ASYNCHRONOUS DIFFUSION ADAPTATION OVER NETWORKS

Xiaochuan Zhao and Ali H. Sayed

Department of Electrical Engineering
University of California, Los Angeles

ABSTRACT

This work studies the asynchronous behavior of diffusion adaptation strategies for distributed optimization over networks. Under the assumed model, agents in the network may stop updating their estimates or may stop exchanging information at random times. It is expected that asynchronous behavior degrades performance. The analysis quantifies by how much performance degrades and reveals that the learning rate and the mean-square stability conditions of the network are influenced by the rates of occurrence of the asynchronous events.

Index Terms— Distributed optimization, diffusion adaptation, asynchronous behavior, adaptive networks.

1. INTRODUCTION

Distributed optimization over multi-agent networks is an important problem in many contexts, including distributed estimation [1–4], distributed machine learning [5], resource allocation [6], flocking, swarming, and distributed inference and decision [7, 8]. Several decentralized solutions, such as consensus strategies [9–16], incremental strategies [17–19], and diffusion strategies [1–3], have been proposed and studied in the literature. Among these schemes, diffusion strategies are attractive because they are scalable, robust, fully-distributed, and able to endow networks with real-time adaptation and learning abilities.

An underlying assumption used by the traditional diffusion strategies developed in [1–3] is that all agents act synchronously. At every iteration i, each agent k must complete its adaptation step before its neighbors start their combination steps. There is an implicit assumption of *coordinated* behavior throughout the network. In this work, we examine the effect of asynchronous events that can occur randomly across the network. These events may occur as a result of random data arrival times, random agent failures, the turning on and off of agents for energy conservation, or the possibility that some agents are more computationally powerful than others so that they can complete their processing more quickly. Some related and useful work on asynchronous processing for consensus-type strategies can be found in [11–16]. We adopt a more general asynchronous model than before and consider the general diffusion strategies developed in [3] for distributed optimization.

Notation: We use lowercase letters to denote vectors, uppercase letters for matrices, plain letters for deterministic variables, and boldface letters for random variables.

2. ASYNCHRONOUS DIFFUSION ADAPTATION

We consider a connected network consisting of N agents, where the kth agent has an individual cost (or utility) function denoted by $J_k(w)\colon \mathbb{R}^M \longmapsto \mathbb{R}$. The objective of the network is to determine the unique $M \times 1$ vector w^o that uniquely optimizes the following problem:

$$\underset{w}{\text{minimize}} \quad J^{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_k(w) \tag{1}$$

where $\{J_k(w)\}$ are assumed to be differentiable and strongly convex. We assume a common minimizer w^o for all $\{J_k(w)\}$, which corresponds to the important situation where all agents are seeking a common objective. This scenario is common in biological networks, such as fish schools moving towards a food source or away from a predator [8].

A diffusion strategy was devised in [3] to solve (1) in a fully distributed manner. We describe here the Adapt-then-Combine (ATC) form of the algorithm. In this strategy, agents combine information from their immediate neighbors and employ updates of the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \widehat{\nabla_w J_k}(\boldsymbol{w}_{k,i-1}) \tag{2}$$

$$\boldsymbol{w}_{k,i} = \sum_{l \in \mathcal{N}_k} a_{lk} \boldsymbol{\psi}_{l,i} \tag{3}$$

where $\widehat{\nabla_w J_k}(\boldsymbol{w}_{k,i-1})$ denotes a perturbed measurement of the true gradient vector, say, of the form:

$$\widehat{\nabla_w J_k}(\boldsymbol{w}_{k,i-1}) = \nabla_w J_k(\boldsymbol{w}_{k,i-1}) + \boldsymbol{v}_{k,i} \tag{4}$$

where $\nabla_w J_k(\boldsymbol{w}_{k,i-1})$ denotes the gradient of $J_k(w)$ evaluated at $\boldsymbol{w}_{k,i-1}$ and $\boldsymbol{v}_{k,i}$ represents gradient noise that may depend on $\boldsymbol{w}_{k,i-1}$. The nonnegative combination weights $\{a_{lk}\}$ are zero whenever node l is not connected to node k, i.e.,

This work was supported in part by NSF grants CCF-0942936 and CCF-1011918. Email: {xzhao, sayed}@ee.ucla.edu.

