Optimization and Analysis of Distributed Averaging with Memory

Boris N. Oreshkin, Mark J. Coates and Michael G. Rabbat
Department of Electrical and Computer Engineering
McGill University, Montréal, Québec, Canada
Email: boris.oreshkin@mail.mcgill.ca, {mark.coates, michael.rabbat}@mcgill.ca

***Abstract*—This paper analyzes the rate of convergence of a distributed averaging scheme making use of memory at each node. In conventional distributed averaging, each node computes an update based on its current state and the current states of their neighbours. Previous work observed the trajectories at each node converge smoothly and demonstrated via simulation that a predictive framework can lead to faster rates of convergence. This paper provides theoretical guarantees for a distributed averaging algorithm with memory. We analyze a scheme where updates are computed as a convex combination of two terms: (i) the usual update using only current states, and (ii) a local linear predictor term that makes use of a node's current and previous states. Although this scheme only requires one additional memory register, we prove that this approach can lead to dramatic improvements in the rate of convergence. For example, on the N-node chain topology, our approach leads to a factor of N improvement over the standard approach, and on the two-dimensional grid, our approach achieves a factor of $\sqrt{N}$ improvement. Our analysis is direct and involves relating the eigenvalues of a conventional (memoryless) averaging matrix to the eigenvalues of the averaging matrix implementing the proposed scheme via a standard linearization of the quadratic eigenvalue problem. The success of our approach relies on each node using the optimal parameter for combining the two update terms. We derive a closed form expression for the optimal parameter as a function of the second largest eigenvalue of a memoryless averaging matrix, which can easily be computed in a decentralized fashion using existing methods, making our approach amenable to a practical implementation.**

I. Introduction

Average consensus has developed into a canonical problem in the distributed signal processing and control communities, due to its applications in cyber-physical and multi-agent systems. See [14] for a survey. Although distributed averaging algorithms for solving the average consensus problem have many attractive properties (e.g., robustness to changing topology and lossy links, fully decentralized, no unnecessary overhead for forming routes), conventional approaches are known to suffer from slow convergence on important network topologies such as two-dimensional grids and random geometric graphs [4], even if the algorithm parameters are optimized for the underlying topology.

In the conventional, memoryless distributed averaging algorithm analyzed by Xiao and Boyd in [20] (and which can be traced back to the seminal work of Tsitsiklis [19]), nodes exchange information with all of their neighbors at each iteration, and then update their local state with a weighted linear combination of the information just received. This update can be expressed as a simple recursion of the form $\mathbf{x}(t+1) = \mathbf{W}\mathbf{x}(t)$, where $x_i(0)$ is the initial value at node i, $x_i(t)$ is the estimate after t iterations, and the matrix $\mathbf{W}$ is the topology-respecting weight matrix summarizing the updates at each node; that is, $W_{i,j} \neq 0$ only if nodes i and j communicate directly. Xiao and Boyd [20] provide conditions on $\mathbf{W}$ which guarantee asymptotic convergence, and they show that the rate of convergence is governed by the spectral radius, $\rho(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)$. As mentioned above, this approach is slow to converge on grid-like topologies and random geometric graphs [4] – topologies which are commonly used to model connectivity in wireless networks. However, visual examination reveals that the local trajectories taken by each node converge smoothly to the consensus state (see, e.g., Fig. 4(d) in [14]). Thus, one would hope that faster convergence could by achieved by predicting the final state given this initial trajectory. Numerical simulations have previously demonstrated that such predictive consensus algorithms indeed converge much faster than memoryless, non-predictive schemes [2], [5], [9], [12], [16]. However, to date, there has been no theoretical characterization of the improvement.

A. Contributions

This paper presents the first theoretical performance guarantees for predictive consensus algorithms. We focus on linear updates of the form $\mathbf{x}(t+1) = \alpha\mathbf{x}^P(t+1) + (1-\alpha)\mathbf{x}^W(t+1)$ that mix the outcomes of a conventional neighborhood averaging, $\mathbf{x}^W(t+1) = \mathbf{W}\mathbf{x}(t)$, with a local linear predictor $\mathbf{x}^P(t+1)$ that uses only one additional memory register at each node. (The precise form of $\mathbf{x}^P(t+1)$ is described in the next section.) Thus, the updated state $x_i(t+1)$ at node i is a function of the previous states $x_j(t)$ at nodes j which are neighbors of i, and i's two previous states, $x_i(t)$ and $x_i(t-1)$. We derive a closed form expression for the mixing parameter $\alpha^\star$ that optimizes the rate of convergence, and we show that the proposed scheme leads to dramatic improvement in rate of convergence. In particular, we show that if the underlying averaging matrix has a spectral radius bounded according to $\rho(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^T) = 1 - \Psi(N)$, for a network of N nodes, then the spectral radius of our predictive scheme is bounded above by $1 - \sqrt{\Psi(N)}$. Consequently, for a two-dimensional grid, the number of iterations required to reach ϵ relative accuracy is decreased by a factor of $\sqrt{N}$ using our approach, and for a chain topology, the number of iterations is decreased by a factor of N. The optimal mixing parameter $\alpha^\star$ is a function of the underlying topology and choice of weights, $\mathbf{W}$, through the second largest eigenvalue of $\mathbf{W}$. There are existing schemes for decentralized spectral analysis (e.g., [4], [11], [15]), and so it is practical to operate the proposed scheme in a fully decentralized fashion.

B. Related Work

A number of approaches have been proposed in the literature for accelerating distributed averaging algorithms. In the context of gossip algorithms (an asynchronous form of distributed averaging), researchers have proposed exchanging information over longer distances [7] and averaging along paths [3]. Although both of these schemes lead to faster convergence rates, they also require that information be exchanged over longer distances (multiple hops). Schemes based on lifting Markov chains have also been proposed [10], [13], but construction of these schemes also requires that nodes know their locations or even more global topological information.

Two main approaches to accelerating the convergence of synchronous distributed averaging algorithms have been identified: optimizing the weight matrix $\mathbf{W}$ [4], [20], and incorporating memory into the distributed averaging algorithm [2], [5], [9], [12], [16], [17]. The spectral radius of the weight matrix governs the worst-case convergence rate in memoryless distributed averaging algorithms, so optimizing the weight matrix corresponds to minimizing the spectral radius subject to connectivity constraints. Although elegant, optimizing weights on random geometric graph topologies provides no additional gains, order-wise, over simpler, fully-decentralized weight constructions [4].

A more promising research direction is based on using local node memory. The idea of using higher-order eigenvalue shaping filters was discussed in [16], but the problem of identifying optimal filter parameters was not solved. In [5] Cao *et al.* proposed a memory-based acceleration framework for gossip algorithms where updates are a weighted sum of previous state values and gossip exchanges, but they provide no solutions or directions for weight vector design or optimization. Johansson and Johansson [9] advocate a similar scheme for distributed consensus averaging. They investigate convergence conditions and use standard solvers to find a numerical solution for the optimal weight vector. Recently, polynomial filtering was introduced for consensus acceleration, with the optimal weight vector again determined numerically [12]. Analytical solutions for the topology-dependent optimal weights have not been considered in previous work [2], [5], [9], [12] and, consequently, there has been no theoretical convergence rate analysis. Aysal et al. proposed the mixing of neighbourhood averaging with a local linear predictor in [2]. The algorithm we analyze belongs to the general framework presented therein. Although the algorithmic framework in [2] allows for multi-tap linear predictors, the analysis focuses entirely on one-tap prediction, which is equivalent to optimizing the original weight matrix without making use of the history at each node. As such, the convergence rate improvement cannot be better than that achieved by optimizing the weight matrix as in [4], [20], [21].

Sundaram and Hadjicostis [17] also investigate how memory can be used in distributed averaging and derive an algorithm that exactly achieves consensus in a finite number of iterations. Each node records the entire history of values $\{x_i(t)\}_{t=0}^{T}$ and, after enough iterations, inverts this history to recover the network average. In order to carry out the inversion, each node needs to know a topology-dependent set of weights. This leads to complicated initialization procedures; the additional memory required at each node grows with the network size. In contrast, the approach analyzed in this paper only requires a constant increase in memory size at each node, independent of the size of the network, albeit, at the price of only achieving asymptotic convergence.

C. Paper Organization

The remainder of this paper is structured as follows. Section II introduces the distributed average consensus framework and outlines the linear prediction-based acceleration methodology. Section III provides the main results, including the optimal value of the mixing parameter for the two-tap predictor and an analysis of the convergence rate improvement. We provide proofs of the main results in Section IV and Section V concludes the paper.

