On the Performance of Exact Diffusion over Adaptive Networks

K. Yuan[†], S. A. Alghunaim[†], B. Ying[†], and A. H. Sayed[‡]

***Abstract*— Various bias-correction methods such as EXTRA, DIGing, and exact diffusion have been proposed recently to solve distributed deterministic optimization problems. These methods employ constant step-sizes and converge linearly to the *exact* solution under proper conditions. However, their performance under stochastic and adaptive settings remains unclear. It is still unknown whether bias-correction is beneficial in stochastic settings. By studying exact diffusion and examining its steady-state performance under stochastic scenarios, this paper provides affirmative results. It is shown that the correction step in exact diffusion can lead to better steady-state performance than traditional methods.**

I. Introduction

This work considers stochastic optimization problems where a collection of K networked agents work cooperatively to solve an aggregate optimization problem of the form

$$w^\star = \arg\min_{w\in\mathbb{R}^M} \sum_{k=1}^{K} J_k(w), \text{ where } J_k(w) = \mathbb{E}\, Q(w;\boldsymbol{x}_k) \quad (1)$$

The local risk function $J_k(w)$ held by agent k is differentiable and strongly convex, and it is constructed as the expectation of some loss function $Q(w;\boldsymbol{x}_k)$. The random variable $\boldsymbol{x}_k$ represents the streaming data received by agent k, and the expectation in $J_k(w)$ is over the distribution of $\boldsymbol{x}_k$. While the cost functions $J_k(w)$ may have *different* local minimizers, all agents seek to determine the *common* global solution $w^\star$ under the constraint that agents can only communicate with their direct neighbors. Problem (1) can find applications in a wide range of areas including wireless sensor networks [1], [2], distributed adaptation and estimation [3]–[5], and distributed statistical learning [6].

There are several techniques that can be used to solve problems of the type (1) such as consensus [7], [8] and diffusion [3]–[5] strategies. The class of diffusion strategies has been shown to be particularly well-suited for online learning with an enhanced stability range over other methods, as well as an improved ability to track drifts in the underlying models and statistics. We therefore focus on this class of algorithms since we are mainly interested in methods that are able to learn and adapt from continuous streaming data. For example, the adapt-then-combine formulation of diffusion takes the following form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\nabla Q(\boldsymbol{w}_{k,i-1};\boldsymbol{x}_{k,i}), \quad \text{(adaptation)} \quad (2)$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{\psi}_{\ell,i}, \quad \text{(combination)} \quad (3)$$

where the subscript k denotes the agent index and i denotes the iteration index. The variable $\boldsymbol{x}_{k,i}$ is the data realization observed by agent k at iteration i. The scalar $a_{\ell k}$ is the weight used by agent k to scale information received from agent ℓ, and $\mathcal{N}_k$ is the set of neighbors of agent k (including k itself). In (2)–(3), variable $\boldsymbol{\psi}_{k,i}$ is an intermediate estimate for $w^\star$ at agent k, while $\boldsymbol{w}_{k,i}$ is the updated estimate. Note that step (2) uses the gradient of the loss function, $Q(\cdot)$, rather than its expected value $J_k(w)$. This is because the statistical properties of the data are not known beforehand. If $J_k(w)$ were known, then we could use its gradient vector in (2). In that case, we would refer to the resulting method as a *deterministic* rather than *stochastic* solution. Throughout this paper, we employ a *constant* step-size μ to enable continuous adaptation and learning in response to drifts in the location of the global minimizer due to changes in the statistical properties of the data. The adaptation and tracking abilities are crucial in many applications — see examples in [4], [5].