$l \notin \mathcal{N}_k$, where $\mathcal{N}_k$ denotes the neighborhood of node k, and they satisfy $\sum_{l\in\mathcal{N}_k} a_{lk} = 1$ for all k. Before proceeding with our model and the subsequent analysis, we state our assumptions on the individual cost functions and gradient noise in a manner similar to [3].

Assumption 1 (Bounded Hessian)**.** *The Hessian matrix of each individual cost function $J_k(w)$ is bounded as:*

$$\lambda_{k,min} I_M \leq \nabla_w^2 J_k(w) \leq \lambda_{k,max} I_M \tag{5}$$

where $0 < \lambda_{k,min} \leq \lambda_{k,max}$. ■

Assumption 2 (Gradient noise)**.** *The gradient noise $\boldsymbol{v}_{k,i}$ satisfies:*

$$\mathbb{E}(\boldsymbol{v}_{k,i}|\boldsymbol{w}_{i-1}) = 0, \quad \mathbb{E}\|\boldsymbol{v}_{k,i}\|^2 \leq \alpha\mathbb{E}\|w^o - \boldsymbol{w}_{k,i-1}\|^2 + \sigma_v^2 \tag{6}$$

where $\boldsymbol{w}_i \triangleq \mathrm{col}\{\boldsymbol{w}_{1,i}, \boldsymbol{w}_{2,i}, \ldots, \boldsymbol{w}_{N,i}\}$, $\alpha \geq 0$, and $\sigma_v^2 \geq 0$ for all i and k. ■

To model asynchronous behavior, we modify the ATC strategy (2)–(3) to the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \boldsymbol{\mu}_k(i)\widehat{\nabla_w J_k}(\boldsymbol{w}_{k,i-1}) \tag{7}$$

$$\boldsymbol{w}_{k,i} = \sum_{l\in\mathcal{N}_k} \boldsymbol{a}_{lk}(i)\boldsymbol{\psi}_{l,i} \tag{8}$$

where $\boldsymbol{\mu}_k(i)$ is now a random step-size for node k at time i and $\{\boldsymbol{a}_{lk}(i)\}$ are random combination weights. We adopt the following assumption on $\{\boldsymbol{\mu}_k(i), \boldsymbol{a}_{lk}(i)\}$.

Assumption 3 (Asynchronous Model)**.**

1. *The step-sizes $\{\boldsymbol{\mu}_k(i)\}$ are distributed as follows:*

$$\boldsymbol{\mu}_k(i) = \begin{cases} \mu_k > 0, & \textit{with probability } p_k \\ 0, & \textit{with probability } 1 - p_k \end{cases} \tag{9}$$

   and they are temporally and spatially independent for different k and i.

2. *The combination weights $\{\boldsymbol{a}_{lk}(i)\}$ are distributed as follows:*

$$\boldsymbol{a}_{lk}(i) = \begin{cases} a_{lk} > 0, & \textit{with probability } q_{lk} \\ 0, & \textit{with probability } 1 - q_{lk} \end{cases} \tag{10}$$

   for all $l \in \mathcal{N}_k\backslash\{k\}$, and they are temporally independent for different i. Node k adjusts its own weight $\boldsymbol{a}_{kk}(i)$ at each iteration by

$$\boldsymbol{a}_{kk}(i) = 1 - \sum_{l\in\mathcal{N}_k\backslash\{k\}} \boldsymbol{a}_{lk}(i) \tag{11}$$

   to ensure $\sum_{l\in\mathcal{N}_k} \boldsymbol{a}_{lk}(i) = 1$.

3. *$\boldsymbol{\mu}_k(i)$ and $\boldsymbol{a}_{lm}(j)$ are mutually-independent for all k, l, m, i, and j; they are also independent of any other random variables.* ■

The second part of Assumption 3 was also used in [14, 16]. It is worth noting that we allow $\{\boldsymbol{a}_{lk}(i)\}$ to be spatially correlated for different l and k. For different realizations of $\{\boldsymbol{\mu}_k(i), \boldsymbol{a}_{lk}(i)\}$, the diffusion strategy (7)–(8) is able to capture various types of asynchronous behavior that may occur.

It was shown in [3] that the synchronous diffusion strategy (2)–(3) is mean-square stable if the step-sizes are sufficiently small. We establish in the sequel that this result still holds for the *asynchronous* diffusion strategy (7)–(8), which means that the diffusion strategy is robust to asynchronous events.