II. Problem Formulation

We assume that a network of N nodes is given, and that the communication topology is specified in terms of a collection of neighborhoods of each node: $\mathcal{N}_i \subseteq \{1, \ldots, N\}$ is the set of nodes with whom node i communicates directly. For $j \in \mathcal{N}_i$, we will also say that there is an edge between i and j, and assume that connectivity is symmetric; i.e., $j \in \mathcal{N}_i$ implies that $i \in \mathcal{N}_j$. We assume that the network is connected, meaning that there is a path (a sequence of adjacent edges) connecting every pair of nodes.

Initially, each node $i = 1, \ldots, N$ has a scalar value $x_i(0) \in \mathbb{R}$, and the goal is to develop a distributed algorithm such that every node computes $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i(0)$. Previous studies (see, e.g., [19] or [20]) have considered linear updates of the form $\mathbf{x}(t+1) = \mathbf{W}\mathbf{x}(t)$, where $\sum_j W_{ij} = 1$, and $W_{i,j} \neq 0$ only if $j \in \mathcal{N}_i$. For this basic setup, Xiao and Boyd [20] have shown that necessary and sufficient conditions on $\mathbf{W}$ which ensure convergence to the average consensus, $\bar{x}\mathbf{1}$, are

$$\mathbf{W}\mathbf{1} = \mathbf{1}, \quad \mathbf{1}^T\mathbf{W} = \mathbf{1}^T, \quad \rho(\mathbf{W} - \mathbf{J}) < 1, \tag{1}$$

where $\mathbf{J}$ is the averaging matrix, $\mathbf{J} = \frac{1}{N}\mathbf{1}\mathbf{1}^T$, and $\rho(\mathbf{A})$ denotes the spectral radius[1] of a matrix $\mathbf{A}$. Algorithms have been identified for locally generating weight matrices that satisfy the required convergence conditions if the underlying graph is connected, *e.g.*, Maximum–degree and Metropolis–Hastings weights [20].

Empirical evidence suggests that the convergence of the algorithm can be significantly improved by using local memory [2], [9], [12]. The idea is to exploit smooth convergence

[1] $\rho(\mathbf{A}) \triangleq \max_{i=1,\ldots,N} |\lambda_i|$, where $\{\lambda_i\}_{i=1}^{N}$ denote the eigenvalues of $\mathbf{A}$.

of the algorithm, using current and past values to predict the future trajectory. In this fashion, the algorithm achieves faster convergence by bypassing intermediate states. Each update becomes a weighted mixture of a prediction and a neighborhood averaging, but the mixture weights must be chosen carefully to ensure convergence.

The simplest case of local memory is two taps (a single tap is equivalent to storing only the current value, as in standard distributed averaging), and this is the case we consider in this paper. For two taps of memory, prediction at node i is based on the previous state value $x_i(t-1)$, the current value $x_i(t)$, and the value achieved by one application of the original averaging matrix, i.e. $x_i^{\mathrm{W}}(t+1) = W_{ii}x_i(t)+\sum_{j\in\mathcal{N}_i} W_{ij}x_j(t)$. The state-update equations at a node become a combination of the predictor and the value derived by application of the consensus weight matrix (this is easily extended for predictors with longer memories; see [2], [9]). In the two-tap memory case, we have:

$$x_i(t+1) = \alpha x_i^{\mathrm{P}}(t+1) + (1-\alpha)x_i^{\mathrm{W}}(t+1) \tag{2a}$$

$$x_i^{\mathrm{W}}(t+1) = W_{ii}x_i(t) + \sum_{j\in\mathcal{N}_i} W_{ij}x_j(t) \tag{2b}$$

$$x_i^{\mathrm{P}}(t+1) = \theta_3 x_i^{\mathrm{W}}(t+1) + \theta_2 x_i(t) + \theta_1 x_i(t-1). \tag{2c}$$

Here $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$ is the vector of predictor coefficients.

The network–wide equations can then be expressed in matrix form by defining

$$\mathbf{W}_3[\alpha] \triangleq (1-\alpha+\alpha\theta_3)\mathbf{W} + \alpha\theta_2\mathbf{I}, \tag{3}$$

$$\mathbf{X}(t) \triangleq [\mathbf{x}(t)^T, \mathbf{x}(t-1)^T]^T, \tag{4}$$

where $\mathbf{I}$ is the identity matrix of the appropriate size, and

$$\mathbf{\Phi}_3[\alpha] \triangleq \begin{bmatrix} \mathbf{W}_3[\alpha] & \alpha\theta_1\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}. \tag{5}$$

Each block of the above matrix has dimensions $N \times N$. We also define $\mathbf{x}(-1) = \mathbf{x}(0)$ so that $\mathbf{X}(0) = [\mathbf{x}(0)^T\mathbf{x}(0)^T]$. The update equation is then simply $\mathbf{X}(t+1) = \mathbf{\Phi}_3[\alpha]\mathbf{X}(t)$.

Aysal et al. describe a method for choosing the predictor coefficients $\boldsymbol{\theta}$ in [2] based on least-squares predictor design. For the two-tap memory case, the predictor coefficients are identified as $\boldsymbol{\theta}_3 = \mathbf{A}^{\dagger T}\mathbf{B}$, where

$$\mathbf{A} \triangleq \begin{bmatrix} -2 & -1 & 0 \\ 1 & 1 & 1 \end{bmatrix}^T, \tag{6}$$

$\mathbf{B} \triangleq [1,1]^T$, and $\mathbf{A}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{A}$. This choice of predictor coefficients satisfies the technical conditions on $\boldsymbol{\theta}$ required in our main results, Theorems 1 and 2, below ($\theta_1+\theta_2+\theta_3 = 1$ and $\theta_3 \geq 1, \theta_2 \geq 0$).

III. Main Results

A. Optimal Mixing Parameter

The mixing parameter α determines the influence of the standard one-step consensus iteration relative to the predictor in (2a). We take as given a foundational weight matrix, $\mathbf{W}$, which respects the underlying topology and satisfies the convergence criteria (1) and proceed to determine the optimal mixing parameter α with respect to $\mathbf{W}$. Before deriving an expression for the optimal α, it is necessary to specify what "optimal" means. Our goal is to minimize convergence time, but it is important to identify how we measure convergence time.

Xiao and Boyd [20] show that selecting weights $\mathbf{W}$ to minimize the spectral radius $\rho(\mathbf{W}-\mathbf{J})$ (while respecting the network topology constraints) leads to the optimal convergence rate for standard distributed averaging. In particular, the spectral radius $\rho(\mathbf{W}-\mathbf{J})$ is the worst-case asymptotic convergence rate. Maximizing asymptotic convergence rate is equivalent to minimizing asymptotic convergence time,

$$\tau_{\text{asym}} \triangleq \frac{1}{\log(\rho(\mathbf{W}-\mathbf{J})^{-1})}, \tag{7}$$

where, asymptotically, τ_{asym} corresponds to the number of iterations required to reduce the error $\|\mathbf{x}(t) - \bar{\mathbf{x}}\|$ by a factor of e^{-1} [20]. An alternative metric is the averaging time, the time required to achieve the prescribed level of accuracy ε while performing the distributed averaging operation:

$$T_{\text{ave}}(\mathbf{W},\varepsilon) \triangleq \sup_{\mathbf{X}(0)\neq\mathbf{0}} \inf_{t\geq 0} \big\{ t : ||\mathbf{X}(t) - \bar{\mathbf{X}}(0)||_2 \leq \varepsilon ||\mathbf{X}(0) - \bar{\mathbf{X}}(0)||_2 \big\}, \tag{8}$$

When $\mathbf{W}$ is symmetric, $\rho(\mathbf{W}-\mathbf{J})$ also defines an upper bound on the averaging time.

The update matrix we propose, (5), is not symmetric and it may not even be contracting. The results of [20] do not apply for such matrices, and the spectral radius $\rho(\mathbf{W}-\mathbf{J})$ cannot, in general, be used to specify an upper bound on averaging time. We can, however, establish a result for the *limiting* ε-averaging time, which is the averaging time for asymptotically small ε. Specifically, in Section IV-A we show that for matrices of the form (5),

$$\lim_{\varepsilon\to 0} \frac{T_{\text{ave}}(\mathbf{\Phi}_3[\alpha],\varepsilon)}{\log\varepsilon^{-1}} < \frac{1}{\log\rho(\mathbf{\Phi}_3[\alpha]-\mathbf{J})^{-1}}. \tag{9}$$

According to this result, the averaging time required to approach the average within ε-accuracy grows at the rate at most $1/\log\rho(\mathbf{\Phi}_3[\alpha]-\mathbf{J})^{-1}$ as $\varepsilon \to 0$. Minimizing the spectral radius is thus a natural optimality criterion. The following theorem establishes the optimal setting of α for a given weight matrix $\mathbf{W}$, as a function of $\lambda_2(\mathbf{W})$, the second largest eigenvalue of $\mathbf{W}$.