Previous studies have shown that both consensus and diffusion methods are able to solve problems of the type (1) well for sufficiently small step-sizes. That is, the squared error $\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$ approaches a small neighborhood around zero for all agents, where $\widetilde{\boldsymbol{w}}_{k,i} = w^\star - \boldsymbol{w}_{k,i}$. Note that these methods do not converge to the *exact* minimizer $w^\star$ of (1) but rather approach a small neighborhood around $w^\star$ with a small steady-state bias under both stochastic and deterministic optimization scenarios. For example, in deterministic settings where the individual costs $J_k(w)$ are known, it is shown in [3] that the squared errors $\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$ generated by the diffusion iterates converge to a $O(\mu^2)$-neighborhood. In this case, this inherent limiting bias is not due to any gradient noise arising from stochastic approximations; it is instead due to the inherent update structure in diffusion and consensus implementations — see the explanations in [9, Sec. III.B]. For stochastic optimization problems, on the other hand, the size of the bias is $O(\mu)$ rather than $O(\mu^2)$ because of the gradient noise.

When high precision is desired, especially in deterministic optimization problems, it would be preferable to remove the $O(\mu^2)$ bias. Motivated by these considerations, the works [9], [10] showed that a simple correction step inserted between the adaptation and combination steps (2) and (3) is sufficient to ensure *exact* convergence of the algorithm to $w^\star$ by all agents — see expression (10) further ahead. In this way, the $O(\mu^2)$ inherent bias is removed completely, and the convergence rate is also improved.

This work was supported in part by NSF grants CCF-1524250 and ECCS-1407712.

[†]Kun Yuan, Sulaiman A. Alghunaim and Bicheng Ying are with the Electrical and Computer Engineering Department, UCLA, CA 90095. Emails:{kunyuan, salghunaim, ybc}@ucla.edu. Bicheng Ying is now with google.

[‡]Ali H. Sayed is with School of Engineering, EPFL, CH-1015 Lausanne, Switzerland e-mail: ali.sayed@epfl.ch.

While the correction of the $O(\mu^2)$ bias is critical in the deterministic setting, it is not clear *whether* it can help in the stochastic and adaptive settings. This motivates us to study exact diffusion in these settings and compare against standard diffusion. To this end, we carry out a higher-order analysis of the error dynamics for both methods, and derive their steady-state performance in both terms $O(\mu)$ and $O(\mu^2)$. In contrast, prior analysis for diffusion have only focused on the $O(\mu)$ term [4], [5]. Our analysis will reveal that the bias in diffusion can get amplified over sparsely-connected graph topologies, and the bias-correction step in exact diffusion can help address this potential deterioration.

A. Our Results

We establish in Theorem 1 that, under sufficiently small step-sizes, the exact diffusion strategy (9)–(11) will converge exponentially fast, at a rate $\rho = 1-O(\mu\nu)$, to a neighborhood around $w^\star$. Moreover, the size of the neighborhood will be characterized as

$$\limsup_{i\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2_{\mathrm{ed}} = O\left(\frac{\mu\sigma^2}{K\nu} + \frac{\delta^2}{\nu^2}\cdot\frac{\mu^2\sigma^2}{1-\lambda}\right) \quad (4)$$

where the subscript ed indicates that $\boldsymbol{w}_{k,i}$ is generated by the exact diffusion method, the quantity σ^2 is a measure of the size of gradient noise, $\lambda \in (0,1)$ is the second largest eigenvalue of the combination matrix $A = [a_{\ell k}]$ which reflects the level of the network connectivity, and ν is the strong convexity constant. In comparison, we will show that the traditional diffusion strategy converges at a similar rate albeit to the following neighborhood:

$$\begin{aligned}\limsup_{i\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2_{\mathrm{d}} \\ = O\left(\frac{\mu\sigma^2}{K\nu} + \frac{\delta^2}{\nu^2}\cdot\frac{\mu^2\sigma^2}{1-\lambda} + \frac{\delta^2}{\nu^2}\cdot\frac{\mu^2 b^2}{(1-\lambda)^2}\right)\end{aligned} \quad (5)$$

where the subscript d indicates that $\boldsymbol{w}_{k,i}$ is generated by the diffusion method, and $b^2 = (1/K)\sum_{k=1}^{K}\|\nabla J_k(w^\star)\|^2$ is a bias constant independent of the gradient noise.