3. MEAN-SQUARE STABILITY ANALYSIS

Let us introduce the error vectors:

$$\widetilde{\boldsymbol{\psi}}_{k,i} \triangleq w^o - \boldsymbol{\psi}_{k,i}, \qquad \widetilde{\boldsymbol{w}}_{k,i} \triangleq w^o - \boldsymbol{w}_{k,i} \tag{12}$$

By (4), the error recursion of (7)–(8) is then given by

$$\widetilde{\boldsymbol{\psi}}_{k,i} = [I_M - \boldsymbol{\mu}_k(i)\boldsymbol{H}_{k,i-1}]\,\widetilde{\boldsymbol{w}}_{k,i-1} + \boldsymbol{\mu}_k(i)\boldsymbol{v}_{k,i} \tag{13}$$

$$\widetilde{\boldsymbol{w}}_{k,i} = \sum_{l\in\mathcal{N}_k} \boldsymbol{a}_{lk}(i)\widetilde{\boldsymbol{\psi}}_{l,i} \tag{14}$$

where $\boldsymbol{H}_{k,i-1}$ is a positive-definite random matrix, defined as

$$\boldsymbol{H}_{k,i-1} \triangleq \int_0^1 \nabla_w^2 J_k\left(w^o - t\cdot\widetilde{\boldsymbol{w}}_{k,i-1}\right) dt \tag{15}$$

Since the squared Euclidean norm $\|\cdot\|^2$ is a convex function, applying Jensen's inequality to (14), the variance of $\widetilde{\boldsymbol{w}}_{k,i}$ can be bounded by

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 \leq \sum_{l\in\mathcal{N}_k} \bar{a}_{lk}\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{l,i}\|^2 \tag{16}$$

where

$$\bar{a}_{lk} \triangleq \mathbb{E}\boldsymbol{a}_{lk}(i) = \begin{cases} q_{lk}a_{lk}, & l\in\mathcal{N}_k\backslash\{k\} \\ 1 - \sum_{l\in\mathcal{N}_k\backslash\{k\}} q_{lk}a_{lk}, & l = k \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

We collect the $\{\boldsymbol{a}_{lk}(i)\}$ and $\{\bar{a}_{lk}\}$ into two $N\times N$ matrices $\boldsymbol{A}_i$ and $\bar{A}$, respectively, such that $\bar{A} = \mathbb{E}\boldsymbol{A}_i$. It can be verified from Assumption 3 that both $\boldsymbol{A}_i$ and $\bar{A}$ are left-stochastic, i.e., $\boldsymbol{A}_i^\mathsf{T}\mathbb{1}_N = \bar{A}^\mathsf{T}\mathbb{1}_N = \mathbb{1}_N$, where $\mathbb{1}_N$ denotes the $N\times 1$ all-one vector. By Assumptions 2 and 3 and (13), we get

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{k,i}\|^2 = \mathbb{E}(\|\widetilde{\boldsymbol{w}}_{k,i-1}\|^2_{\boldsymbol{\Sigma}_{k,i-1}}) + p_k\mu_k^2\,\mathbb{E}\|\boldsymbol{v}_{k,i}\|^2 \tag{18}$$

$$\boldsymbol{\Sigma}_{k,i-1} \triangleq (I_M - \bar{\mu}_k\boldsymbol{H}_{k,i-1})^\mathsf{T}(I_M - \bar{\mu}_k\boldsymbol{H}_{k,i-1}) + p_k(1-p_k)\mu_k^2\boldsymbol{H}_{k,i-1}^\mathsf{T}\boldsymbol{H}_{k,i-1} \tag{19}$$

where

$$\bar{\mu}_k \triangleq \mathbb{E}\,\boldsymbol{\mu}_k(i) = p_k \mu_k \tag{20}$$

From Assumption 1 and expression (15), we obtain

$$0 \le \boldsymbol{\Sigma}_{k,i-1} \le \gamma_k^2 I_M \tag{21}$$

where

$$\gamma_k^2 \triangleq \max\left\{(1-\bar{\mu}_k\lambda_{k,\max})^2, (1-\bar{\mu}_k\lambda_{k,\min})^2\right\} + p_k(1-p_k)\mu_k^2\lambda_{k,\max}^2 \tag{22}$$

