22] J. M. Mooij and H. J. Kappen, Sufficient Conditions for Convergence of Loopy Belief Propagation, Apr. 2005, preprint.

[23] G. Forney, F. Kschischang, and B. Marcus, "Iterative decoding of tail-biting trelisses," in *Proc. 1998 Information Theory Workshop*, Killarney, Ireland, Jun. 1998.

[24] S. M. Aji, G. B. Horn, and R. J. McEliece, "On the convergence of iterative decoding on graphs with a single cycle," in *Proc. Conf. Information Science and Systems*, Princeton, NJ, 1998.

[25] Y. Weiss and W. T. Freeman, "Correctness of local probability propagation in graphical models with loops," *Neural Comput.*, vol. 12, pp. 1–41, 2000.

[26] T. Richardson and R. Urbanke, "The capacity of low-density parity check codes under message-passing decoding," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[27] M. Bayati, D. Shah, and M. Sharma, "Maximum weight matching via max-product belief propagation," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 1763–1767.

[28] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *Neural Comput.*, vol. 13, pp. 2173–2200, 2001.

[29] J. K. Johnson, D. M. Malioutov, and A. S. Willsky, "Walk-sum interpretation and analysis of Gaussian belief propagation," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 579–586.

[30] C. C. Moallemi and B. Van Roy, Convergence of the Min-Sum Message Passing Algorithm for Quadratic Optimization, Mar. 2006, preprint.

[31] V. Saligrama, M. Alanyali, and O. Savas, Asynchronous Distributed Detection in Sensor Networks, 2005, preprint.

[32] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1997.

[33] M. J. Wainwright, T. Jaakkola, and A. S. Willsky, "Tree-based reparameterization framework for analysis of sum-product and related algorithms," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1120–1146, May 2003.

[34] S. Boyd, P. Diaconis, J. Sun, and L. Xiao, "Fastest mixing Markov chain on a path," *Amer. Math. Monthly*, vol. 113, no. 1, pp. 70–74, Jan. 2006.

[35] S. Boyd, P. Diaconis, P. Parillo, and L. Xiao, "Symmetry analysis of reversible Markov chains," *Internet Math.*, vol. 2, no. 1, pp. 31–71, 2005.

[36] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Mixing times for random walks on geometric random graphs," in *Proc. SIAM Workshop on Analytic Algorithms and Combinatorics*, Vancouver, BC, Canada, jan. 2005.

[37] P. Diaconis, S. Holmes, and R. Neal, "Analysis of a nonreversible Markov chain sampler," *Ann. Appl. Probab.*, vol. 10, no. 3, pp. 726–752, 2000.

[38] S. Roch, "Bounded fastest mixing," *Electron. Commun. Probab.*, vol. 10, pp. 282–296, 2005.