Expressions (4) and (5) have the following important implications. First, it is obvious that diffusion suffers from an inherent bias term $\mu^2 b^2/(1-\lambda)^2$, which is independent of the gradient noise σ^2. In contrast, exact diffusion removes this bias. In fact, in the deterministic setting where the gradient noise $\sigma^2 = 0$, it is observed from (4) and (5) that diffusion converges to a $O(\mu^2)$-neighborhood around the global solution $w^\star$ while exact diffusion converges exactly to $w^\star$. This result is consistent with [3], [4], [10], [11].

Second, it is observed from (4) and (5) that exact diffusion has generally better steady-state mean-square-error performance than diffusion when $b \neq 0$. The superiority of exact diffusion is more obvious when the bias term $\mu^2 b^2/(1-\lambda)^2$ is significant, which can happen when the bias b^2 is large, or the network is sparsely-connected (which happens when λ is close to 1). Under these scenarios, if the step-size is moderately (but not extremely) small such that

$$c_2(1-\lambda)^2\sigma^2/b^2 \leq \mu \leq c_1(1-\lambda) \quad (6)$$

where c_1 and c_2 are constants given in Sec. IV-A, then exact diffusion will perform better than diffusion in steady-state.

Third, the superiority of exact diffusion over diffusion will vanish as step-size μ approaches 0. This is because $O(\mu\sigma^2/K\nu)$ will dominate all other $O(\mu^2)$ terms when μ is sufficiently small, i.e.,

$$\limsup_{i\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|\boldsymbol{w}_{k,i} - w^\star\|^2_{\mathrm{ed}} = O\Big(\mu\sigma^2/K\nu\Big), \quad (7)$$

$$\limsup_{i\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|\boldsymbol{w}_{k,i} - w^\star\|^2_{\mathrm{d}} = O\Big(\mu\sigma^2/K\nu\Big). \quad (8)$$

The "sufficiently" small μ can be roughly characterized as $\mu \leq c_3(1-\lambda)^{2+x}$, where x is any positive constant.

B. Related work

In addition to exact diffusion, there exist other bias-correction methods such as EXTRA [12], ESOM [13], NIDS [14], and gradient-tracking methods such as Aug-DGM [15], NEXT/SONATA [16], [17], DIGing [18], [19], and push-pull methods [20], [21]. All these methods can converge linearly to the exact solution under the deterministic setting, but their performance (especially their advantage over diffusion or consensus) in the stochastic and adaptive settings remains unclear. The work [22] studies a gradient-tracking method under stochastic and adaptive setting and shows its superiority over consensus via numerical simulations. However, it does not analytically discuss *when* and *why* bias-correction methods can outperform consensus. Another useful work is [23], which establishes the convergence property of exact diffusion for stochastic non-convex cost functions and decaying step-sizes. It proves exact diffusion is less sensitive to the data variance across the network than diffusion and is endowed with a better convergence rate when the data variance is large. Different from [23], our bound in (5) shows that even small data variance (i.e., small b^2) can be significantly amplified by a bad network connectivity, see the example graphs discussed in Sec. IV-A. This implies that the superiority of exact diffusion does not only rely on its robustness to data variance, but, more importantly, to the network connectivity as well. We will further clarify in this paper scenarios where exact diffusion and diffusion have the same performance in steady state.

Notation. Throughout the paper we use $\mathrm{col}\{x_1,\cdots,x_K\}$ and $\mathrm{diag}\{x_1,\cdots,x_K\}$ to denote a column vector and a diagonal matrix formed from $x_1,\cdots,x_K$. The notation $\mathbb{1}_K = \mathrm{col}\{1,\cdots,1\} \in \mathbb{R}^K$ and $I_K \in \mathbb{R}^{K\times K}$ is an identity matrix. The Kronecker product is denoted by "$\otimes$".

