Network-based consensus averaging with general noisy channels

Ram Rajagopal* Martin J. Wainwright*,†
ramr@eecs.berkeley.edu wainwrig@stat.berkeley.edu
Department of Statistics†, and
Department of Electrical Engineering and Computer Sciences*
University of California, Berkeley
Berkeley, CA 94720

Technical Report
Department of Statistics, UC Berkeley

May 2008

Abstract

This paper focuses on the consensus averaging problem on graphs under general noisy channels. We study a particular class of distributed consensus algorithms based on damped updates, and using the ordinary differential equation method, we prove that the updates converge almost surely to exact consensus for finite variance noise. Our analysis applies to various types of stochastic disturbances, including errors in parameters, transmission noise, and quantization noise. Under a suitable stability condition, we prove that the error is asymptotically Gaussian, and we show how the asymptotic covariance is specified by the graph Laplacian. For additive parameter noise, we show how the scaling of the asymptotic MSE is controlled by the spectral gap of the Laplacian.

Keywords: Distributed averaging; sensor networks; message-passing; consensus protocols; gossip algorithms; stochastic approximation; graph Laplacian.

1 Introduction

Consensus problems, in which a group of nodes want to arrive at a common decision in a distributed manner, have a lengthy history, dating back to seminal work from over twenty years ago [8, 5, 18]. A particular type of consensus estimation is the distributed averaging problem, in which a group of nodes want to compute the average (or more generally, a linear function) of a set of values. Due to its applications in sensor and wireless networking, this distributed averaging problem has been the focus of substantial recent research. The distributed averaging problem can be studied either in continuous-time [16], or in the discrete-time setting (e.g., [13, 19, 6, 3, 9]). In both cases, there is now a fairly good understanding of the conditions under which various distributed averaging algorithms converge, as well as the rates of convergence for different graph structures.

The bulk of early work on consensus has focused on the case of perfect communication between nodes. Given that noiseless communication may be an unrealistic assumption for sensor networks, a more recent line of work has addressed the issue of noisy communication links. With imperfect observations, many of the standard consensus protocols might fail to reach an agreement. Xiao et

al. [20] observed this phenomenon, and opted to instead redefine the notion of agreement, obtaining a protocol that allows nodes to obtain a steady-state agreement, whereby all nodes are able to track but need not obtain consensus agreement. Schizas et al. [17] study distributed algorithms for optimization, including the consensus averaging problem, and establish stability under noisy updates, in that the iterates are guaranteed to remain within a ball of the correct consensus, but do not necessarily achieve exact consensus. Kashyap et al. [12] study consensus updates with the additional constraint that the value stored at each node must be integral, and establish convergence to quantized consensus. Fagnani and Zampieri [10] study the case of packet-dropping channels, and propose various updates that are guaranteed to achieve consensus. Yildiz and Scaglione [21] suggest coding strategies to deal with quantization noise, but do not establish convergence. In related work, Aysal et al [3] used probabilistic forms of quantization to develop algorithms that achieve consensus in expectation, but not in an almost sure sense.

In the current paper, we address the discrete-time average consensus problem for general stochastic channels. Our main contribution is to propose and analyze simple distributed protocols that are guaranteed to achieve exact consensus in an almost sure (sample-path) sense. These exactness guarantees are obtained using protocols with decreasing step sizes, which smooths out the noise factors. The framework described here is based on the classic ordinary differential equation method [15], and allows for the analysis of several different and important scenarios, namely:

- Noisy storage: stored values at each node are corrupted by noise, with known covariance structure.
- Noisy transmission: messages across each edge are corrupted by noise, with known covariance structure.
- Bit constrained channels: dithered quantization is applied to messages prior to transmission.

To the best of our knowledge, this is the first paper to analyze protocols that can achieve arbitrarily small mean-squared error (MSE) for distributed averaging with noise. By using stochastic approximation theory [4, 14], we establish almost sure convergence of the updates, as well as asymptotic normality of the error under appropriate stability conditions. The resulting expressions for the asymptotic variance reveal how different graph structures—ranging from ring graphs at one extreme, to expander graphs at the other—lead to different variance scaling behaviors, as determined by the eigenspectrum of the graph Laplacian [7].

The remainder of this paper is organized as follows. We begin in Section 2 by describing the distributed averaging problem in detail, and defining the class of stochastic algorithms studied in this paper. In Section 3, we state our main results on the almost-sure convergence and asymptotic normality of our protocols, and illustrate some of their consequences for particular classes of graphs. In particular, we illustrate the sharpness of our theoretical predictions by comparing them to simulation results, on various classes of graphs. Section 4 is devoted to the proofs of our main results, and we conclude the paper with discussion in Section 5. (This work was presented in part at the Allerton Conference on Control, Computing and Communication in September 2007.)

Comment on notation: Throughout this paper, we use the following standard asymptotic notation: for a functions f and g, the notation $f(n) = \mathcal{O}(g(n))$ means that $f(n) \leq Cg(n)$ for some

constant $C < \infty$; the notation $f(n) = \Omega(g(n))$ means that $f(n) \geq C' g(n)$ for some constant $C' > 0$, and $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$.

2 Problem set-up

In this section, we describe the distributed averaging problem, and specify the class of stochastic algorithms studied in this paper.

2.1 Consensus matrices and stochastic updates

Consider a set of $m = |V|$ nodes, each representing a particular sensing and processing device. We model this system as an undirected graph $G = (V, E)$, with processors associated with nodes of the graph, and the edge set $E \subset V \times V$ representing pairs of processors that can communicate directly. For each node v, we let $N(v) := \{u \in V \mid (v, u) \in E\}$ be its neighborhood set.

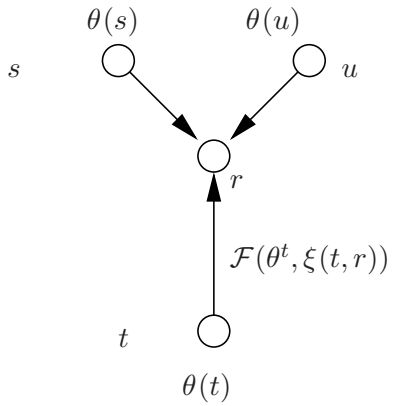


Figure 1. Illustration of the distributed protocol. Each node $t \in V$ maintains an estimate $\theta(t)$. At each round, for a fixed reference node $r \in V$, each neighbor $t \in N(r)$ sends the message $\mathcal{F}(\theta^t, \xi(t, r))$ along the edge $t \to r$.