Theorem 1 (Optimal mixing parameter)**.** *Suppose $\theta_3+\theta_2+\theta_1 = 1$ and $\theta_3 \geq 1$, $\theta_2 \geq 0$. Suppose further that $|\lambda_N(\mathbf{W})| \leq |\lambda_2(\mathbf{W})|$, where the eigenvalues $\lambda_1(\mathbf{W}),\ldots,\lambda_N(\mathbf{W})$ are labelled in decreasing order. Then the solution of the optimization problem*

$$\alpha^\star = \arg\min_\alpha \rho(\mathbf{\Phi}_3[\alpha]-\mathbf{J}) \tag{10}$$

is given by the following:

$$\alpha^\star = \frac{-((\theta_3-1)\lambda_2(\mathbf{W})^2 + \theta_2\lambda_2(\mathbf{W}) + 2\theta_1)}{(\theta_2+(\theta_3-1)\lambda_2(\mathbf{W}))^2} - \frac{2\sqrt{\theta_1^2+\theta_1\lambda_2(\mathbf{W})\left(\theta_2+(\theta_3-1)\lambda_2(\mathbf{W})\right)}}{(\theta_2+(\theta_3-1)\lambda_2(\mathbf{W}))^2} \tag{11}$$

Proofs of all results are deferred to Section IV below. A brief discussion of the conditions of this theorem is warranted. The conditions on the predictor weights are technical conditions that ensure convergence is achieved. Three factors motivate our belief that these are not overly-restricting: (i) these conditions are satisfied if we employ the least-squares predictor weights design strategy of [2]; (ii) the conditions are relatively natural for a linear predictor that is based on an estimate of slope; (iii) in Section III-B we show that the choice of weights does not have a significant effect on the convergence properties.

The condition on the weight matrix, $|\lambda_N(\mathbf{W})| \leq |\lambda_2(\mathbf{W})|$, significantly reduces the complexity of the proof. Most distributed algorithms for constructing weight matrices (e.g., Metropolis-Hastings (MH) or max-degree) lead to $\mathbf{W}$ that satisfy the condition, but they are not guaranteed to do so. We can ensure that the condition is satisfied by applying a completely local adjustment to any weight matrix. The mapping $\mathbf{W} \mapsto 1/2(\mathbf{I}+\mathbf{W})$ transforms any stochastic matrix $\mathbf{W}$ into a stochastic matrix with all positive eigenvalues [4]; this mapping can be carried out locally, without any knowledge of the global properties of $\mathbf{W}$, and without affecting the order-wise asymptotic convergence rate as $N \to \infty$.

B. Convergence Rate Analysis

We begin with our main result for the convergence rate of two-tap predictor-based accelerated consensus. Theorem 2 indicates how the spectral radius of the accelerated operator $\mathbf{\Phi}_3[\alpha]$ is related to the spectral radius of the foundational weight matrix $\mathbf{W}$ (in terms of upper bounds on these quantities). Since the (limiting) asymptotic convergence time is governed by the spectral radius, this relationship characterizes the improvement in convergence rate that can be obtained.

Theorem 2 (Convergence rate)**.** *Suppose the assumptions of Theorem 1 hold. Suppose further that the original matrix* $\mathbf{W}$ *satisfies* $\rho(\mathbf{W}-\mathbf{J}) \leq 1-\Psi(N)$ *for some function* $\Psi : \mathbb{N} \to (0,1)$ *of the network size* N *decreasing to* 0*. Then the matrix* $\mathbf{\Phi}_3[\alpha^\star]$ *satisfies* $\rho(\mathbf{\Phi}_3[\alpha^\star]-\mathbf{J}) \leq 1-\sqrt{\Psi(N)}$.

In order to explore how fast the spectral radius, $\rho(\mathbf{\Phi}_3[\alpha^\star]-\mathbf{J}) = \sqrt{-\alpha^\star\theta_1}$, (see Section IV-C for details) goes to one as $N \to \infty$, we can take its asymptotic Taylor series expansion:

$$\rho(\mathbf{\Phi}_3[\alpha^*]-\mathbf{J}) = 1 - \sqrt{\frac{2(\theta_3-1)+\theta_2}{\theta_3-1+\theta_2}}\sqrt{\Psi(N)} + \mathcal{O}(\Psi(N)). \tag{12}$$

From this expression, we see that the bound presented in Theorem 2 correctly captures the convergence rate of the accelerated consensus algorithm. Alternatively, leaving only two terms in the expansion above, $\rho(\mathbf{\Phi}_3[\alpha^*]-\mathbf{J}) = 1 - \Omega(\sqrt{\Psi(N)})$, we see that the bound presented is rate optimal in Landau notation.

We can also use (12) to provide guidelines for choosing asymptotically optimal prediction parameters θ_3 and θ_2. In particular, it is clear that the coefficient $\gamma(\theta_2,\theta_3) = \sqrt{[2(\theta_3-1)+\theta_2]/[\theta_3-1+\theta_2]}$ should be maximized to minimize the spectral radius $\rho(\mathbf{\Phi}_3[\alpha^\star]-\mathbf{J})$. It is straightforward to verify that setting $\theta_2 = 0$ and $\theta_3 = 1+\epsilon$ for any $\epsilon > 0$ satisfies the assumptions of Theorem 1 and also satisfies $\gamma(0,1+\epsilon) > \gamma(\theta_2,1+\epsilon)$ for any positive θ_2. Since $\gamma(0,1+\epsilon) = \sqrt{2}$ is independent of ϵ (or θ_3) we conclude that setting $(\theta_1,\theta_2,\theta_3) = (-\epsilon,0,1+\epsilon)$ satisfies the assumptions of Theorem 1 and asymptotically yields the optimal limiting ε-averaging time for the proposed approach, as $N \to \infty$.

For a chain graph (path of N vertices) the eigenvalues of the normalized graph Laplacian $\mathcal{L}$ are given by $\lambda_i(\mathcal{L}) = 1-\cos(\pi i/(N-1)), i = 0,1,\ldots,N-1$ [6]. It is straightforward to verify that for the Metropolis-Hastings (MH) weight matrix a similar expression holds: $\lambda_i(\mathbf{W}_{\text{MH}}) = 1/3 + 2/3\cos(\pi(i-1)/N), i = 1,2,\ldots,N$. Thus, in this case, $\rho(\mathbf{W}_{\text{MH}}-\mathbf{J}) = 1/3+2/3\cos(\pi/N)$. For sufficiently large N, this results in $\rho(\mathbf{W}_{\text{MH}}-\mathbf{J}) = 1-\frac{\pi^2}{3}\frac{1}{N^2}+\mathcal{O}(1/N^4)$, and thus we find that $\rho(\mathbf{\Phi}_3[\alpha^\star]-\mathbf{J}) = 1-\mathcal{O}(1/N)$. In terms of the number of the limiting ε-averaging time described above, this corresponds to a factor of N decrease in the number of iterations required to achieve a relative error less than ε. Similarly, for a network with two-dimensional grid topology, taking $\mathbf{W}$ to be the transition matrix for a natural random walk on the grid (a minor perturbation of the MH weights) it is known [1] that $(1-\lambda_2(\mathbf{W}))^{-1} = \Theta(N)$. Thus, for a two-dimensional grid, the proposed algorithm leads to $\rho(\mathbf{\Phi}_3[\alpha^\star]-\mathbf{J}) = 1-\mathcal{O}(N^{-1/2})$, or an improvement by a factor of $N^{1/2}$ iterations.

IV. Proofs of Main Results and Discussion

A. Limiting ε-averaging Time

To begin, we need to motivate choosing α to minimize the spectral radius $\rho(\mathbf{\Phi}_3[\alpha]-\mathbf{J})$ since, unlike in the memoryless setting, it does not bound the step-wise rate of convergence. In fact, since $\mathbf{\Phi}_3[\alpha]$ is not symmetric, $\mathbf{\Phi}_3[\alpha]^t$ does not even converge to $\mathbf{J}$ as $t \to \infty$, as in the memoryless setting. However, we will show that: (i) for the proposed construction, $\mathbf{\Phi}_3[\alpha]^t$ does converge to a matrix limit $\bar{\mathbf{\Phi}}$; (ii) that the limiting averaging time is governed by $\rho(\mathbf{\Phi}_3[\alpha]-\bar{\mathbf{\Phi}})$; and (iii) that $\rho(\mathbf{\Phi}_3[\alpha]-\bar{\mathbf{\Phi}}) = \rho(\mathbf{\Phi}_3[\alpha]-\mathbf{J})$.