II. Exact Diffusion Strategy

A. Exact Diffusion Recursions

The exact diffusion strategy from [9], [10] was originally proposed to solve deterministic optimization problems. We adapt it to solve stochastic optimization problems by replacing the gradient of the local cost $J_k(w)$ by the gradient of the corresponding loss function. That is, we now use:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\nabla Q(\boldsymbol{w}_{k,i-1};\boldsymbol{x}_{k,i}), \text{ (adaptation)} \quad (9)$$

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{\psi}_{k,i} + \boldsymbol{w}_{k,i-1} - \boldsymbol{\psi}_{k,i-1}, \qquad \text{(correction)} \quad (10)$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k} \bar{a}_{\ell k}\boldsymbol{\phi}_{\ell,i}. \qquad \text{(combination)} \quad (11)$$

Observe that the fusion step (11) now employs the corrected iterates from (10) rather than the intermediate iterates from (9). The recursions (9)–(11) can start from any $\boldsymbol{w}_{k,-1}$, but we need to set $\boldsymbol{\psi}_{k,-1} = \boldsymbol{w}_{k,-1}$ for all k in initialization. Note that the weight $\bar{a}_{\ell k}$ is different from $a_{\ell k}$ used in diffusion recursion (3). If we let $A = [a_{\ell k}] \in \mathbb{R}^{K\times K}$ and $\bar{A} = [\bar{a}_{\ell k}] \in \mathbb{R}^{K\times K}$ denote the combination matrices used in diffusion and exact diffusion respectively, the relation between them is $\bar{A} = (A + I_K)/2$. In the paper, we assume A (and hence $\bar{A}$) are symmetric and doubly stochastic.

As explained in [9], [10], exact diffusion is essentially a primal-dual method. We can describe its operation more succinctly by collecting the iterates and gradients from across the network into global vectors. Specifically, we introduce

$$\boldsymbol{\mathcal{W}}_i = \begin{bmatrix} \boldsymbol{w}_{1,i} \\ \vdots \\ \boldsymbol{w}_{K,i} \end{bmatrix}, \ \nabla\mathcal{Q}(\boldsymbol{\mathcal{W}}_{i-1};\boldsymbol{\mathcal{X}}_i) = \begin{bmatrix} \nabla Q(\boldsymbol{w}_{1,i-1};\boldsymbol{x}_{1,i}) \\ \vdots \\ \nabla Q(\boldsymbol{w}_{K,i-1};\boldsymbol{x}_{K,i}) \end{bmatrix} \quad (12)$$

$\mathcal{A} = A\otimes I_K$ and $\overline{\mathcal{A}} = (\mathcal{A} + I_{KM})/2$. Then recursions (9) – (11) lead to the second-order recursion

$$\boldsymbol{\mathcal{W}}_i = \overline{\mathcal{A}}\Big(2\boldsymbol{\mathcal{W}}_{i-1} - \boldsymbol{\mathcal{W}}_{i-2} - \mu\nabla\mathcal{Q}(\boldsymbol{\mathcal{W}}_{i-1};\boldsymbol{\mathcal{X}}_i) + \mu\nabla\mathcal{Q}(\boldsymbol{\mathcal{W}}_{i-2};\boldsymbol{\mathcal{X}}_{i-1})\Big). \quad (13)$$

We can rewrite this update in a primal-dual form as follows. First, since the combination matrix $\bar{A}$ is symmetric and doubly stochastic, it holds that $I - \bar{A}$ is positive semi-definite. By decomposing $I - \bar{A} = U\Sigma U^{\mathsf{T}}$ and defining $V = U\Sigma^{1/2}U^{\mathsf{T}} \in \mathbb{R}^{K\times K}$, where Σ is a non-negative diagonal matrix, we know that V is also positive semi-definite and $V^2 = I - \bar{A}$. Furthermore, if we let $\mathcal{V} = V\otimes I_K$ then $\mathcal{V}^2 = I_{KM} - \overline{\mathcal{A}}$ holds. With these relations, it can be verified[1] that recursion (13) is equivalent to