Suppose that each vertex v makes a real-valued measurement $x(v)$, and consider the goal of computing the average $\overline{x} = \frac{1}{m} \sum_{v \in V} x(v)$. We assume that $|x(v)| \leq x_{\max}$ for all $v \in V$, as dictated by physical constraints of sensing. For iterations $n = 0, 1, 2, \ldots$, let $\theta^n = \{\theta^n(v),\ v \in V\}$ represent an m-dimensional vector of estimates. Solving the distributed averaging problem amounts to having θ^n converge to $\theta^* := \overline{x}\,\vec{1}$, where $\vec{1} \in \mathbb{R}^m$ is the vector of all ones. Various algorithms for distributed averaging [16, 6] are based on symmetric consensus matrices $L \in \mathbb{R}^{m \times m}$ with the properties:

$$L(v, v') \neq 0 \quad \text{only if } (v, v') \in E \tag{1a}$$

$$L\vec{1} = \vec{0}, \qquad \text{and} \tag{1b}$$

$$L \succeq 0. \tag{1c}$$

The simplest example of such a matrix is the *graph Laplacian*, defined as follows. Let $A \in \mathbb{R}^{m \times m}$

be the adjacency matrix of the graph G, i.e. the symmetric matrix with entries

$$A_{ij} \;=\; \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

and let $D = \text{diag}\{d_1, d_2, \ldots, d_m\}$ where $d_i := |N(i)|$ is the degree of node i. Assuming that the graph is connected (so that $d_i \geq 1$ for all i), the graph Laplacian is given by

$$L(G) \;=\; I - D^{-1/2} A D^{-1/2}. \tag{3}$$

Our analysis applies to the (rescaled) graph Laplacian, as well as to various weighted forms of graph Laplacian matrices [7].

Given a fixed choice of consensus matrix L, we consider the following family of updates, generating the sequence $\{\theta^n,\ n = 0, 1, 2 \ldots\}$ of m-dimensional vectors. The updates are designed to respect the neighborhood structure of the graph G, in the sense that at each iteration, the estimate $\theta^{n+1}(r)$ at a *receiving node* $r \in V$ is a function of only[1] the estimates $\{\theta^n(t),\ t \in N(r)\}$ associated with *transmitting nodes* t in the neighborhood of node r. In order to model noise and uncertainty in the storage and communication process, we introduce random variables $\xi(t,r)$ associated with the transmission link from t to r; we allow for the possibility that $\xi(t,r) \neq \xi(r,t)$, since the noise structure might be asymmetric.

With this set-up, we consider algorithms that generate a stochastic sequence $\{\theta^n, n = 0, 1, 2, \ldots\}$ in the following manner:

1. At time step $n = 0$, initialize $\theta^0(v) = x(v)$ for all $v \in V$.

2. For time steps $n = 0, 1, 2, \ldots$, each node $t \in V$ computes the random variables

$$Y^{n+1}(r,t) = \begin{cases} \theta^n(t), & \text{if } t = r \\ \mathcal{F}(\theta^n(t), \xi^{n+1}(t,r)) & \text{if } (t,r) \in E \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

   where $\mathcal{F}$ is the *communication-noise function* defining the model.

3. Generate estimate $\theta^{n+1} \in \mathbb{R}^m$ as

$$\theta^{n+1} \;=\; \theta^n + \epsilon_n \left[- \left(L \odot Y^{n+1} \right) \vec{1} \right], \tag{5}$$

   where $\odot$ denotes the Hadamard (elementwise) product between matrices, and $\epsilon_n > 0$ is a decaying step size parameter.

See Figure 1 for an illustration of the message-passing update of this protocol. In this paper, we focus on step size parameters ϵ_n that scale as $\epsilon_n = \Theta(1/n)$. On an elementwise basis, the update (5)

[1] In fact, our analysis is easily generalized to the case where $\theta^{n+1}(r)$ depends only on vertices $t \in N'(r)$, where $N'(r)$ is a (possibly random) subset of the full neighborhood set $N(v)$. However, to bring our results into sharp focus, we restrict attention to the case $N'(r) = N(r)$.

takes the form

$$\theta^{n+1}(r) \;=\; \theta^n(r) - \epsilon_n \left[L(r,r)\theta^n(r) + \sum_{t \in N(r)} L(r,t)\, \mathcal{F}(\theta^n(t), \xi^{n+1}(t,r)) \right].$$

2.2 Communication and noise models

It remains to specify the form of the the function $\mathcal{F}$ that controls the communication and noise model in the local computation step in equation (4).

Noiseless real number model: The simplest model, as considered by the bulk of past work on distributed averaging, assumes noiseless communication of real numbers. This model is a special case of the update (4) with $\xi^n(t,r) = 0$, and

$$\mathcal{F}(\theta^n(t), \xi^{n+1}(t,r)) = \theta^n(t). \tag{6}$$

Additive edge-based noise model (AEN**):** In this model, the term $\xi^n(t,r)$ is zero-mean additive random noise variable that is associated with the transmission $t \to r$, and the communication function takes the form

$$\mathcal{F}(\theta^n(t), \xi^{n+1}(t,r)) = \theta^n(t) + \xi^{n+1}(t,r). \tag{7}$$

We assume that the random variables $\xi^{n+1}(t,r)$ and $\xi^{n+1}(t',r)$ are independent for distinct edges (t',r) and (t,r), and identically distributed with zero-mean and variance $\sigma^2 = \mathrm{Var}(\xi^{n+1}(t,r))$.

Additive node-based noise model (ANN**):** In this model, the function $\mathcal{F}$ takes the same form (7) as the edge-based noise model. However, the key distinction is that for each $v' \in V$, we assume that

$$\xi^{n+1}(t,r) \;=\; \xi^{n+1}(t) \qquad \text{for all } r \in N(t), \tag{8}$$

where $\xi^{n+1}(t)$ is a single noise variable associated with node t, with zero mean and variance $\sigma^2 = \mathrm{Var}(\xi^n(t))$. Thus, the random variables $\xi^{n+1}(t,r)$ and $\xi^{n+1}(t,r')$ are all *identical* for all edges out-going from the transmitting node t.