Before stating our first result we must introduce some notation. For now, assume we are given a matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$ with $\bar{\mathbf{\Phi}} = \lim_{t\to\infty}\mathbf{\Phi}^t$. We will address conditions for existence of the limit below. For a given initialization vector $\mathbf{X}(0) \in \mathbb{R}^n$, let $\tilde{\mathbf{X}}(0) = \bar{\mathbf{\Phi}}\mathbf{X}(0)$, and define the set of non-trivial initialization vectors $\mathcal{X}_{0,\mathbf{\Phi}} \triangleq \{\mathbf{X}(0) \in \mathbb{R}^n : \mathbf{X}(0) \neq \tilde{\mathbf{X}}(0)\}$. Since we have not yet established that $\tilde{\mathbf{X}}(0) = \bar{\mathbf{X}}(0) \triangleq \mathbf{J}\mathbf{X}(0)$, we keep the discussion general and use the following definition of the averaging time:

$$T_{\text{ave}}(\mathbf{\Phi},\varepsilon) \triangleq \sup_{\mathbf{X}(0)\in\mathcal{X}_{0,\mathbf{\Phi}}} \inf_{t\geq 0} \Big\{ t : ||\mathbf{X}(t)-\tilde{\mathbf{X}}(0)||_2 \leq \varepsilon||\mathbf{X}(0)-\tilde{\mathbf{X}}(0)||_2 \Big\}. \tag{13}$$

In order to obtain the desired result linking $T_{\text{ave}}(\mathbf{\Phi},\varepsilon)$ and the spectral radius $\rho(\mathbf{\Phi}_3[\alpha]-\mathbf{J})$ of the proposed algorithm, we will make use of the following more general result about limiting rates for convergent matrices.

Theorem 3. *Let $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$ be given, with limit $\lim_{t\to\infty} \boldsymbol{\Phi}^t = \bar{\boldsymbol{\Phi}}$, and assume that $\rho(\boldsymbol{\Phi} - \bar{\boldsymbol{\Phi}}) > 0$. Then*

$$\lim_{\varepsilon\to 0} \frac{T_{ave}(\boldsymbol{\Phi}, \varepsilon)}{\log \varepsilon^{-1}} < \frac{1}{\log \rho(\boldsymbol{\Phi} - \bar{\boldsymbol{\Phi}})^{-1}}. \tag{14}$$

We omit the full proof here due to lack of space. The basic idea is to relate the definition of $T_{\text{ave}}(\boldsymbol{\Phi}, \varepsilon)$ above in terms of the quantity $\|(\boldsymbol{\Phi} - \bar{\boldsymbol{\Phi}})^t\|^{1/t}_{\mathbf{X}(0)}$, where $||\boldsymbol{\Phi}||_{\mathbf{X}(0)} = ||\boldsymbol{\Phi}(\mathbf{X}(0) - \tilde{\mathbf{X}}(0))||_2/||\mathbf{X}(0) - \tilde{\mathbf{X}}(0)||_2$, and then to apply Gelfand's formula [8] to obtain the final result, linking $T_{\text{ave}}(\boldsymbol{\Phi}, \varepsilon)$ and $\rho(\boldsymbol{\Phi} - \bar{\boldsymbol{\Phi}})$. Accomplishing these steps requires a number of technicalities; see [15] for the full details.

In order to apply the above result to the proposed algorithm, we must establish that $\boldsymbol{\Phi}_3[\alpha]$ satisfies the conditions of Theorem 3. In doing so, we will also show that (i) for $\boldsymbol{\Phi} = \boldsymbol{\Phi}_3[\alpha]$, the limit $\bar{\boldsymbol{\Phi}}\mathbf{X}(0) = \mathbf{J}\mathbf{X}(0)$, so our approach indeed converges to the average consensus, and (ii) that the limiting averaging time is characterized by a function of $\rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$, which motivates choosing α to optimize this expression. (Recall, in this setting $\mathbf{J}$ is the $2N \times 2N$ matrix with all entries equal to $1/2N$.) Note that in the following proposition, the conditions on $\boldsymbol{\theta}$ are the same as in Theorem 1 (and were discussed in Section III), and the condition on α is necessary for $\boldsymbol{\Phi}_3[\alpha]^t$ to have a limit as $t \to \infty$, as will be established in Section IV-B.

Proposition 1. *Let $\boldsymbol{\Phi}_3[\alpha]$ be defined as in (5) and assume that $\theta_1 + \theta_2 + \theta_3 = 1$, $\theta_3 \geq 1$, $\theta_2 \geq 0$, and $\alpha \in [0, -\theta_1^{-1})$. Then:*

(a) $\bar{\boldsymbol{\Phi}}_3[\alpha] = \lim_{t\to\infty} \boldsymbol{\Phi}_3[\alpha]^t$ *exists, with* $\bar{\boldsymbol{\Phi}}_3[\alpha]\mathbf{X}(0) = \mathbf{J}\mathbf{X}(0)$ *for all* $\mathbf{X}(0)$ *defined in* (4),

(b) $\rho(\boldsymbol{\Phi}_3[\alpha] - \bar{\boldsymbol{\Phi}}_3[\alpha]) > 0$, *and*

(c) $\lim\limits_{\varepsilon\to 0} \frac{T_{ave}(\boldsymbol{\Phi}_3[\alpha],\varepsilon)}{\log \varepsilon^{-1}} < \frac{1}{\log \rho(\boldsymbol{\Phi}_3[\alpha]-\mathbf{J})^{-1}}$.

Proof: Proof of part (a). In Theorem 1 in [9], Johansson and Johansson show that the necessary and sufficient conditions for the consensus algorithm of the form $\boldsymbol{\Phi}_3[\alpha]$ to converge to the average are (JJ1) $\boldsymbol{\Phi}_3[\alpha]\mathbf{1} = \mathbf{1}$; (JJ2) $\mathbf{g}^T\boldsymbol{\Phi}_3[\alpha] = \mathbf{g}^T$ for vector $\mathbf{g}^T = [\beta_1 \mathbf{1}^T \beta_2 \mathbf{1}^T]$ with weights satisfying $\beta_1 + \beta_2 = 1$; and (JJ3) $\rho(\boldsymbol{\Phi}_3[\alpha] - \frac{1}{N}\mathbf{1}\mathbf{g}^T) < 1$. If these conditions hold then we also have $\bar{\boldsymbol{\Phi}}_3[\alpha] = \frac{1}{N}\mathbf{1}\mathbf{g}^T$ [9] implying $\tilde{\mathbf{X}}(0) = \bar{\mathbf{X}}(0)$. Condition (JJ1) is easily verified after straightforward algebraic manipulations using the definition of $\boldsymbol{\Phi}_3[\alpha]$ in (5), the assumption that $\theta_1 + \theta_2 + \theta_3 = 1$, and recalling that $\mathbf{W}$ satisfies $\mathbf{W}\mathbf{1} = \mathbf{1}$ by design. To address condition (JJ2), we set $\beta_1 = 1/(1 + \alpha\theta_1)$ and $\beta_2 = \alpha\theta_1/(1 + \alpha\theta_1)$. Clearly, $\beta_1 + \beta_2 = 1$, and it is also easy to verify condition (JJ2) by plugging these values into the definition of $\mathbf{g}$, and using the same properties of $\boldsymbol{\Phi}_3[\alpha]$, the θ_i's, and $\mathbf{W}$ as above.