$$\begin{cases} \boldsymbol{\mathcal{W}}_i = \overline{\mathcal{A}}\big(\boldsymbol{\mathcal{W}}_{i-1} - \mu\nabla\mathcal{Q}(\boldsymbol{\mathcal{W}}_{i-1};\boldsymbol{\mathcal{X}}_i)\big) - \mathcal{V}\boldsymbol{\mathcal{Y}}_{i-1}, \\ \boldsymbol{\mathcal{Y}}_i = \boldsymbol{\mathcal{Y}}_{i-1} + \mathcal{V}\boldsymbol{\mathcal{W}}_i, \end{cases} \quad (14)$$

where $\mathcal{Y}_i \in \mathbb{R}^{KM}$ plays the role of a dual variable. The analysis in [9], [10] explains how the correction term in (10) guarantees *exact* convergence to $w^\star$ by all agents in deterministic optimization problems where the true gradient $\nabla J_k(w)$ is available. In the following sections, we will examine the convergence of exact diffusion (9)–(11) in the stochastic setting.

III. Error Dynamics of Exact Diffusion

To establish the error dynamics of exact diffusion, we first introduce some standard assumptions.

*Assumption 1 (*Conditions on cost functions*):* Each $J_k(w)$ is ν-strongly convex and twice differentiable, and its Hessian matrix satisfies

$$\nu I_M \le \nabla^2 J_k(w) \le \delta I_M, \quad \forall\ k. \quad (15)$$

■

*Assumption 2 (*Conditions on combination matrix*):* The network is undirected and strongly connected, and the combination matrix A satisfies

$$A = A^{\mathsf{T}}, \quad A\mathbb{1}_K = \mathbb{1}_K, \quad \mathbb{1}_K^{\mathsf{T}}A = \mathbb{1}_K^{\mathsf{T}}. \quad (16)$$

■

Since the network is strongly connected, it holds that

$$1 = \lambda_1(\bar{A}) > \lambda_2(\bar{A}) \ge \cdots \ge \lambda_K(\bar{A}) > 0. \quad (17)$$

To establish the optimality condition for problem (1), we introduce the following notation:

$$\mathcal{w} = \mathrm{col}\{w_1,\cdots,w_K\} \in \mathbb{R}^{KM}, \quad (18)$$

$$\nabla\mathcal{J}(\mathcal{w}) = \mathrm{col}\{\nabla J_1(w_1),\cdots,\nabla J_K(w_K)\}, \quad (19)$$

where w_k in (18) is the k-th block entry of vector $\mathcal{w}$. With the above notation, the following lemma from [10] states the optimality condition for problem (1).

*Lemma 1 (*Optimality Condition*):* Under Assumption 1, if some block vectors $(\mathcal{w}^\star, \mathcal{y}^\star)$ exist that satisfy:

$$\mu\overline{\mathcal{A}}\nabla\mathcal{J}(\mathcal{w}^\star) + \mathcal{V}\mathcal{y}^\star = 0, \quad (20)$$

$$\mathcal{V}\mathcal{w}^\star = 0. \quad (21)$$

then it holds that each block entries in $\mathcal{w}^\star$ satisfy:

$$w_1^\star = w_2^\star = \cdots = w_N^\star = w^\star \quad (22)$$

where $w^\star$ is the unique solution to problem (1). ■

A. Error Dynamics

We define the gradient noise at agent k as

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \nabla Q(\boldsymbol{w}_{k,i-1};\boldsymbol{x}_{k,i}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \quad (23)$$

and introduce the network vectors:

$$\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}) = \mathrm{col}\{\boldsymbol{s}_{1,i}(\boldsymbol{w}_{1,i-1}),\cdots,\boldsymbol{s}_{K,i}(\boldsymbol{w}_{K,i-1})\} \quad (24)$$

$$\nabla\mathcal{J}(\boldsymbol{\mathcal{W}}_{i-1}) = \mathrm{col}\{\nabla J_1(\boldsymbol{w}_{1,i-1}),\cdots,\nabla J_K(\boldsymbol{w}_{K,i-1})\} \quad (25)$$