Bit-constrained communication (BC**):** Suppose that the channel from node v' to v is bit-constrained, so that one can transmit at most B bits, which is then subjected to random dithering. Under these assumptions, the communication function $\mathcal{F}$ takes the form

$$\mathcal{F}(\theta(v'), \xi(v',v)) \;=\; Q_B\left(\theta(v') + \xi(v',v)\right), \tag{9}$$

where $Q_B(\cdot)$ represents the B-bit quantization function with maximum value M and $\xi(v',v)$ is random dithering. We assume that the random dithering is applied prior to transmission across the channel out-going from vertex v', so that $\xi(v',v) = \xi(v')$ is the same random variable across all neighbors $v \in N(v')$.

3 Main result and consequences

In this section, we first state our main result, concerning the stochastic behavior of sequence $\{\theta^n\}$ generated by the updates (5). We then illustrate its consequences for the specific communication and noise models described in Section 2.2, and conclude with a discussion of behavior for specific graph structures.

3.1 Statement of main result

Consider the factor $L \odot Y$ that drives the updates (5). An important element of our analysis is the conditional covariance of this update factor, denoted by $\Sigma = \Sigma_\theta$ and given by

$$\Sigma_\theta \;\; := \;\; \mathbb{E}\left[(L \odot Y(\theta, Z))\ \vec{1}\ \vec{1}^T\ (L \odot Y(\theta, Z))^T \;\mid\; \theta\right] - L\,\theta(L\theta)^T. \tag{10}$$

A little calculation shows that the $(i,j)^{th}$ element of this matrix is given by

$$\Sigma_\theta(i,j) \;\; = \;\; \sum_{k,\ell=1}^{m} L(i,k)L(j,\ell)\ \mathbb{E}\left[Y(i,k)Y(j,\ell) - \theta(k)\theta(\ell) \;\mid\; \theta\right]. \tag{11}$$

Moreover, the eigenstructure of the consensus matrix L plays an important role in our analysis. Since it is symmetric and positive semidefinite, we can write

$$L \;\; = \;\; UJU^T, \tag{12}$$

where U be an $m \times m$ orthogonal matrix with columns defined by unit-norm eigenvectors of L, and $J := \mathrm{diag}\{\lambda_1(L), \ldots, \lambda_m(L)\}$ is a diagonal matrix of eigenvalues, with

$$0 \;=\; \lambda_1(L) < \lambda_2(L) \leq \ldots < \lambda_m(L). \tag{13}$$

It is convenient to let $\widetilde{U}$ denote the $m \times (m-1)$ matrix with columns defined by eigenvectors associated with positive eigenvalues of L— that is, excluding column $U_1 = \vec{1}/\|\vec{1}\|_2$, associated with the zero-eigenvalue $\lambda_1(L) = 0$. With this notation, we have

$$\widetilde{J} \;=\; \mathrm{diag}\{\lambda_2(L), \ldots, \lambda_m(L)\} \;\; = \;\; \widetilde{U}^T L \widetilde{U}. \tag{14}$$

Theorem 1. *Consider the random sequence $\{\theta^n\}$ generated by the update (5) for some communication function $\mathcal{F}$, consensus matrix L, and step size parameter $\epsilon_n = \Theta(1/n)$.*

(a) In all cases, the sequence $\{\theta^n\}$ is a strongly consistent estimator of $\theta^* = \bar{x}\vec{1}$, meaning that $\theta^n \to \theta^*$ almost surely (a.s.).

(b) Furthermore, if the second smallest eigenvalue of the consensus matrix L satisfies $\lambda_2(L) > 1/2$ then

$$\sqrt{n}(\theta_n - \theta^*) \xrightarrow{d} N\left(0, U^T \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{P} \end{bmatrix} U\right), \tag{15}$$

where the $(m-1)\times(m-1)$ *matrix* $\widetilde{P}$ *is the solution of the continuous time Lyapunov equation*

$$\left(\widetilde{J}-\frac{I}{2}\right)\widetilde{P}+\widetilde{P}\left(\widetilde{J}-\frac{I}{2}\right)^{T} \;=\; \widetilde{\Sigma}_{\theta^*} \tag{16}$$

where $\widetilde{J}$ *is the diagonal matrix* (14)*, and* $\widetilde{\Sigma}_{\theta^*}=\widetilde{U}^T\Sigma_{\theta^*}\widetilde{U}$ *is the transformed version of the conditional covariance* (10)*.*

Theorem 1(a) asserts that the sequence $\{\theta^n\}$ is a strongly consistent estimator of the average. As opposed to weak consistency, this result guarantees that for almost any realization of the algorithm, the associated sample path converges to the exact consensus solution. Theorem 1(b) establishes that for appropriate choices of consensus matrices, the rate of MSE convergence is of order $1/n$, since the $\sqrt{n}$-rescaled error converges to a non-degenerate Gaussian limit. Such a rate is to be expected in the presence of sufficient noise, since the number of observations received by any given node (and hence the inverse variance of estimate) scales as n. The solution of the Lyapunov equation (16) specifies the precise form of this asymptotic covariance, which (as we will see) depends on the graph structure.

3.2 Some consequences

Theorem 1 can be specialized to particular noise and communication models. Here we derive some of its consequences for the AEN, ANN and BC models. For any model for which Theorem 1(b) holds, we define the average mean-squared error as

$$\mathrm{AMSE}(L;\theta^*) \;:=\; \frac{1}{m}\,\mathrm{trace}(\widetilde{P}(\theta^*)), \tag{17}$$

corresponding to asymptotic error variance, averaged over nodes of the graph.

Corollary 1 (Asymptotic MSE for specific models)**.** *Given a consensus matrix* L *with second-smallest eigenvalue* $\lambda_2(L)>\frac{1}{2}$*, the sequence* $\{\theta^n\}$ *is a strongly consistent estimator of the average* θ^**, with asymptotic MSE characterized as follows:*

*(a) For the additive edge-based noise (*AEN*) model* (7)*:*

$$\mathrm{AMSE}(L;\theta^*) \;\leq\; \frac{\sigma^2}{m}\sum_{i=2}^{m}\left[\frac{\max_{j=1,\ldots,m}\sum_{k\neq j}L^2(j,k)}{2\lambda_i(L)-1}\right]. \tag{18}$$

*(b) For the additive node-based noise (*ANN*) model* (8) *and the bit-constrained (*BC*) model* (9)*:*

$$\mathrm{AMSE}(L;\theta^*) \;=\; \frac{\sigma^2}{m}\sum_{i=2}^{m}\left[\frac{[\lambda_i(L)]^2}{2\lambda_i(L)-1}\right], \tag{19}$$

where the variance term σ^2 *is given by the quantization noise* $\mathbb{E}\left[Q_B(\theta+\xi)^2-\theta^2 \mid \theta\right]$ *for the* BC *model, and the noise variance* $\mathrm{Var}(\xi(v'))$ *for the* ANN *model.*