By Assumption 2, applying Lemma 3 from [3] and substituting (21) into (18), we can bound the variance of $\widetilde{\boldsymbol{\psi}}_{k,i}$ by

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{k,i}\|^2 \le (\gamma_k^2 + \alpha p_k \mu_k^2)\,\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i-1}\|^2 + p_k\mu_k^2\sigma_v^2 \tag{23}$$

Introduce the global mean-square-deviation (MSD) vector:

$$\widetilde{w}_i \triangleq \text{col}\{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{1,i}\|^2, \mathbb{E}\|\widetilde{\boldsymbol{w}}_{2,i}\|^2, \dots, \mathbb{E}\|\widetilde{\boldsymbol{w}}_{N,i}\|^2\} \tag{24}$$

From (16) and (23), it can be verified that

$$\widetilde{w}_i \preceq \bar{A}^{\mathsf{T}}\Gamma\widetilde{w}_{i-1} + \bar{A}^{\mathsf{T}}\Omega\,\mathbb{1}_N \tag{25}$$

where $\preceq$ denotes element-wise ordering and

$$\Gamma \triangleq \text{diag}\{\gamma_1^2 + \alpha p_1 \mu_1^2, \dots, \gamma_N^2 + \alpha p_N \mu_N^2\} \tag{26}$$

$$\Omega \triangleq \text{diag}\{p_1\mu_1^2\sigma_v^2, \dots, p_N\mu_N^2\sigma_v^2\} \tag{27}$$

Then, using Lemma 4 from [3] and extending the argument from its Appendix A, we arrive at a sufficient condition on the step-sizes for the mean-square stability of the asynchronous diffusion strategy (7)–(8):

$$\boxed{\mu_k < \min\left\{\frac{2\lambda_{k,\max}}{\lambda_{k,\max}^2 + \alpha}, \frac{2\lambda_{k,\min}}{p_k\lambda_{k,\min}^2 + (1-p_k)\lambda_{k,\max}^2 + \alpha}\right\}} \tag{28}$$

Since $0 < p_k < 1$, we get $p_k\lambda_{k,\min}^2 + (1-p_k)\lambda_{k,\max}^2 \ge \lambda_{k,\min}^2$. Thus, the bound (28) is less than or equal to the bound (67) in [3], meaning that the dynamic range of each step-size shrinks due to the asynchronous behavior. If condition (28) is satisfied, then, as $i \to \infty$, it can be shown that

$$\limsup_{i\to\infty}\|\widetilde{w}_i\|_\infty \le \frac{\|\Omega\|_\infty}{1-\|\Gamma\|_\infty} = \frac{\max_k(p_k\mu_k^2\sigma_v^2)}{1-\max_k(\gamma_k^2+\alpha p_k\mu_k^2)} \tag{29}$$

where $\|\cdot\|_\infty$ denotes the ℓ_∞ norm (or the maximum absolute row sum). When the step-sizes $\{\mu_k\}$ are sufficiently small, we can further establish the following inequality:

$$\boxed{\limsup_{i\to\infty}\|\widetilde{w}_i\|_\infty \le \frac{\sigma_v^2}{2}\,\frac{\max_k(p_k)}{\min_k(p_k\lambda_{k,\min})}\,\frac{(\max_k\mu_k)^2}{\min_k\mu_k}} \tag{30}$$

The bound (30) implies that, if step-sizes $\{\mu_k\}$ are sufficiently small, the MSD at each node k, i.e., $\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, can become sufficiently small. This result is clear when the step-sizes are uniform, say, $\mu_k = \mu$. Then, the right-hand side of (30) is of the order of μ.

4. STEADY-STATE PERFORMANCE

So far we established that the steady-state estimators $\{\boldsymbol{w}_{k,i}\}$ of the asynchronous diffusion strategy (7)–(8) converge with high probability to a 2-norm ball $\mathbb{B}(w^o, r) \triangleq \{w \in \mathbb{R}^{M\times 1};\ \|w^o - w\| < r\}$ that is centered at the optimal solution w^o with radius r. The value of r, according to expression (30), is proportional to the step-sizes and can be sufficiently small. It is worth noting that $\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$ may not converge to a *fixed* value (meaning mean-square convergence) but instead may drift and fluctuate in the range $[0, r^2]$. Nevertheless, if r is sufficiently small, we can still find an *approximate* MSD for the asynchronous diffusion strategy (7)–(8). So let us introduce a small step-size assumption.