In order to verify that condition (JJ3) holds, we will show here that $\rho(\boldsymbol{\Phi}_3[\alpha] - \frac{1}{N}\mathbf{1}\mathbf{g}^T) = \rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$. In Section IV-B we show that $\rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J}) < 1$ if $\alpha \in [0, -\theta_1^{-1})$, and thus condition (JJ3) is also satisfied under the assumptions of the proposition. To show that $\rho(\boldsymbol{\Phi}_3[\alpha] - \frac{1}{N}\mathbf{1}\mathbf{g}^T) = \rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$, we prove a stronger result, namely that $\boldsymbol{\Phi}_3[\alpha] - \frac{1}{N}\mathbf{1}\mathbf{g}^T$ and $\boldsymbol{\Phi}_3[\alpha] - \mathbf{J}$ have the same eigenspectra. Consider the eigenvector $\mathbf{v}_i$ of $\boldsymbol{\Phi}_3[\alpha]$ with corresponding eigenvalue $\lambda_i(\boldsymbol{\Phi}_3[\alpha])$. This pair solves the eigenvalue problem, $\boldsymbol{\Phi}_3[\alpha]\mathbf{v}_i = \lambda_i(\boldsymbol{\Phi}_3[\alpha])\mathbf{v}_i$. Equivalently, expanding the definition of $\boldsymbol{\Phi}_3[\alpha]$, we have

$$\begin{bmatrix} \mathbf{W}_3[\alpha] & \alpha\theta_1\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \mathbf{v}_i = \lambda_i(\boldsymbol{\Phi}_3[\alpha]) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{v}_i. \tag{15}$$

We observe that (15) fits a modification of the first companion form of the linearization of a Quadratic Eigenvalue Problem (QEP) (see Section 3.4 in [18]). The QEP has general form $(\lambda^2\mathbf{M}+\lambda\mathbf{C}+\mathbf{K})\mathbf{u} = \mathbf{0}$, where $\mathbf{u}$ is the eigenvector associated with this QEP. The linearization of interest to us has the form:

$$\begin{bmatrix} -\mathbf{C} & -\mathbf{K} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \lambda\mathbf{u} \\ \mathbf{u} \end{bmatrix} - \lambda \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \lambda\mathbf{u} \\ \mathbf{u} \end{bmatrix} = \mathbf{0}. \tag{16}$$

The correspondence is clear if we make the associations: $\mathbf{M} = \mathbf{I}$, $\mathbf{C} = -\mathbf{W}_3[\alpha]$ and $\mathbf{K} = -\alpha\theta_1\mathbf{I}$, $\lambda = \lambda_i(\boldsymbol{\Phi}_3[\alpha])$ and $\mathbf{v}_i = [\lambda_i(\boldsymbol{\Phi}_3[\alpha])\mathbf{u}^T\mathbf{u}^T]^T$. Eigenvectors $\mathbf{v}_i$ that solve (15) thus have special structure and are related to $\mathbf{u}_i$, the solution to the QEP,

$$(\lambda_i(\boldsymbol{\Phi}_3[\alpha])^2\mathbf{I} - \lambda_i(\boldsymbol{\Phi}_3[\alpha])\mathbf{W}_3[\alpha] - \alpha\theta_1\mathbf{I})\mathbf{u}_i = \mathbf{0}. \tag{17}$$

Because the first and third terms above are scaled identity matrices and the definition of $\mathbf{W}_3[\alpha]$ (see (3)) also involves scaled identity matrices, we can simplify this last equation to find that any solution $\mathbf{u}_i$ must also be an eigenvector of $\mathbf{W}$.

We have seen above, when verifying condition (JJ1), that $\mathbf{1}$ is an eigenvector of $\boldsymbol{\Phi}_3[\alpha]$ with corresponding eigenvalue $\lambda_i(\boldsymbol{\Phi}_3[\alpha]) = 1$. Likewise, we know that[2] $\mathbf{W}\mathbf{1} = \mathbf{1}$, and so this agrees with the structure of $\mathbf{v}_i$ identified above. Observe that, from the definition of $\mathbf{g}$ and because $\beta_1 + \beta_2 = 1$, we have $(\frac{1}{N}\mathbf{1}\mathbf{g}^T)\mathbf{1} = \mathbf{1}$. Thus, $(\boldsymbol{\Phi}_3[\alpha] - \frac{1}{N}\mathbf{1}\mathbf{g}^T)\mathbf{1} = \mathbf{0}$. Similarly, recalling that $\mathbf{J} = \frac{1}{2N}\mathbf{1}\mathbf{1}^T$, we have $\mathbf{J}\mathbf{1} = \mathbf{1}$, and thus $(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})\mathbf{1} = \mathbf{0}$. By design, $\mathbf{W}$ is a doubly stochastic matrix, and all eigenvectors $\mathbf{u}$ of $\mathbf{W}$ with $\mathbf{u} \neq \mathbf{1}$ are orthogonal to $\mathbf{1}$. It follows that $(\frac{1}{N}\mathbf{1}\mathbf{g}^T)\mathbf{v}_i = \mathbf{0}$ for corresponding eigenvectors $\mathbf{v}_i = [\lambda_i(\boldsymbol{\Phi}_3[\alpha])\mathbf{u}^T\mathbf{u}^T]^T$ of $\boldsymbol{\Phi}_3[\alpha]$, and thus $(\boldsymbol{\Phi}_3[\alpha] - \frac{1}{N}\mathbf{1}\mathbf{g}^T)\mathbf{v}_i = \boldsymbol{\Phi}_3[\alpha]\mathbf{v}_i = \lambda_i(\boldsymbol{\Phi}_3[\alpha])\mathbf{v}_i$. Similarly, $\mathbf{J}\mathbf{v}_i = 0$ if $\mathbf{v}_i \neq \mathbf{1}$, and $(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})\mathbf{v}_i = \lambda_i(\boldsymbol{\Phi}_3[\alpha])\mathbf{v}_i$. Therefore, we conclude that the matrices $(\boldsymbol{\Phi}_3[\alpha] - \bar{\boldsymbol{\Phi}}_3[\alpha])$ and $(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$ have identical eigenspectra, and thus $\rho(\boldsymbol{\Phi}_3[\alpha] - \frac{1}{N}\mathbf{1}\mathbf{g}^T) = \rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$.

In Section IV-B we show that $\rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J}) < 1$ if $\alpha \in [0, -\theta_1^{-1})$, and thus the assumptions of the proposition, taken together with the analysis just conducted, verify that condition (JJ3) is also satisfied. Therefore, the limit $\lim_{t\to\infty} \boldsymbol{\Phi}_3[\alpha]^t = \bar{\boldsymbol{\Phi}}_3[\alpha] = \frac{1}{N}\mathbf{1}\mathbf{g}^T$ exists, and $\bar{\boldsymbol{\Phi}}_3[\alpha]\mathbf{X}(0) = \mathbf{J}\mathbf{X}(0)$ for all $\mathbf{X}(0)$ defined in (4).

Proofs of parts (b) and (c). In the proof of Lemma 1 (see Section IV-B), it is shown that $\rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J}]) \geq -\alpha\theta_1$. Thus, if $\alpha > 0$ and $\theta_1 < 0$, then part (b) holds. The assumptions $\theta_1 + \theta_2 + \theta_3 = 1$, $\theta_3 \geq 1$, and $\theta_2 \geq 0$ imply that $\theta_1 \leq 0$, and by assumption, $\alpha \geq 0$. If $\alpha = 0$ or $\theta_1 = 0$, then the

[2] We abuse notation here, using $\mathbf{1}$ to denote the vector of all 1's, where the dimension is not explicitly indicated but should be clear from the context.

proposed predictive consensus scheme reduces to memoryless consensus with weight matrix $\mathbf{W}$ (and the statement follows directly from the results of [4], [20]). Thus, part (b) of the proposition follows from the assumptions and the analysis in Lemma 1 below. By proving parts (a) and (b), we have verified the assumptions of Theorem 3 above. Applying the result of this Theorem, together with the equivalence of $\rho(\boldsymbol{\Phi}_3[\alpha] - \frac{1}{N}\mathbf{1}\mathbf{g}^T)$ and $\rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$, gives the claim in part (c), thereby completing the proof. ■

B. Proof of Theorem 1: Optimal Mixing Parameter

In order to minimize the spectral radius of $\boldsymbol{\Phi}_3[\alpha]$ we need to know its eigenvalues. These can be calculated by solving the eigenvalue problem (15). We can multiply (17) by $\mathbf{u}_i^T$ on the left to obtain a quadratic equation that links the individual eigenvalues $\lambda_i(\boldsymbol{\Phi}_3[\alpha])$ and $\lambda_i(\mathbf{W}_3[\alpha])$:

$$\mathbf{u}_i^T(\lambda_i(\boldsymbol{\Phi}_3[\alpha])^2\mathbf{I} - \lambda_i(\boldsymbol{\Phi}_3[\alpha])\mathbf{W}_3[\alpha] - \alpha\theta_1\mathbf{I})\mathbf{u}_i = 0$$

$$\lambda_i(\boldsymbol{\Phi}_3[\alpha])^2 - \lambda_i(\mathbf{W}_3[\alpha])\lambda_i(\boldsymbol{\Phi}_3[\alpha]) - \alpha\theta_1 = 0. \quad (18)$$