Proof. The essential ingredient controlling the asymptotic MSE is the conditional covariance matrix Σ_{θ^*}, which specifies $\widetilde{P}$ via the Lyapunov equation (16). For analyzing model AEN, it is useful to establish first the following auxiliary result. For each $i = 1, \ldots, m-1$, we have

$$\widetilde{P}_{ii} \;\; \leq \;\; \frac{|||\Sigma_{\theta^*}|||_2}{2\,\lambda_{i+1}(L) - 1}, \tag{20}$$

where $|||\Sigma_{\theta^*}|||_2 = |||\Sigma|||_2$ is the spectral norm (maximum eigenvalue for a positive semdefinite symmetric matrix). To see this fact, note that

$$\widetilde{U}^T \Sigma \widetilde{U} \;\preceq\; \widetilde{U}^T \left[|||\Sigma|||_2 I\right] \widetilde{U} \;=\; |||\Sigma|||_2 I.$$

Since $\widetilde{P}$ satisfies the Lyapunov equation, we have

$$\left(\widetilde{J} - \frac{I}{2}\right) \widetilde{P} + \widetilde{P} \left(\widetilde{J} - \frac{I}{2}\right)^T \;\; \preceq \;\; |||\Sigma|||_2 I.$$

Note that the diagonal entries of the matrix $\left(\widetilde{J} - \frac{I}{2}\right) \widetilde{P} + \widetilde{P} \left(\widetilde{J} - \frac{I}{2}\right)^T$ are of the form $(2\lambda_{i+1} - 1)\; \widetilde{P}_{ii}$. The difference between the RHS and LHS matrices constitute a positive semidefinite matrix, which must have a non-negative diagonal, implying the claimed inequality (20).

In order to use the bound (20), it remains to compute or upper bound the spectral norm $|||\Sigma|||_2$, which is most easily done using the elementwise representation (11).

(a) For the AEN model (7), we have

$$\mathbb{E}\left[Y(i,k)Y(j,\ell) - \theta_k \theta_\ell \;\mid\; \theta\right] \;\; = \;\; \mathbb{E}\left[\xi(i,k)\xi(j,\ell)\right]. \tag{21}$$

Since we have assumed that the random variables $\xi(i,k)$ on each edge (i,k) are i.i.d., with zero-mean and variance σ^2, we have

$$\mathbb{E}\left[Y(i,k)Y(j,\ell) - \theta(k)\theta(\ell) \;\mid\; \theta\right] \;\; = \;\; \begin{cases} \sigma^2 & \text{if } (i,k) = (j,\ell) \text{ and } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, from the elementwise expression (11), we conclude that Σ is diagonal, with entries

$$\Sigma(j,j) \;\; = \;\; \sigma^2 \sum_{k \neq j} L^2(k,j),$$

so that $|||\Sigma|||_2 = \sigma^2 \max_{j=1,\ldots,m} \sum_{k \neq j} L_{jk}^2$, which establishes the claim (18).

(b) For the BC model (9), we have

$$\mathbb{E}\left[Y(i,k)Y(j,\ell) - \theta(k)\theta(\ell) \;\mid\; \theta\right] \;\; = \;\; \begin{cases} \sigma^2_{\text{qnt}} & \text{if } i = j \text{ and } k = \ell \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

where $\sigma^2_{\text{qnt}} := \mathbb{E}\left[Q_B(\theta + \xi)^2 - \theta^2 \;\mid\; \theta\right]$ is the quantization noise. Therefore, we have $\Sigma(\theta^*) = \sigma^2_{\text{qnt}} L^2$, and using the fact that $\widetilde{U}$ consists of eigenvectors of L (and hence also L^2, the Lyapunov equation (16) takes the form

$$\left(\widetilde{J} - \frac{I}{2}\right) \widetilde{P} + \widetilde{P} \left(\widetilde{J} - \frac{I}{2}\right)^T \;\; = \;\; \sigma^2_{\text{qnt}}\, (\widetilde{J})^2,$$

which has the explicit diagonal solution $\widetilde{P}$ with entries $\widetilde{P}_{ii} = \frac{\sigma^2_{\text{qnt}} \lambda^2_{i+1}(L)}{2\lambda_{i+1}(L)-1}$. Computing the asymptotic MSE $\frac{1}{m}\sum_{i=1}^{m-1} \widetilde{P}_{ii}$ yields the claim (19). The proof of the same claim for the ANN model is analogous. □

3.3 Scaling behavior for specific graph classes

We can obtain further insight by considering Corollary 1 for specific graphs, and particular choices of consensus matrices L. For a fixed graph G, consider the graph Laplacian $L(G)$ defined in equation (3). It is easy to see that $L(G)$ is always positive semi-definite, with minimal eigenvalue $\lambda_1(L(G)) = 0$, corresponding to the constant vector. For a connected graph, the second smallest eigenvector $L(G)$ is strictly positive [7]. Therefore, given an undirected graph G that is connected, the most straightforward manner in which to obtain a consensus matrix L satisfying the conditions of Corollary 1 is to rescale the graph Laplacian $L(G)$, as defined in equation (3), by its second smallest eigenvalue $\lambda_2(L(G))$, thereby forming the rescaled consensus matrix

$$R(G) \quad := \quad \frac{1}{\lambda_2(L(G))}\, L(G). \tag{23}$$

with $\lambda_2(R(G)) = 1 > \frac{1}{2}$.

With this choice of consensus matrix, let us consider the implications of Corollary 1(b), in application to the additive node-based noise (ANN) model, for various graphs. We begin with a simple lemma, proved in Appendix A, showing that, up to constants, the scaling behavior of the asymptotic MSE is controlled by the second smallest eigenvalue $\lambda_2(L(G))$.

Lemma 1. *For any connected graph G, using the rescaled Laplacian consensus matrix* (23)*, the asymptotic MSE for the* ANN *model* (8) *satisfies the bounds*

$$\frac{\sigma^2}{2\lambda_2(L(G))} \;\leq\; \text{AMSE}(R(G); \theta^*) \;\leq\; \frac{\sigma^2}{\lambda_2(L(G))}, \tag{24}$$

where $\lambda_2(L(G))$ is the second smallest eigenvalue of the graph.