Assumption 4 (Small step-sizes)**.** *The step-sizes are small enough such that the radius r of the 2-norm ball $\mathbb{B}(w^o, r)$ is also sufficiently small, i.e.,* $r \ll 1$. ■

Based on Assumption 4, the individual cost function $J_k(w)$ can be approximated by a quadratic function that is obtained by truncating the higher terms in its Taylor series expansion, i.e.,

$$J_k(w) \approx J_k(w^o) + \|w^o - w\|^2_{\nabla_w^2 J_k(w^o)},\ \ w \in \mathbb{B}(w^o, r) \tag{31}$$

Then, the original minimization problem (1) is *approximately* equivalent to the following problem

$$\underset{w\in\mathbb{B}(w^o,r)}{\text{minimize}}\ \sum_{k=1}^{N}\|w^o - w\|^2_{\nabla_w^2 J_k(w^o)} \tag{32}$$

We can use the energy conservation technique [20] to evaluate the steady-state MSD for problem (32). Let us denote

$$H_k \triangleq \nabla_w^2 J_k(w^o) \tag{33}$$

Then, the variance relation for the network error vector $\widetilde{\boldsymbol{w}}_i$ can be obtained from (13)–(14) to be

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2_\Sigma = \mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2_{\Sigma'} + \mathbb{E}\|\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{\mathcal{M}}_i\boldsymbol{v}_i\|^2_\Sigma \tag{34}$$

where

$$\boldsymbol{\mathcal{A}}_i \triangleq \boldsymbol{A}_i \otimes I_M \tag{35}$$

$$\boldsymbol{\mathcal{M}}_i \triangleq \text{diag}\{\boldsymbol{\mu}_1(i)I_M, \boldsymbol{\mu}_2(i)I_M, \dots, \boldsymbol{\mu}_N(i)I_M\} \tag{36}$$

$$\boldsymbol{v}_i \triangleq \text{col}\{\boldsymbol{v}_{1,i}, \boldsymbol{v}_{2,i}, \dots, \boldsymbol{v}_{N,i}\} \tag{37}$$

$$\mathcal{H} \triangleq \text{diag}\{H_1, H_2, \dots, H_N\} \tag{38}$$

$$\Sigma' \triangleq \mathbb{E}\,(I_{NM} - \boldsymbol{\mathcal{M}}_i\mathcal{H})\boldsymbol{\mathcal{A}}_i\Sigma\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}(I_{NM} - \boldsymbol{\mathcal{M}}_i\mathcal{H}) \tag{39}$$

Under Assumptions 2 and 4, when $\boldsymbol{w}_{k,i} \in \mathbb{B}(w^o, r)$, or, equivalently, $\widetilde{\boldsymbol{w}}_{k,i} \in \mathbb{B}(0, r)$, the first term on the right-hand side of (6) becomes negligible. Therefore, we shall assume that the gradient noise $\boldsymbol{v}_{k,i}$ is independent of any other random variable and its moments are given by

$$\mathbb{E}\,\boldsymbol{v}_{k,i} = 0, \qquad \mathbb{E}\,\boldsymbol{v}_{k,i}\boldsymbol{v}_{k,i}^{\mathsf{T}} = R_{v,k} \ge 0 \tag{40}$$

We collect the $\{R_{v,k}\}$ into the network covariance matrix

$$\mathcal{R}_v \triangleq \mathbb{E}\,\boldsymbol{v}_i\boldsymbol{v}_i^{\mathsf{T}} = \mathrm{diag}\{R_{v,1}, R_{v,2}, \ldots, R_{v,N}\} \tag{41}$$