It then follows that

$$\nabla\mathcal{Q}(\boldsymbol{\mathcal{W}}_{i-1};\boldsymbol{\mathcal{X}}_i) = \nabla\mathcal{J}(\boldsymbol{\mathcal{W}}_{i-1}) + \boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}). \quad (26)$$

Next, we introduce the error vectors

$$\widetilde{\boldsymbol{\mathcal{W}}}_i = \mathcal{w}^\star - \boldsymbol{\mathcal{W}}_i, \quad \widetilde{\boldsymbol{\mathcal{Y}}}_i = \mathcal{y}^\star - \boldsymbol{\mathcal{Y}}_i \quad (27)$$

where $(\mathcal{w}^\star, \mathcal{y}^\star)$ are optimal solutions satisfying (20)–(21). By combining (14), (20), (21), (26) and (27), we reach

$$\begin{cases} \widetilde{\boldsymbol{\mathcal{W}}}_i = \overline{\mathcal{A}}\big[\widetilde{\boldsymbol{\mathcal{W}}}_{i-1} + \mu(\nabla\mathcal{J}(\boldsymbol{\mathcal{W}}_{i-1}) - \nabla\mathcal{J}(\boldsymbol{\mathcal{W}}^\star))\big] \\ \qquad\qquad -\mathcal{V}\widetilde{\boldsymbol{\mathcal{Y}}}_{i-1} + \mu\overline{\mathcal{A}}\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}), \\ \widetilde{\boldsymbol{\mathcal{Y}}}_i = \widetilde{\boldsymbol{\mathcal{Y}}}_{i-1} + \mathcal{V}\widetilde{\boldsymbol{\mathcal{W}}}_i. \end{cases} \quad (28)$$

Since each $J_k(w)$ is twice-differentiable (see Assumption 1), we can appeal to the mean-value theorem from Lemma D.1 in [4], which allows us to express each difference in (28) in terms of Hessian matrices for any $k = 1, 2, \ldots, N$:

$$\nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(w^\star) = -\boldsymbol{H}_{k,i-1}\widetilde{\boldsymbol{w}}_{k,i-1},$$

where

$$\boldsymbol{H}_{k,i-1} \triangleq \int_0^1 \nabla^2 J_k\big(w^\star - r\widetilde{\boldsymbol{w}}_{k,i-1}\big)dr \in \mathbb{R}^{M\times M} \quad (29)$$

[1] To verify it, one can substitute the second recursion in (14) into the first recursion to remove $\boldsymbol{\mathcal{Y}}_i$ and arrive at (13).

We introduce the block diagonal matrix

$$\boldsymbol{\mathcal{H}}_{i-1} \triangleq \mathrm{diag}\{\boldsymbol{H}_{1,i-1}, \boldsymbol{H}_{2,i-1}, \cdots, \boldsymbol{H}_{K,i-1}\} \tag{30}$$

so that

$$\nabla \mathcal{J}(\boldsymbol{\mathcal{W}}_{i-1}) - \nabla \mathcal{J}(\mathcal{W}^\star) = -\boldsymbol{\mathcal{H}}_{i-1}\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}. \tag{31}$$

Substituting (31) into the first recursion in (28), we reach

$$\begin{cases} \widetilde{\boldsymbol{\mathcal{W}}}_i = \overline{\mathcal{A}}(I_{KM} - \mu\boldsymbol{\mathcal{H}}_{i-1})\widetilde{\boldsymbol{\mathcal{W}}}_{i-1} - \mathcal{V}\widetilde{\boldsymbol{\mathcal{Y}}}_{i-1} + \mu\overline{\mathcal{A}}\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}), \\ \widetilde{\boldsymbol{\mathcal{Y}}}_i = \widetilde{\boldsymbol{\mathcal{Y}}}_{i-1} + \mathcal{V}\widetilde{\boldsymbol{\mathcal{W}}}_i. \end{cases} \tag{32}$$

Next, if we substitute the first recursion in (32) into the second one, and recall that $\mathcal{V}^2 = I_{KM} - \overline{\mathcal{A}}$, we reach the following error dynamics.