Combined with known results from spectral graph theory [7], Lemma 1 allows us to make specific predictions about the number of iterations required, for a given graph topology of a given size m, to reduce the asymptotic MSE to any $\delta > 0$: in particular, the required number of iterations scales as

$$n \quad = \quad \Theta\left(\frac{\sigma^2}{\lambda_2(L(G))}\,\frac{1}{\delta}\right). \tag{25}$$

Note that this scaling is similar but different from the scaling of noiseless updates [6, 9], where the MSE is (with high probability) upper bounded by δ for $n = \Theta(\frac{\log(1/\delta)}{-\log(1-\lambda_2(L(G)))})$, which scales as

$$n \quad = \quad \Theta\left(\frac{\log(1/\delta)}{\lambda_2(L(G))}\right), \tag{26}$$

for a decaying spectral gap $\lambda_2(L(G)) \to 0$.

3.4 Illustrative simulations

We illustrate the predicted scaling (25) by some simulations on different classes of graphs. For all experiments reported here, we set the step size parameter $\epsilon_n = \frac{1}{n+100}$. The additive offset serves to ensure stability of the updates in very early rounds, due to the possibly large gain specified by the rescaled Laplacian (23). We performed experiments for a range of graph sizes, for the additive node noise (ANN) model (8), with noise variance $\sigma^2 = 0.1$ in all cases. For each graph size m, we measured the number of iterations n required to reach a fixed level δ of mean-squared error.

3.4.1 Cycle graph

Consider the ring graph C_m on m vertices, as illustrated in Figure 2(a). Panel (b) provides a log-log plot of the MSE versus the iteration number n; each trace corresponds to a particular sample path. Notice how the MSE over each sample coverges to zero. Moreover, since Theorem 1 predicts that the MSE should drop off as $1/n$, the linear rate shown in this log-log plot is consistent. Figure 2(c) plots the number of iterations (vertical axis) required to achieve a given constant MSE versus the size of the ring graph (horizontal axis). For the ring graph, it can be shown (see Chung [7]) that the second smallest eigenvalue scales as $\lambda_2(L(C_m)) = \Theta(1/m^2)$, which implies that the number of iterations to achieve a fixed MSE for a ring graph with m vertices should scale as $n = \Theta(m^2)$. Consistent with this prediction, the plot in Figure 2(c) shows a quadratic scaling; in particular, note the excellent agreement between the theoretical prediction and the data.

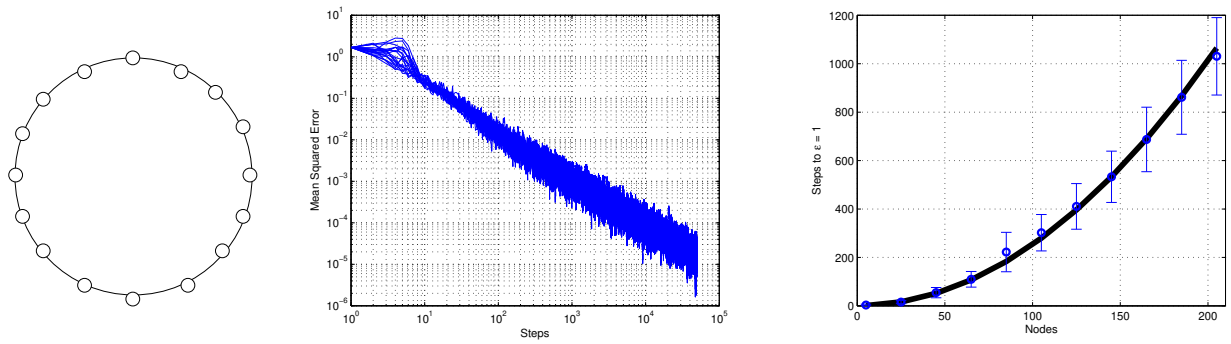
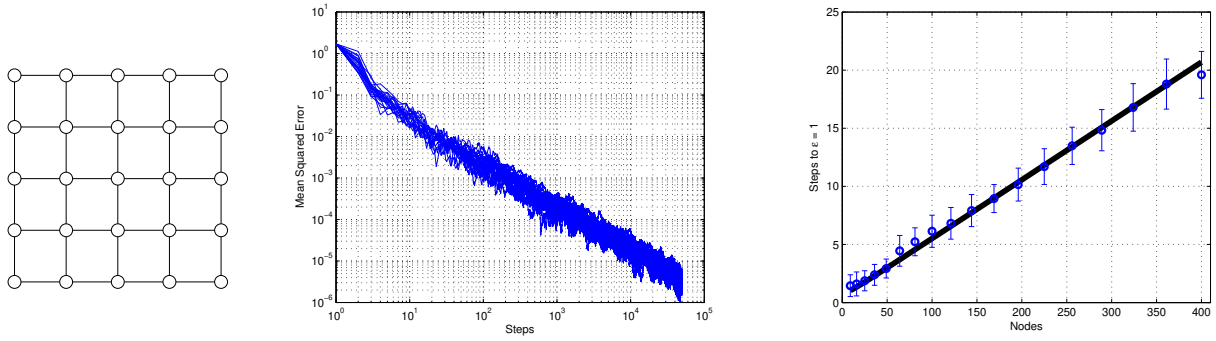


Figure 2. Comparison of empirical simulations to theoretical predictions for the ring graph in panel (a). (b) Sample path plots of log MSE versus log iteration number: as predicted by the theory, the log MSE scales linearly with log iterations. (c) Plot of number of iterations (vertical axis) required to reach a fixed level of MSE versus the graph size (horizontal axis). For the ring graph, this quantity scales quadratically in the graph size, consistent with Corollary 1.

3.4.2 Lattice model

Figure 3(a) shows the two-dimensional four nearest-neighbor lattice graph with m vertices, denoted F_m. Again, panel (b) corresponds to a log-log plot of the MSE versus the iteration number n, with each trace corresponding to a particular sample path, again showing a linear rate of convergence

to zero. Panel (c) shows the number of iterations required to achieve a constant MSE as a function of the graph size. For the lattice, it is known [7] that $\lambda_2(L(F_m)) = \Theta(1/m)$, which implies that the critical number of iterations should scale as $n = \Theta(m)$. Note that panel (c) shows linear scaling, again consistent with the theory.



(a) (b) (c)

Figure 3. Comparison of empirical simulations to theoretical predictions for the four nearest-neighbor lattice (panel (a)). (b) Sample path plots of log MSE versus log iteration number: as predicted by the theory, the log MSE scales linearly with log iterations. (c) Plot of number of iterations (vertical axis) required to reach a fixed level of MSE versus the graph size (horizontal axis). For the lattice, graph, this quantity scales linearly in the graph size, consistent with Corollary 1.