Recall $\boldsymbol{\Phi}_3[\alpha]$ is a $2N \times 2N$ matrix, and so $\boldsymbol{\Phi}_3[\alpha]$ has, in general, $2N$ eigenvalues – twice as many as $\mathbf{W}_3[\alpha]$. These eigenvalues are the solutions of the quadratic (18), and are given by

$$\lambda_i^*(\boldsymbol{\Phi}_3[\alpha]) = \frac{1}{2}\left(\lambda_i(\mathbf{W}_3[\alpha]) + \sqrt{\lambda_i(\mathbf{W}_3[\alpha])^2 + 4\alpha\theta_1}\right) \quad (19)$$

$$\lambda_i^{**}(\boldsymbol{\Phi}_3[\alpha]) = \frac{1}{2}\left(\lambda_i(\mathbf{W}_3[\alpha]) - \sqrt{\lambda_i(\mathbf{W}_3[\alpha])^2 + 4\alpha\theta_1}\right).$$

With these expressions for the eigenvalues of $\boldsymbol{\Phi}_3[\alpha]$, we are in a position to formulate the problem of minimizing the spectral radius of the matrix $(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$, $\alpha^\star = \arg\min_{\alpha} \rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$. It can be shown that this problem is equivalent to

$$\alpha^\star = \arg\min_{\alpha \geq 0} \rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J}) \quad (20)$$

The simplest way to demonstrate this is to show that $\rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J}) \geq \rho(\boldsymbol{\Phi}_3[0] - \mathbf{J})$ for any $\alpha < 0$. Indeed, by the definition of the spectral radius we have that $\rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J}) \geq \lambda_2^*(\boldsymbol{\Phi}_3[\alpha])$ and $\rho(\boldsymbol{\Phi}_3[0] - \mathbf{J}) = \lambda_2(\mathbf{W})$ since $\boldsymbol{\Phi}_3[0] = \mathbf{W}$. Hence it is enough to demonstrate $\lambda_2^*(\boldsymbol{\Phi}_3[\alpha]) \geq \lambda_2(\mathbf{W})$. Consider the inequality $\lambda_2^*(\boldsymbol{\Phi}_3[\alpha]) - \lambda_2(\mathbf{W}) \geq 0$. Replacying $\lambda_i^*(\boldsymbol{\Phi}_3[\alpha])$ with its definition, (19), rearranging terms and squaring both sides gives $\alpha\theta_1 \geq \lambda_2(\mathbf{W})^2 - \lambda_2(\mathbf{W})\lambda_2(\mathbf{W}_3[\alpha])$. From the definition of $\mathbf{W}_3[\alpha]$ in (3), it follows that $\lambda_2(\mathbf{W}_3[\alpha]) = (1-\alpha+\alpha\theta_3)\lambda_2(\mathbf{W}) + \alpha\theta_1$. Using this relation leads to the expression $\alpha(\theta_1 + (\theta_3 - \alpha)\lambda_2(\mathbf{W})^2 + \theta_1\lambda_2(\mathbf{W})) \geq 0$. Under our assumptions, we have $\theta_3 - 1 \geq 1$, $\theta_2 \geq 0$ and $\theta_1 \leq 0$. Thus $\theta_1 + (\theta_3 - 1)\lambda_2^2 + \theta_2\lambda_2 \leq \theta_1 + \theta_3 - 1 + \theta_2 = 0$. This implies that if $\alpha < 0$, the last inequality holds leading to $\lambda_2^*(\boldsymbol{\Phi}_3[\alpha]) \geq \lambda_2(\mathbf{W})$. Thus for any $\alpha < 0$ the spectral radius $\rho(\boldsymbol{\Phi}_3[\alpha] - \mathbf{J})$ cannot decrease, and so we may focus on optimizing over $\alpha \geq 0$.

Now, the proof of Theorem 1 boils down to examining how varying α affects the eigenvalues of $\boldsymbol{\Phi}_3[\alpha]$ on a case-by-case basis. We show that the second eigenvalues, $\lambda_2^*(\boldsymbol{\Phi}_3[\alpha])$ and $\lambda_2^{**}(\boldsymbol{\Phi}_3[\alpha])$, dominate all other pairs, $\lambda_j^*(\boldsymbol{\Phi}_3[\alpha])$ and $\lambda_j^{**}(\boldsymbol{\Phi}_3[\alpha])$, for $j = 1$ and $j > 2$, allowing us to focus on the second eigenvalues, from which the proof follows. Along the way, we establish conditions on α which guarantee stability of the proposed predictive consensus methodology.

To begin, we reformulate the optimization problem in terms of the eigenvalues of $\boldsymbol{\Phi}_3[\alpha]$. We first consider $\lambda_1^*(\boldsymbol{\Phi}_3[\alpha])$ and $\lambda_1^{**}(\boldsymbol{\Phi}_3[\alpha])$. Substituting $\lambda_1(\mathbf{W}_3[\alpha]) = (1 - \alpha + \alpha\theta_3) + \alpha\theta_2$ we obtain the relationship $\sqrt{\lambda_1^2(\mathbf{W}_3[\alpha]) + 4\alpha\theta_1} = |1 + \alpha\theta_1|$ and using the condition $\theta_1 \leq 0$, we conclude that

$$\lambda_1^*(\boldsymbol{\Phi}_3[\alpha]), \lambda_1^{**}(\boldsymbol{\Phi}_3[\alpha])$$
$$= \begin{cases} 1, -\alpha\theta_1 & \text{if } \; 1 + \alpha\theta_1 \geq 0 \Rightarrow \alpha \leq -\theta_1^{-1} \\ -\alpha\theta_1, 1 & \text{if } \; 1 + \alpha\theta_1 < 0 \Rightarrow \alpha > -\theta_1^{-1}. \end{cases} \quad (21)$$

We note that $\alpha > -\theta_1^{-1}$ implies $|\lambda_1^{**}(\boldsymbol{\Phi}_3[\alpha])| > 1$, leading to divergence of the linear recursion involving $\boldsymbol{\Phi}_3[\alpha]$, and thus conclude that the potential solution is restricted to the range $\alpha \leq -\theta_1^{-1}$. Focusing on this setting, we write $\lambda_1^*(\boldsymbol{\Phi}_3[\alpha]) = 1$ and $\lambda_1^{**}(\boldsymbol{\Phi}_3[\alpha]) = -\alpha\theta_1$. We can now reformulate the problem (20) in terms of the eigenvalues of $\boldsymbol{\Phi}_3[\alpha]$:

$$\alpha^\star = \arg\min_{\alpha \geq 0} \max_{i=1,2,\ldots N} \mathcal{J}_i[\alpha, \lambda_i(\mathbf{W})] \quad (22)$$

where

$$\mathcal{J}_i[\alpha, \lambda_i(\mathbf{W})] = \begin{cases} |\lambda_1^{**}(\boldsymbol{\Phi}_3[\alpha])|, & i = 1 \\ \max(|\lambda_i^*(\boldsymbol{\Phi}_3[\alpha])|, |\lambda_i^{**}(\boldsymbol{\Phi}_3[\alpha])|) & i > 1. \end{cases} \quad (23)$$

We now state a lemma that characterizes the functions $\mathcal{J}_i[\alpha, \lambda_i(\mathbf{W})]$.

Lemma 1. *Under the assumptions of Theorem 1,*

$$\mathcal{J}_i[\alpha, \lambda_i(\mathbf{W})]$$
$$= \begin{cases} \alpha^{1/2}(-\theta_1)^{1/2} & \textit{if } \; \alpha \in [\alpha_i^*, \theta_1^{-1}] \\ \frac{1}{2}\left(|\lambda_i(\mathbf{W}_3[\alpha])| + \sqrt{\lambda_i(\mathbf{W}_3[\alpha])^2 + 4\alpha\theta_1}\right) & \textit{if } \; \alpha \in [0, \alpha_i^*) \end{cases} \quad (24)$$

where

$$\alpha_i^* = \frac{-((\theta_3 - 1)\lambda_i(\mathbf{W})^2 + \theta_2\lambda_i(\mathbf{W}) + 2\theta_1)}{(\theta_2 + (\theta_3 - 1)\lambda_i(\mathbf{W}))^2} - \frac{2\sqrt{\theta_1^2 + \theta_1\lambda_i(\mathbf{W})\,(\theta_2 + (\theta_3 - 1)\lambda_i(\mathbf{W}))}}{(\theta_2 + (\theta_3 - 1)\lambda_i(\mathbf{W}))^2} \quad (25)$$