*Lemma 2 (*ERROR DYNAMICS): Under Assumption 1, the error dynamics for the exact diffusion recursions (9)–(11) is as follows:

$$\begin{bmatrix} \widetilde{\boldsymbol{\mathcal{W}}}_i \\ \widetilde{\boldsymbol{\mathcal{Y}}}_i \end{bmatrix} = \Big(\underbrace{\begin{bmatrix} \overline{\mathcal{A}} & -\mathcal{V} \\ \mathcal{V}\overline{\mathcal{A}} & \overline{\mathcal{A}} \end{bmatrix}}_{\triangleq\, \mathcal{B}} - \mu \underbrace{\begin{bmatrix} \overline{\mathcal{A}}\boldsymbol{\mathcal{H}}_{i-1} & 0 \\ \mathcal{V}\overline{\mathcal{A}}\boldsymbol{\mathcal{H}}_{i-1} & 0 \end{bmatrix}}_{\triangleq\, \boldsymbol{\mathcal{T}}_{i-1}} \Big) \begin{bmatrix} \widetilde{\boldsymbol{\mathcal{W}}}_{i-1} \\ \widetilde{\boldsymbol{\mathcal{Y}}}_{i-1} \end{bmatrix} + \mu \underbrace{\begin{bmatrix} \overline{\mathcal{A}} \\ \mathcal{V}\overline{\mathcal{A}} \end{bmatrix}}_{\triangleq\, \mathcal{B}_\ell} \boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}), \tag{33}$$

and $\boldsymbol{\mathcal{H}}_i$ is defined in (30). ■

B. Transformed Error Dynamics

The direct convergence analysis of recursion (33) is still challenging. To facilitate the analysis, we identify a convenient change of basis and transform (33) into another equivalent form that is easier to handle. To this end, we introduce a fundamental decomposition from [10] here.

*Lemma 3 (*FUNDAMENTAL DECOMPOSITION): Under Assumptions 1 and 2, the matrix $\mathcal{B}$ defined in (33) can be decomposed as

$$\mathcal{B} = \underbrace{\begin{bmatrix} \mathcal{R}_1 & \mathcal{R}_2 & c\mathcal{X}_R \end{bmatrix}}_{\mathcal{X}} \underbrace{\begin{bmatrix} I_M & 0 & 0 \\ 0 & I_M & 0 \\ 0 & 0 & \mathcal{D}_1 \end{bmatrix}}_{\mathcal{D}} \underbrace{\begin{bmatrix} \mathcal{L}_1^{\mathsf{T}} \\ \mathcal{L}_2^{\mathsf{T}} \\ \frac{1}{c}\mathcal{X}_L \end{bmatrix}}_{\mathcal{X}^{-1}} \tag{34}$$

where c can be any positive constant, and $\mathcal{D} \in \mathbb{R}^{2KM \times 2KM}$ is a diagonal matrix. Moreover, we have

$$\mathcal{R}_1 = \begin{bmatrix} \mathcal{I} \\ 0 \end{bmatrix} \in \mathbb{R}^{2KM \times M}, \quad \mathcal{R}_2 = \begin{bmatrix} 0 \\ \mathcal{I} \end{bmatrix} \in \mathbb{R}^{2KM \times M}, \tag{35}$$

$$\mathcal{L}_1 = \begin{bmatrix} \frac{1}{K}\mathcal{I} \\ 0 \end{bmatrix} \in \mathbb{R}^{2KM \times M}, \ \mathcal{L}_2 = \begin{bmatrix} 0 \\ \frac{1}{K}\mathcal{I} \end{bmatrix} \in \mathbb{R}^{2KM \times M}, \tag{36}$$

$$\mathcal{X}_R \in \mathbb{R}^{2KM \times 2(K-1)M}, \qquad \mathcal{X}_L \in \mathbb{R}^{2(K-1)M \times 2KM}. \tag{37}$$

where $\mathcal{I} = \mathbb{1}_K \otimes I_M$. Also, the matrix $\mathcal{D}_1$ is a diagonal matrix with complex entries. The magnitudes of the diagonal entries are all strictly less than 1. ■

By multiplying $\mathcal{X}^{-1}$ to both sides of the error dynamics (33) and simplifying we arrive at the following result.