3.4.3 Expander graphs

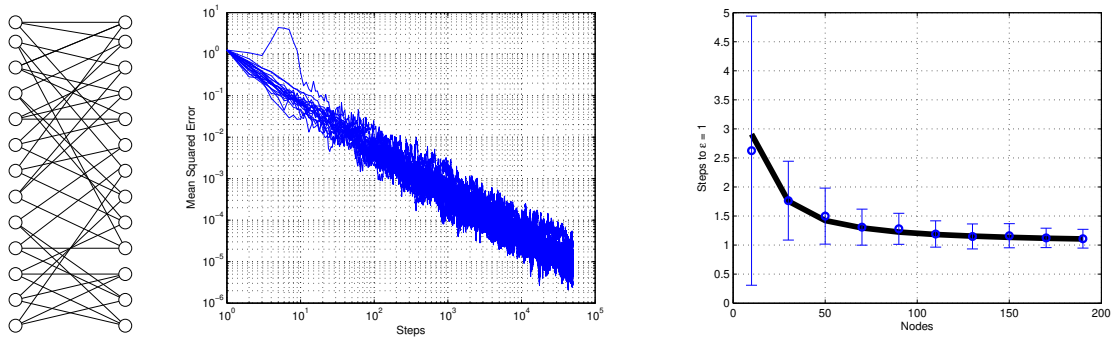
Consider a bipartite graph $G = (V_1, V_2, E)$, with $m = |V_1| + |V_2|$ vertices and edges joining only vertices in V_1 to those in V_2, and constant degree d; see Figure 4(a) for an illustration with $d = 3$. A bipartite graph of this form is an expander [1, 2, 7] with parameters $\alpha, \delta \in (0, 1)$, if for all subsets $S \subset V_1$ of size $|S| \leq \alpha |V_1|$, the neighborhood set of S—namely, the subset

$$N(S) \quad := \quad \{t \in V_2 \ \mid \ (s,t) \quad \text{for some } s \in S\},$$

has cardinality $|N(S)| \geq \delta d |S|$. Intuitively, this property guarantees that each subset of V_1, up to some critical size, "expands" to a relatively large number of neighbors in V_2. (Note that the maximum size of $|N(S)|$ is $d|S|$, so that δ close to 1 guarantees that the neighborhood size is close to its maximum, for all possible subsets S.) Expander graphs have a number of interesting theoretical properties, including the property that $\lambda_2(L(K_m)) = \Theta(1)$—that is, a bounded spectral gap [1, 7].

In order to investigate the behavior of our algorithm for expanders, we construct a random bipartite graph as follows: for an even number of nodes m, we split them into two subsets $V_i, i = 1, 2$, each of size $m/2$. We then fix a degree d, construct a random matching on $d\frac{m}{2}$ nodes, and use it connect the vertices in V_1 to those in V_2. This procedure forms a random bipartite d-regular graph; using the probabilistic method, it can be shown to be an edge-expander with with probability $1 - o(1)$, as the graph size tends to infinity [1, 11].

Given the constant spectral gap $\lambda_2(L(K_m)) = \Theta(1)$, the scaling in number of iterations to achieve constant MSE is $n = \Theta(1)$. This theoretical prediction is compared to simulation results in



(a) (b) (c)

Figure 4. Comparison of empirical simulations to theoretical predictions for the bipartite expander graph in panel (a). (b) Sample path plots of log MSE versus log iteration number: as predicted by the theory, the log MSE scales linearly with log iterations. (c) Plot of number of iterations (vertical axis) required to reach a fixed level of MSE versus the graph size (horizontal axis). For an expander, this quantity remains essentially constant with the graph size, consistent with Corollary 1.

Figure 4; note how the number of iterations soon settles down to a constant, as predicted by the theory.

4 Proof of Theorem 1

We now turn to the proof of Theorem 1. The basic idea is to relate the behavior of the stochastic recursion (5) to an ordinary differential equation (ODE), and then use the ODE method [15] to analyze its properties. The ODE involves a function $t \mapsto \theta_t \in \mathbb{R}^m$, with its specific structure depending on the communication and noise model under consideration. For the AEN and ANN models, the relevant ODE is given by

$$\frac{d\theta_t}{dt} = -L\theta_t. \tag{27}$$

For the BC model, the approximating ODE is given by

$$\frac{d\theta_t}{dt} = -L\, C_M(\theta_t) \qquad \text{with } C_M(u) := \begin{cases} u & \text{if } |u| < M \\ -M & \text{if } u \leq -M \\ +M & \text{if } u \geq +M. \end{cases} \tag{28}$$

In both cases, the ODE must satisfy the initial condition $\theta_0(v) = x(v)$.

4.1 Proof of Theorem 1(a)

The following result connects the discrete-time stochastic process $\{\theta^n\}$ to the deterministic ODE solution, and establishes Theorem 1(a):

Lemma 2. *The ODEs* (27) *and* (28) *each have* $\theta^* = \overline{x}\vec{1}$ *as their unique stable fixed point. Moreover, for all* $\delta > 0$*, we have*

$$\mathbb{P}\left(\limsup_{n\to\infty} \|\theta_n - \theta_{t_n}\| > \delta\right) \;=\; 0, \qquad \text{for } t_n = \textstyle\sum_{k=1}^{n} \frac{1}{k}, \tag{29}$$

which implies that $\theta^n \to \theta^*$ *almost surely.*

Proof. We prove this lemma by using the ODE method and stochastic approximation—in particular, Theorem 1 from Kushner and Yin [14], which connects stochastic recursions of the form (5) to the ordinary differential equation $d\theta_t/dt = \mathbb{E}_\xi[n\ (L \odot Y(\theta_t, \xi)) \mid \theta_t]$. Using the definition of Y in terms of $\mathcal{F}$, for the AEN and ANN models, we have

$$\mathbb{E}_\xi\left[\mathcal{F}(\theta(v), \xi(v,r)) \mid \theta(v)\right] \;=\; \theta(v),$$

from which we conclude that with the stepsize choice $\epsilon_m = \Theta(1/m)$, we have

$$\mathbb{E}_\xi[n\ (L \odot Y(\theta_t, \xi)) \mid \theta_t] \;=\; -L\theta_t.$$

By our assumptions on the eigenstructure of L, the system $d\theta_t/dt = -L\,\theta_t$ is globally asymptotically stable, with a line of fixed points $\{\theta \in \mathbb{R}^m \mid L\theta = 0\}$. Given the initial condition $\theta_0(v) = x(v)$, we conclude that $\theta^* = \overline{x}\vec{1}$ is the unique asymptotically fixed point of the ODE, so that the claim (29) follows from Kushner and Yin [14].