Proof: For $i = 2, 3, \ldots N$, the eigenvalues $\lambda_i^*(\boldsymbol{\Phi}_3[\alpha])$ and $\lambda_i^{**}(\boldsymbol{\Phi}_3[\alpha])$ can admit two distinct forms; when the expression under the square root in (19) is less then zero, the respective eigenvalues are complex, and when this expression is positive, the eigenvalues are real. In the region where the eigenvalues are complex,

$$\max(|\lambda_i^*(\boldsymbol{\Phi}_3[\alpha])|, |\lambda_i^{**}(\boldsymbol{\Phi}_3[\alpha])|)$$
$$= \frac{1}{2}\left[\lambda_i(\mathbf{W}_3[\alpha])^2 + \imath^2\left(\sqrt{\lambda_i(\mathbf{W}_3[\alpha])^2 + 4\alpha\theta_1}\right)^2\right]^{1/2}$$
$$= \alpha^{1/2}(-\theta_1)^{1/2}. \quad (26)$$

We note that (26) is a strictly increasing function of α. Recalling that $\lambda_i(\mathbf{W}_3[\alpha]) = (1 + \alpha(\theta_3 - 1))\lambda_i(\mathbf{W}) + \alpha\theta_2$

and solving the quadratic $\lambda_i(\mathbf{W}_3[\alpha])^2 + 4\alpha\theta_1 = 0$, we can identify region, $[\alpha_i^*, \alpha_i^{**}]$, where the eigenvalues are complex. The upper boundary of this region is

$$\alpha_i^{**} = \frac{-((\theta_3-1)\lambda_i(\mathbf{W})^2 + \theta_2\lambda_i(\mathbf{W}) + 2\theta_1)}{(\theta_2 + (\theta_3-1)\lambda_i(\mathbf{W}))^2} + \frac{2\sqrt{\theta_1^2 + \theta_1\lambda_i(\mathbf{W})\,(\theta_2 + (\theta_3-1)\lambda_i(\mathbf{W}))}}{(\theta_2 + (\theta_3-1)\lambda_i(\mathbf{W}))^2} \tag{27}$$

Relatively straightforward algebraic manipulation of (25) and (27) leads to the following conclusion: if $\lambda_i(\mathbf{W}) \in [-1, 1]$, $\theta_2 \geq 0$ and $\theta_3 \geq 1$, then $0 \leq \alpha_i^* \leq -\theta_1^{-1} \leq \alpha_i^{**}$. This implies that (26) holds in the region $[\alpha_i^*, -\theta_1^{-1}]$.

On the interval $\alpha \in [0, \alpha_i^*)$, the expression under the square root in (19) is positive, and the corresponding eigenvalues are real. Thus,

$$\max(|\lambda_i^*(\mathbf{\Phi}_3[\alpha])|, |\lambda_i^{**}(\mathbf{\Phi}_3[\alpha])|) = \frac{1}{2}\begin{cases} \left|\lambda_i(\mathbf{W}_3[\alpha]) + \sqrt{\lambda_i(\mathbf{W}_3[\alpha])^2 + 4\alpha\theta_1}\right| & \text{if } \lambda_i(\mathbf{W}_3[\alpha]) \geq 0 \\ \left|-\lambda_i(\mathbf{W}_3[\alpha]) + \sqrt{\lambda_i(\mathbf{W}_3[\alpha])^2 + 4\alpha\theta_1}\right| & \text{if } \lambda_i(\mathbf{W}_3[\alpha]) < 0, \end{cases} \tag{28}$$

or equivalently, $\max(|\lambda_i^*(\mathbf{\Phi}_3[\alpha])|, |\lambda_i^{**}(\mathbf{\Phi}_3[\alpha])|) = \frac{1}{2}\left(|\lambda_i(\mathbf{W}_3[\alpha])| + \sqrt{\lambda_i(\mathbf{W}_3[\alpha])^2 + 4\alpha\theta_1}\right)$. These results establish the expression for $\mathcal{J}_i[\alpha, \lambda_i(\mathbf{W})]$ in the lemma. ■

The previous lemma provided a characterization of $\mathcal{J}_i[\alpha, \lambda_i(\mathbf{W})]$. The following lemma establishes that we can simplify the optimization and focus solely on $\mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})]$.

Lemma 2. *Under the assumptions of Theorem 1, $\mathcal{J}_i[\alpha, \lambda_i(\mathbf{W})] \leq \mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})]$ and $\alpha_i^*[\lambda_i(\mathbf{W})] \leq \alpha_2^*[\lambda_2(\mathbf{W})]$ for $i = 1$ and $i = 3, 4, \ldots, N$ over the range $\alpha \in [0, -\theta_1^{-1}]$.*

We omit the proof of this lemma due to space limitations (see [15] for details). The remainder of the proof of Theorem 1 proceeds as follows. From Lemmas 1 and 2, the optimization problem (10) simplifies to: $\alpha^\star = \arg\min_{\alpha \geq 0} \mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})]$. We shall now show that α_2^* is a global minimizer of this function. Consider the derivative of $\mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})]$ w.r.t. α on $[0, \alpha_2^*)$:

$$\frac{\partial}{\partial\alpha}\mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})] = \frac{2\theta_1 + (\theta_2 + (\theta_3-1)\,\lambda_2(\mathbf{W}))\,(\lambda_2(\mathbf{W}) + \alpha\,(\theta_2 + (\theta_3-1)\,\lambda_2(\mathbf{W})))}{\sqrt{4\alpha\theta_1 + (\lambda_2(\mathbf{W}) + \alpha\,(\theta_2 + (\theta_3-1)\,\lambda_2(\mathbf{W})))^2}} + (\theta_2 + (\theta_3-1)\,\lambda_2(\mathbf{W}))\,\mathrm{sgn}\,[\lambda_2(\mathbf{W}) + \alpha\,(\theta_2 + (\theta_3-1)\,\lambda_2(\mathbf{W}))]\,.$$

Denote the first term in this sum by $\varphi_1(\lambda_2(\mathbf{W}), \alpha)$ and the second by $\varphi_2(\lambda_2(\mathbf{W}), \alpha)$. It can be shown that $|\varphi_1(\lambda_2(\mathbf{W}), \alpha)| \geq |\varphi_2(\lambda_2(\mathbf{W}), \alpha)|$ for any $\lambda_2(\mathbf{W}) \in [-1, 1]$ and $\alpha \in [0, \alpha_2^*)$ by directly solving the inequality. We conclude that the sign of the derivative on $\alpha \in [0, \alpha_2^*)$ is completely determined by the sign of $\varphi_1(\lambda_2(\mathbf{W}), \alpha)$ for $\lambda_2(\mathbf{W}) \in [-1, 1]$. On $\alpha \in [0, \alpha_2^*)$, the sign of $\varphi_1(\lambda_2(\mathbf{W}), \alpha)$ is determined by the sign of its numerator. The transition point for the numerator's sign occurs at:

$$\alpha^+ = -\frac{2\theta_1 + \lambda_2(\mathbf{W})(\theta_2 + (\theta_3-1)\lambda_2(\mathbf{W}))}{(\theta_2 + (\theta_3-1)\lambda_2(\mathbf{W}))^2},$$

and by showing that $\alpha^+ \geq -\theta_1^{-1}$, we can establish that this transition point is at or beyond α_2^*. This indicates that $\varphi_1(\lambda_2(\mathbf{W}), \alpha) \leq 0$ if $\alpha \in [0, \alpha_2^*)$. We observe that $\mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})]$ is nonincreasing on $\alpha \in [0, \alpha_2^*)$ and nondecreasing on $\alpha \in [\alpha_2^*, -\theta_1^{-1})$ (as established in Lemma 1). We conclude that α_2^* is a global minimum of the function $\mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})]$, thereby proving Theorem 1 and establishing $\mathcal{J}_2[\alpha^\star, \lambda_2(\mathbf{W})] = |\lambda_2^*(\mathbf{\Phi}_3[\alpha^\star])| = \sqrt{-\alpha^\star\theta_1}$.

Note that the last argument also implies that $\mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})] \leq \lambda_2(\mathbf{W})$ on $\alpha \in [0, \alpha_2^*]$ and $\mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})] < 1$ on $\alpha \in (\alpha_2^*, -\theta_1^{-1})$ since $\mathcal{J}_2[\alpha, \lambda_2(\mathbf{W})]$ is non-increasing on the former interval, it is non-decreasing on the latter interval and $\mathcal{J}_2[-\theta_1^{-1}, \lambda_2(\mathbf{W})] = 1$. This fact demonstrates that the matrix $\mathbf{\Phi}_3[\alpha]$ is convergent if $\alpha \in [0, -\theta_1^{-1})$ in the sense that we have $\rho(\mathbf{\Phi}_3[\alpha] - \mathbf{J}) < 1$.