In order to exploit the block structure of the weighting matrix Σ' in (39), let us introduce the block vectorization operation $\mathrm{bvec}(\Sigma)$ [1, 21], which transforms an $NM \times NM$ matrix Σ into an $N^2M^2 \times 1$ vector by stacking the columns from each block of Σ on top of each other. Specifically, we first divide Σ into $N \times N$ blocks that are denoted by $\{\Sigma_{lk}\}$ and whose sizes are $M \times M$. The $N^2M^2 \times 1$ vector $\mathrm{bvec}(\Sigma)$ is then defined as

$$\begin{aligned}\mathrm{bvec}(\Sigma) \triangleq \mathrm{col}\{&\mathrm{vec}(\Sigma_{11}), \mathrm{vec}(\Sigma_{21}), \ldots, \mathrm{vec}(\Sigma_{N1}),\\ &\mathrm{vec}(\Sigma_{12}), \mathrm{vec}(\Sigma_{22}), \ldots, \mathrm{vec}(\Sigma_{N2}), \ldots,\\ &\mathrm{vec}(\Sigma_{1N}), \mathrm{vec}(\Sigma_{2N}), \ldots, \mathrm{vec}(\Sigma_{NN})\}\end{aligned} \tag{42}$$

where $\mathrm{vec}(\cdot)$ denotes the vector formed by vertically stacking the columns of its matrix argument. We consider the block Kronecker product of two block matrices $\mathcal{X}$ and $\mathcal{Y}$ [1, 21], and denote it by $\mathcal{X} \otimes_b \mathcal{Y}$. The (l,k)th block of $\mathcal{X} \otimes_b \mathcal{Y}$ is defined as

$$[\mathcal{X} \otimes_b \mathcal{Y}]_{lk} \triangleq \begin{bmatrix} X_{lk} \otimes Y_{11} & X_{lk} \otimes Y_{12} & \ldots & X_{lk} \otimes Y_{1N} \\ X_{lk} \otimes Y_{21} & X_{lk} \otimes Y_{22} & \ldots & X_{lk} \otimes Y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ X_{lk} \otimes Y_{N1} & X_{lk} \otimes Y_{N2} & \ldots & X_{lk} \otimes Y_{NN} \end{bmatrix} \tag{43}$$

where $\{X_{lk}, Y_{lk}\}$ are the (l,k)th blocks of $\{\mathcal{X}, \mathcal{Y}\}$, respectively, and $\otimes$ denotes the Kronecker product. It is shown in [21] that

$$\mathrm{bvec}(\mathcal{X}\Sigma\mathcal{Y}) = (\mathcal{Y} \otimes_b \mathcal{X}^{\mathsf{T}})\,\mathrm{bvec}(\Sigma) \tag{44}$$

Moreover, it can be verified that

$$\mathrm{Tr}(\mathcal{X}\mathcal{Y}) = [\mathrm{bvec}(\mathcal{X}^{\mathsf{T}})]^{\mathsf{T}} \cdot \mathrm{bvec}(\mathcal{Y}) \tag{45}$$

Let us introduce the vector notation:

$$\sigma \triangleq \mathrm{bvec}(\Sigma), \qquad \sigma' \triangleq \mathrm{bvec}(\Sigma') \tag{46}$$

and the $N^2M^2 \times N^2M^2$ matrix $\mathcal{F}$:

$$\mathcal{F} \triangleq \mathbb{E}\,[\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}(I_{NM} - \boldsymbol{\mathcal{M}}_i\mathcal{H})] \otimes_b [\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}(I_{NM} - \boldsymbol{\mathcal{M}}_i\mathcal{H})] \tag{47}$$

Then, from (39) and (44), we can get

$$\sigma' = \mathcal{F} \cdot \sigma \tag{48}$$

Under Assumptions 3 and 4, it can be shown that $\mathcal{F}$ is stable, i.e., $|\rho(\mathcal{F})| < 1$, if the step-sizes $\{\mu_k\}$ also satisfy condition (28). By (36), (41), and (45), we get

$$\mathbb{E}\|\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{\mathcal{M}}_i\boldsymbol{v}_i\|_{\Sigma}^2 = \mathrm{Tr}(\mathcal{Y}\Sigma) = [\mathrm{bvec}\,(\mathcal{Y})]^{\mathsf{T}}\,\mathrm{bvec}(\Sigma) \tag{49}$$

where

$$\mathcal{Y} \triangleq \mathbb{E}\boldsymbol{\mathcal{A}}_i^{\mathsf{T}}\boldsymbol{\mathcal{M}}_i\mathcal{R}_v\boldsymbol{\mathcal{M}}_i\boldsymbol{\mathcal{A}}_i \tag{50}$$