For the BC model, the analysis is somewhat more involved, since the quantization function saturates the output at $\pm M$. For the dithered quantization model (9), we have

$$\mathbb{E}_\xi[n\ (L \odot Y(\theta_t, \xi)) \mid \theta_t] \;=\; -L\,C_M(\theta_t),$$

where $C_M(\cdot)$ is the saturation function (28). We now claim that θ^* is also the unique asymptotically stable fixed point of the ODE $d\theta_t/dt = -L\,C_M(\theta_t)$ subject to the initial condition $\theta_0(v) = x(v)$. Consider the eigendecomposition $L = UJU^T$, where $J = \mathrm{diag}\{0, \lambda_2(L), \ldots, \lambda_m(L)\}$. Define the rotated variable $\gamma_t := U^T\theta_t$, so that the ODE (28) can be re-written as

$$d\gamma_t(1)/dt \;=\; 0 \tag{30a}$$

$$d\gamma_t(k)/dt \;=\; -\lambda_k(L)U_k^T C_M(U\gamma_t), \qquad \text{for } k = 2, \ldots, m, \tag{30b}$$

where U_k denotes the k^{th} column of U.

Note that $U_1 = \vec{1}/\|\vec{1}\|_2$, since it is associated with the eigenvalue $\lambda_1(L) = 0$. Consequently, the solution to equation (30a) takes the form

$$\gamma_t(1) = U_1^T\theta_0 \;=\; \sqrt{m}\,\overline{x}, \tag{31}$$

with unique fixed point $\gamma^*(1) = \sqrt{m}\,\overline{x}$, where $\overline{x} := \frac{1}{m}\sum_{i=1}^m x(i)$ is the average value,

A fixed point $\gamma^* \in \mathbb{R}^m$ for equations (30b) requires that $U_k^T C_M(U\gamma^*) = 0$, for $k = 2, \ldots, m$. Given that the columns of U form an orthogonal basis, this implies that $C_M(U\gamma^*) = \alpha\vec{1}$ for some constant $\alpha \in \mathbb{R}$, or equivalently (given the connection $U\gamma^* = \theta^*$)

$$C_M(\theta^*) \;=\; \alpha\vec{1}. \tag{32}$$

Given the piecewise linear nature of the saturation function, this equality implies either that the fixed point satisfies the elementwise inequality $\theta^* > M$ (if $\alpha = M$); or the elementwise inequality $\theta^* < -M$ (if $\alpha = -M$); or as the final option, the $\theta^* = \alpha$ when $\alpha \in (-M, +M)$. But from equation (31), we know that $\gamma^*(1) = \sqrt{m}\,\overline{x} \in [-M\sqrt{m},\ +M\sqrt{m}]$. But we also have $\gamma^*(1) = \frac{\vec{1}^T}{\sqrt{m}}\theta^*$ by definition, so that putting together the pieces yields

$$-M \ < \ \frac{\vec{1}\,\theta^*}{m} \ < \ M, \tag{33}$$

Thus the only possibility is that $\theta^* = \alpha\vec{1}$ for some constant $\alpha \in (-M, +M)$, and the relation $U\gamma^* = \alpha\vec{1}$ implies that $\alpha = \gamma^*(1)/\sqrt{m} = \overline{x}$, which establishes the claim. □

4.2 Proof of Theorem 1(b)

We analyze the update (5) using results from Benveniste et al [4]. In particular, given the stochastic iteration $\theta^{n+1} = \theta^n + \epsilon_n H(\theta^n, Y^{n+1})$, define the expectation $h(\theta) = \mathbb{E}\left[H(\theta, X)\right]$, its Jacobian matrix $\nabla h(\theta)$, and the covariance matrix $\Sigma(\theta) = \mathbb{E}\left[(H(\theta, X) - h(\theta))(H(\theta, X) - h(\theta))^T\right]$. Then Theorem 3 (p. 110) of Benveniste et al [4] asserts that as long as the eigenvalues $\lambda(\nabla h(\theta))$ are strictly below $-1/2$. then

$$\sqrt{n}\,(\theta_n - \theta^*) \stackrel{d}{\to} N(0, Q), \tag{34}$$

where the covariance matrix Q is the unique solution to the Lyapunov equation

$$\left(\frac{I}{2} + \nabla h(\theta^*)\right) Q + Q \left(\frac{I}{2} + \nabla h(\theta^*)\right)^T + \Sigma_{\theta^*} = 0. \tag{35}$$

We begin by computing the conditional distribution $h(\theta)$; for the models AEN and ANN it takes the form

$$h(\theta) \ \ = \ \ -L\,\theta \tag{36}$$

since the conditional expectation of the random matrix Y is given by $\mathbb{E}[Y \mid \theta] = \theta\,\vec{1}$. For the BC model, since the quantization is finite with maximum value M, the expectation is given by

$$h(\theta) = -L\,C_M(\theta) \tag{37}$$

where the saturation function was defined previously (28). In addition, we computed form of the the covariance matrix Σ_{θ^*} previously (10). Finally, we note that

$$\nabla h(\theta^*) \ \ = \ \ -L \tag{38}$$

for all three models. (This fact is immediate for models AEN and ANN; for the BC model, note that Theorem 1(a) guarantees that θ^* falls in the middle linear portion of the saturation function.)