C. Proof of Theorem 2: Convergence Rate

Proof: According to the discussion in Sections III-A and IV-B , we have

$$\begin{aligned}\rho(\mathbf{\Phi}_3[\alpha^\star] - \mathbf{J}) &= |\lambda_2^*(\mathbf{\Phi}_3[\alpha^\star])| = (\alpha^\star|\theta_1|)^{1/2} \\ &= |\theta_1|^{1/2}\left[\frac{-((\theta_3-1)\lambda_2^2 + \theta_2\lambda_2 + 2\theta_1)}{(\theta_2 + (\theta_3-1)\lambda_2)^2} - \frac{-2\sqrt{\theta_1^2 + \theta_1\lambda_2\,(\theta_2 + (\theta_3-1)\lambda_2)}}{(\theta_2 + (\theta_3-1)\lambda_2)^2}\right]^{1/2}.\end{aligned}$$

In order to prove the claim, we consider two cases: $\lambda_2(\mathbf{W}) = 1 - \Psi(N)$, and $\lambda_2(\mathbf{W}) < 1 - \Psi(N)$.

First, we suppose that $\lambda_2(\mathbf{W}) = 1 - \Psi(N)$ and show that $\rho(\mathbf{\Phi}_3[\alpha^\star] - \mathbf{J})^2 - (1 - \sqrt{\Psi(N)})^2 \leq 0$. Denoting $\Psi(N) = \delta$ and substituting $\lambda_2(\mathbf{W}) = 1 - \delta$ and $\theta_1 = 1 - \theta_2 - \theta_3$, we obtain

$$\rho(\mathbf{\Phi}_3[\alpha^\star] - \mathbf{J})^2 - (1 - \sqrt{\Psi(N)})^2 = -\left(\sqrt{\delta} - 1\right)^2 \frac{A}{B}.$$

Where the numerator is

$$\begin{aligned}A = (\theta_3 - 1)\,(\delta^2 - \delta) + 2\sqrt{\delta}\,(\theta_3 + \theta_2 - 1) \\ - 2\sqrt{\delta\,(\theta_2 + (2-\delta)\,(\theta_3 - 1))}\,(\theta_3 + \theta_2 - 1)\end{aligned}$$

and the denominator

$$\begin{aligned}B &= [(2-\delta)\delta+1](1-\theta_3)-(1+\delta)\theta_2\\ &\quad -2\sqrt{\delta\,(\theta_3+\theta_2-1)\,((\theta_3-1)(2-\delta)+\theta_2)}.\end{aligned}$$

It is clear from the assumptions that the expressions under square roots are non-negative. Furthermore, the denominator is negative since $1-\theta_3<0$, $\theta_2>0$ and $\delta\in(0,1)$. Finally, note that $(\theta_3-1)\left(\delta^2-\delta\right)\le 0$ and $2\sqrt{\delta}\,(\theta_3+\theta_2-1)\ge 0$. Thus, to see that the numerator is non-positive, observe that

$$\begin{aligned}&[\sqrt{\delta}\,(\theta_3+\theta_2-1)]^2\\ &\quad -\left[\sqrt{\delta\,(\theta_2+(2-\delta)\,(\theta_3-1))\,(\theta_3+\theta_2-1)}\right]^2\\ &\quad =(\delta-1)\delta(\theta_3-1)(\theta_3+\theta_2-1)\le 0.\end{aligned}$$

Thus, we have $\rho(\mathbf{\Phi}_3[\alpha^\star]-\mathbf{J})^2-(1-\sqrt{\Psi(N)})^2\le 0$, implying that $\rho(\mathbf{\Phi}_3[\alpha^\star]-\mathbf{J})\le 1-\sqrt{\Psi(N)}$ if $\lambda_2(\mathbf{W})=1-\Psi(N)$.

Now suppose $\lambda_2(\mathbf{W})<1-\Psi(N)$. We have seen in Lemma 2 that $\alpha_i^*[\lambda_i(\mathbf{W})]$ is an increasing function of $\lambda_i(\mathbf{W})$, implying $\alpha_2^*[\lambda_2(\mathbf{W})]\le\alpha_2^*[1-\Psi(N)]$. Since $\rho(\mathbf{\Phi}_3[\alpha^\star]-\mathbf{J})=(\alpha^\star|\theta_1|)^{1/2}=(\alpha_2^*[\lambda_2(\mathbf{W})]|\theta_1|)^{1/2}$ is an increasing function of $\alpha_2^*[\lambda_2(\mathbf{W})]$, the claim of theorem follows. ■

V. Concluding Remarks

This paper provides theoretical performance guarantees for accelerated distributed averaging algorithms using node memory. We consider acceleration based on local linear prediction and focus on the setting where each node uses two memory taps. We derived the optimal value of the mixing parameter for the accelerated averaging algorithm, which can be utilized to initialize the proposed algorithm using a fully-decentralized scheme for estimating the spectral radius. An important contribution of this paper is the derivation of upper bounds on the spectral radius of the accelerated consensus matrix. This bound relates the spectral radius growth rate of the original matrix with that of the accelerated consensus matrix. We believe that this result applies to the general class of distributed averaging algorithms using node state prediction, and shows that, even in its simplified form and even at the theoretical level, accelerated consensus may provide considerable improvement in convergence rate. Our current work involves investigating similar theoretical gains in asynchronous averaging algorithms (gossip), and finding ways to extend this analysis to the more general setting where more than one additional memory tap is used at each node.

References

[1] D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs. Manuscript in preparation; available at http://www.stat.berkeley.edu/ aldous/RWG/book.html, Last accessed March, 2009.

[2] T. Aysal, B. Oreshkin, and M. Coates. Accelerated distributed average consensus via localized node state prediction. *IEEE Trans. Sig. Proc.*, 57(4):1563–76, Apr. 2009.

[3] F. Benezit, A. Dimakis, P. Thiran, and M. Vetterli. Gossip along the way: Order-optimal consensus through randomized path averaging. In *Proc. Allerton Conf. on Comm., Control, and Computing*, Urbana-Champaign, IL, Sep. 2007.

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

[5] M. Cao, D. A. Spielman, and E. M. Yeh. Accelerated gossip algorithms for distributed computation. In *Proc. 44th Annual Allerton Conf. Communication, Control, and Computation*, Monticello, IL, USA, Sep. 2006.

[6] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, Feb. 1997.

[7] A. Dimakis, A. Sarwate, and M. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Trans. Signal Processing*, 56(3):1205–1216, Mar. 2008.

[8] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge, 1985.

[9] B. Johansson and M. Johansson. Faster linear iterations for distributed averaging. In *Proc. IFAC World Congress*, Seoul, South Korea, Jul. 2008.

[10] K. Jung, D. Shah, and J. Shin. Fast gossip through lifted Markov chains. In *Proc. Allerton Conf. on Comm., Control, and Computing*, Urbana-Champaign, IL, Sep. 2007.

[11] D. Kempe and F. McSherry. A decentralized algorithm for spectral analysis. In *Proc. ACM Symp. Theory of Computing*, Chicago, IL, USA, June 2004.

[12] E. Kokiopoulou and P. Frossard. Polynomial filtering for fast convergence in distributed consensus. *IEEE Trans. Sig. Process.*, 57(1):342–354, Jan. 2009.

[13] W. Li and H. Dai. Location-aided distributed averaging algorithms: Performance lower bounds and cluster-based variant. In *Proc. Allerton Conf. on Comm., Control, and Computing*, Urbana-Champaign, IL, Sep. 2007.

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

[15] B. Oreshkin, M. Coates, and M. Rabbat. Optimization and analysis of distributed averaging with short node memory. Technical report, Dept. of Electrical and Computer Engineering, McGill University, Mar. 2009. available at http://www.tsp.ece.mcgill.ca/Networks/publications-techreport.html.

[16] D. Scherber and H. Papadopoulos. Locally constructed algorithms for distributed computations in ad-hoc networks. In *Proc. ACM/IEEE Int. Symp. Information Processing in Sensor Networks*, Berkeley, CA, USA, Apr. 2004.

[17] S. Sundaram and C. Hadjicostis. Distributed consensus and linear function calculation in networks: An observability perspective. In *Proc. IEEE/ACM Int. Symp. Information Proc. in Sensor Networks*, Cambridge, MA, USA, Apr. 2007.

[18] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Rev.*, 43(2):235–286, 2001.

[19] J. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. PhD thesis, Massachusetts Institute of Technology, Nov. 1984.

[20] L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Sys. and Control Letters*, 53(1):65–78, Sep. 2004.

[21] L. Xiao, S. Boyd, and S. Lall. A scheme for robust distributed sensor fusion based on average consensus. In *Proc. IEEE/ACM Int. Symp. on Information Processing in Sensor Networks*, Los Angeles, CA, Apr. 2005.