Then, from (48), relation (34) can be rewritten as

$$\boxed{\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{\sigma}^2 = \mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|_{\mathcal{F}\sigma}^2 + [\mathrm{bvec}\,(\mathcal{Y})]^{\mathsf{T}}\,\sigma} \tag{51}$$

where the notation $\|\cdot\|_{\sigma} \equiv \|\cdot\|_{\Sigma}$. From (51) we see that the spectral radius of the matrix $\mathcal{F}$, i.e., $\rho(\mathcal{F})$, determines the convergence rate. Although finding $\rho(\mathcal{F})$ is generally nontrivial, it can be bounded, under Assumption 4, by

$$\begin{aligned}\rho(\mathcal{F}) &\leq \max_k \left[\|I_M - \bar{\mu}_k H_k\|_2^2 + p_k(1-p_k)\mu_k^2\|H_k\|_2^2\right]\\ &\approx 1 - 2\min\nolimits_k(\bar{\mu}_k\lambda_{k,\min})\end{aligned} \tag{52}$$

From (52), we see that the effective step-sizes are $\{\bar{\mu}_k = p_k\mu_k < \mu_k\}$, meaning that the learning rate for each node in the network is reduced. In steady-state, when $i \to \infty$, relation (51) becomes

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|_{(I_{N^2M^2} - \mathcal{F})\sigma}^2 = [\mathrm{bvec}\,(\mathcal{Y})]^{\mathsf{T}}\,\sigma \tag{53}$$

Recall that we are free to choose Σ, so let us select it such that $(I_{N^2M^2} - \mathcal{F})\mathrm{bvec}(\Sigma) = \mathrm{bvec}(I_{NM}/N)$. Then, the network MSD for problem (32), which approximates the MSD for problem (1) under Assumption 4, can be evaluated through (53) as

$$\boxed{\mathrm{MSD} = \frac{1}{N}\,[\mathrm{bvec}\,(\mathcal{Y})]^{\mathsf{T}}\,(I_{N^2M^2} - \mathcal{F})^{-1}\mathrm{bvec}(I_{NM})} \tag{54}$$

5. SIMULATION RESULTS

We examine the theoretical result (54) by simulation. We consider the topology of Fig. 1 with $N = 20$ nodes. The individual cost function for node k is chosen as the mean-squared-error $J_k(w) = \mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2$, where the $M \times 1$ unknown parameter $w^o \in \mathbb{C}^{M\times 1}$ of length $M = 3$ is randomly generated. We assume a linear data model: $\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i}w^o + \boldsymbol{v}_k(i)$, where the scalar measurement data $\boldsymbol{d}_k(i) \in \mathbb{C}$ and the $1 \times M$ regression data $\boldsymbol{u}_{k,i} \in \mathbb{C}^{1\times M}$ are accessible to node k at time i. The regression data $\boldsymbol{u}_{k,i}$ are white but nonuniform across the network, i.e., $R_{u,k} \triangleq \mathbb{E}\boldsymbol{u}_{k,i}^*\boldsymbol{u}_{k,i} = \sigma_{u,k}^2 I_M$, where $\{\sigma_{u,k}^2\}$ are randomly generated. The variances of the zero-mean noise signals $\{\boldsymbol{v}_k(i)\}$ are denoted by $\sigma_{v,k}^2 \triangleq \mathbb{E}|\boldsymbol{v}_k(i)|^2$ and they are randomly generated. The step-sizes $\mu_k = 0.03$ are uniform across the network. We simulated the diffusion algorithm with the uniform combination rule, i.e., $a_{lk} = \frac{1}{|\mathcal{N}_k|}$ for $l \in \mathcal{N}_k$, where $|\mathcal{N}_k|$ denotes the cardinal of the set $\mathcal{N}_k$. Four different cases are simulated: i) 70% idle: $p_k = q_{lk} = 0.7$; ii) 40% idle: $p_k = q_{lk} = 0.4$; iii) 10% idle: $p_k = q_{lk} = 0.1$; and iv) no idle nodes: $p_k = q_{lk} = 1$ (this case corresponds to the traditional synchronous diffusion (2)–(3)). The network MSD curves are averaged over 50 experiments and are plotted in Fig. 2. The simulation results exhibit a good match with theory.

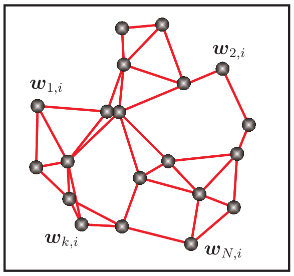

Fig. 1. An adaptive network with $N = 20$ nodes.