We cannot immediately conclude that asymptotic normality (34) holds, because the matrix L has a zero eigenvalue ($\lambda_1(L) = 0$). However, let us decompose $L = UJU^T$ where U is the matrix with unit norm columns as eigenvectors, and $J = \mathrm{diag}\{0, \lambda_2(L), \ldots, \lambda_m(L)\}$. Let $\widetilde{U}$ denote the

$m \times (m-1)$ matrix obtained by deleting the first column of U. Defining the $(m-1)$ vector $\beta^n = \widetilde{U}^T \theta^n$, we can rewrite the update in $(m-1)$-dimensional space as

$$\beta^{n+1} \;=\; \beta^n + \frac{1}{n}\left[-U^T\left(L \odot Y^{n+1}(\theta^n)\right)\vec{1}\right], \tag{39}$$

for which the new effective h function is given by $\tilde{h}(\beta) = -\widetilde{J}\beta$, with $\widetilde{J} = \mathrm{diag}\{\lambda_2(L), \ldots, \lambda_m(L)\}$. Since $\lambda_2(L) > \frac{1}{2}$ by assumption, the asymptotic normality (34) applies to this reduced iteration, so that we can conclude that

$$\sqrt{n}\,(\beta^n - \beta^*) \;\xrightarrow{d}\; N(0, \widetilde{P})$$

where $\widetilde{P}$ solves the Lyapunov equation

$$\left(\widetilde{J} - \frac{I}{2}\right)\widetilde{P} + \widetilde{P}\left(\widetilde{J} - \frac{I}{2}\right)^T \;=\; \widetilde{U}^T \Sigma_{\theta^*} \widetilde{U}.$$

We conclude by noting that the asymptotic covariance of θ^n is related to that of β^n by the relation

$$P \;=\; U^T \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{P} \end{bmatrix} U, \tag{40}$$

from which Theorem 1(b) follows.

5 Discussion

This paper analyzed the convergence and asymptotic behavior of distributed averaging algorithms on graphs with general noise models. Using suitably damped updates, we showed that it is possible to obtain exact consensus, as opposed to approximate or near consensus, even in the presence of noise. We guaranteed almost sure convergence of our algorithms under fairly general conditions, and moreover, under suitable stability conditions, we showed that the error is asymptotically normal, with a covariance matrix that can be predicted from the structure of the consensus operator. We provided a number of simulations that illustrate the sharpness of these theoretical predictions. Although the current paper has focused exclusively on the averaging problem, the methods of analysis in this paper are applicable to other types of distributed inference problems, such as computing quantiles or order statistics, as well as computing various types of M-estimators. Obtaining analogous results for more general problems of distributed statistical inference is an interesting direction for future research.

Acknowledgements

This work was presented in part at the Allerton Conference on Control, Computing and Communication, September 2007. Work funded by NSF-grants DMS-0605165 and CCF-0545862 CAREER to MJW. The authors thank Pravin Varaiya and Alan Willsky for helpful comments.

A Proof of Lemma 1

We begin by noting that for the normalized graph Laplacian $L(G)$, it is known that for any graph, the second smallest eigenvalue satisfies the upper bound $\lambda_2(L(G)) \leq m/(m-1) \leq 1$. Moreover, we have $\text{trace}(L(G)) = m$. See Lemma 1.7 in Chung [7] for proofs of these claims.

Using these facts, we establish Lemma 1 as follows. Recall that by construction, we have $R(G) = \frac{L(G)}{\lambda_2(L(G))}$, so that the second smallest eigenvalue of $R(G)$ is $\lambda_2(R(G)) = 1$, and the remaining eigenvalues are greater than or equal to one. Applying Corollary 1 to the ANN model, we have

$$\begin{aligned}
\text{AMSE}(L;\theta^*) &= \frac{\sigma^2}{m} \sum_{i=2}^{m} \left[\frac{[\lambda_i(R(G))]^2}{2\lambda_i(R(G)) - 1}\right], \\
&= \frac{\sigma^2}{m\ \lambda_2(L(G))} \sum_{i=2}^{m} \left[\frac{[\lambda_i(L(G))]^2}{2\lambda_i(L(G)) - \lambda_2(L(G))}\right] \\
&\geq \frac{\sigma^2}{2\lambda_2(L(G))\ m}\ \text{trace}(L(G)) \\
&= \frac{\sigma^2}{2\lambda_2(L(G))}
\end{aligned}$$

using the fact that $\text{trace}(L(G)) = m$.

In the other direction, using the fact that $\lambda_2(R(G)) \geq 1$ and the bound $\frac{x^2}{2x-1} \leq x$ for $x \geq 1$. we have

$$\begin{aligned}
\text{AMSE}(L;\theta^*) &= \frac{\sigma^2}{m} \sum_{i=2}^{m} \left[\frac{[\lambda_i(R(G))]^2}{2\lambda_i(R(G)) - 1}\right], \\
&\leq \frac{\sigma^2}{m}\,\text{trace}(R(G)) \\
&= \frac{\sigma^2}{\lambda_2(L(G))\,m}\,\text{trace}(L(G)) \\
&= \frac{\sigma^2}{\lambda_2(L(G))}.
\end{aligned}$$

References

[1] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986.

[2] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley Interscience, New York, 2000.

[3] T. C. Aysal, M. Coates, and M. Rabbat. Distributed average consensus using probabilistic quantization. In *IEEE Workshop on Stat. Sig. Proc.*, Madison, WI, August 2007.

[4] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, New York, NY, 1990.

[5] V. Borkar and P. Varaiya. Asymptotic agreement in distributed estimation. *IEEE Trans. Auto. Control*, 27(3):650–655, 1982.

[6] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

[7] F.R.K. Chung. *Spectral Graph Theory*. American Mathematical Society, Providence, RI, 1991.

[8] M. H. deGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, March 1974.

[9] A. G. Dimakis, A. Sarwate, and M. J. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Trans. Signal Processing*, 53:1205–1216, March 2008.

[10] F. Fagnani and S. Zampieri. Average consensus with packet drop communication. *SIAM J. on Control and Optimization*, 2007. To appear.

[11] J. Feldman, T. Malkin, R. A. Servedio, C. Stein, and M. J. Wainwright. LP decoding corrects a constant fraction of errors. *IEEE Trans. Information Theory*, 53(1):82–89, January 2007.

[12] A. Kashyap, T. Basar, and R. Srikant. Quantized consensus. *Automatica*, 43:1192–1203, 2007.

[13] D. Kempe, A. Dobra, and J. Gehrke. Gossip-based computation of aggregate information. *Proc. 44th Ann. IEEE FOCS*, pages 482–491, 2003.

[14] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York, NY, 1997.

[15] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions in Automatic Control*, 22:551–575, 1977.

[16] R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[17] I. D. Schizas, A. Ribeiro, and G. B. Giannakis. Consensus in ad hoc WSNs with noisy links: Part I distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350–364, 2008.

[18] J. Tsitsiklis. *Problems in decentralized decision-making and computation*. PhD thesis, Department of EECS, MIT, 1984.

[19] L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 52:65–78, 2004.

[20] L. Xiao, S. Boyd, and S.-J. Kim. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, 67(1):33–46, 2007.

[21] M. E. Yildiz and A. Scaglione. Differential nested lattice encoding for consensus problems. In *Info. Proc. Sensor Networks (IPSN)*, Cambridge, MA, April 2007.