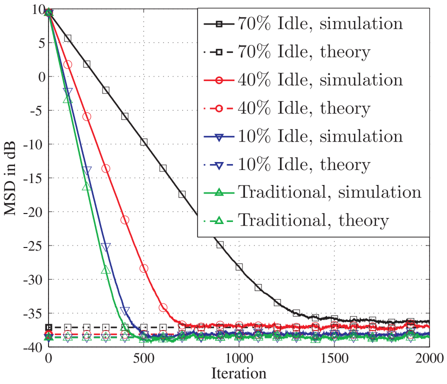

Fig. 2. Network MSD curves for the asyn. diffusion (7)–(8).

We also compare the asynchronous diffusion algorithm (7)–(8) with the corresponding synchronous diffusion (2)–(3) by setting the step-sizes and the combination weights of the latter algorithm to $\mu'_k = \bar{\mu}_k$ and $a'_{lk} = \bar{a}_{lk}$, respectively, where $\bar{\mu}_k$ and $\bar{a}_{lk}$ are given by (20) and (17). We used the diffusion algorithm with the uniform combination rule and simulated the case with 70% idleness, i.e., $p_k = q_{lk} = 0.7$. The network MSD curves are averaged over 50 experiments and are plotted in Fig. 3. Although both algorithms, asynchronous diffusion (7)–(8) and synchronous diffusion (2)–(3), show the same convergence rate, the asynchronous version, as expected, suffers degradation in its MSD performance due to the additional randomness throughout the adaptation process.

6. REFERENCES

[1] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, July 2008.

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

[3] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," to appear in *IEEE Trans. Signal Process.*, 2012, see also arXiv:1111.0034v3 [math.OC] at http://arxiv.org/abs/1111.0034, Oct. 2011.

[4] A. H. Sayed, "Diffusion adaptation over networks," available online at http://arxiv.org/abs/1205.4220 as manuscript arXiv:1205.4220v1 [cs.MA], May 2012.

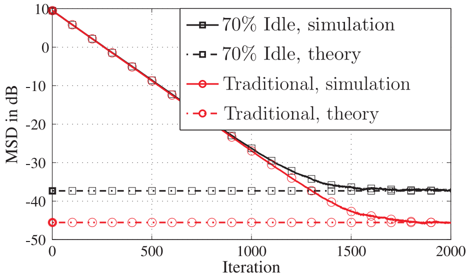

Fig. 3. Comparison of network MSD under 70% idleness.

[5] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction," in *Proc. Int. Conf. Machine Learn. (ICML)*, Bellevue, WA, 2011.

[6] D. Gesbert, S. G. Kiani, A. Gjendemsjo, and G. E. Oien, "Adaptation, coordination, and distributed resource allocation in interference-limited wireless netowrks," *Proc. IEEE*, vol. 95, no. 12, pp. 2393–2409, Dec. 2007.

[7] F. S. Cattivelli and A. H. Sayed, "Modeling bird flight formations using diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2038–2051, May 2011.

[8] S-Y. Tu and A. H. Sayed, "Mobile adaptive networks," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 649–664, Aug. 2011.

[9] J. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Trans. Autom. Control*, vol. 29, no. 1, pp. 42–50, Jan. 1984.

[10] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[11] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, Sept. 1986.

[12] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, June 2006.

[13] S. Kar and J. M. F. Moura, "Sensor networks with random links: Topology design for distributed consensus," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3315–3326, July 2008.

[14] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, Jan. 2009.

[15] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor netowrks: quantized data and random link failures," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1383–1400, Mar. 2010.

[16] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 4, pp. 772–790, Aug. 2011.

[17] D. P. Bertsekas, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, 1997.

[18] A. Nedic and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM J. Optim.*, vol. 12, no. 1, pp. 109–138, 2001.

[19] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 223–229, Aug. 2007.

[20] A. H. Sayed, *Adaptive Filters*, Wiley, NJ, 2008.

[21] R. H. Koning, H. Neudecker, and T. Wansbeek, "Block Kronecker products and the vecb operator," *Linear Algebra Appl.*, vol. 149, pp. 165–184, Apr. 1991.