*Lemma 4 (*TRANSFORMED ERROR DYNAMICS): Under Assumption 1 and 2, the transformed error dynamics for exact diffusion recursions (9)–(11) is as follows:

$$\begin{bmatrix} \bar{\boldsymbol{z}}_i \\ \check{\boldsymbol{z}}_i \end{bmatrix} = \begin{bmatrix} I_M - \frac{\mu}{K}\sum_{k=1}^K \boldsymbol{H}_{k,i-1} & -\frac{c\mu}{K}\mathcal{I}^{\mathsf{T}}\boldsymbol{\mathcal{H}}_{i-1}\mathcal{X}_{R,u} \\ -\frac{\mu}{c}\mathcal{X}_L\boldsymbol{\mathcal{T}}_{i-1}\mathcal{R}_1 & \mathcal{D}_1 - \mu\mathcal{X}_L\boldsymbol{\mathcal{T}}_{i-1}\mathcal{X}_R \end{bmatrix} \times \begin{bmatrix} \bar{\boldsymbol{z}}_{i-1} \\ \check{\boldsymbol{z}}_{i-1} \end{bmatrix} + \mu \begin{bmatrix} \frac{1}{K}\mathcal{I}^{\mathsf{T}} \\ \frac{1}{c}\mathcal{X}_L\mathcal{B}_\ell \end{bmatrix} \boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1}). \tag{38}$$

where $\mathcal{X}_{R,u} \in \mathbb{R}^{KM \times 2(K-1)M}$ is the upper part of matrix $\mathcal{X}_R = [\mathcal{X}_{R,u}; \mathcal{X}_{R,d}]$. The relation between the original and transformed error vectors are

$$\begin{bmatrix} \widetilde{\boldsymbol{\mathcal{W}}}_i \\ \widetilde{\boldsymbol{\mathcal{Y}}}_i \end{bmatrix} = \begin{bmatrix} \mathcal{R}_1 & c\mathcal{X}_R \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{z}}_i \\ \check{\boldsymbol{z}}_i \end{bmatrix}. \tag{39}$$

■

IV. MEAN-SQUARE CONVERGENCE

Using the transformed error dynamics derived in (38), we can now analyze the mean-square convergence of exact diffusion (9)–(11) in the stochastic and adaptive setting. To begin with, we introduce the filtration

$$\boldsymbol{\mathcal{F}}_{i-1} = \mathrm{filtration}\{\boldsymbol{w}_{k,-1}, \boldsymbol{w}_{k,0}, \cdots, \boldsymbol{w}_{k,i-1}, \text{ all } k\}. \tag{40}$$

The following assumption is standard on the gradient noise process (see [4], [22]).

*Assumption 3 (*CONDITIONS ON GRADIENT NOISE): It is assumed that the first and second-order conditional moments of the individual gradient noises for any k and i satisfy

$$\mathbb{E}[\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})|\boldsymbol{\mathcal{F}}_{i-1}] = 0, \tag{41}$$

$$\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \le \beta_k^2\|\widetilde{\boldsymbol{w}}_{k,i-1}\|^2 + \sigma_k^2 \tag{42}$$

for some constants β_k and σ_k. Moreover, we assume each $\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})$ is independent for any k, i given $\boldsymbol{\mathcal{F}}_{i-1}$. ■

With Assumption 3, it can be verified that

$$\mathbb{E}[\boldsymbol{s}_i(\boldsymbol{\mathcal{W}}_{i-1})|\boldsymbol{\mathcal{F}}_{i-1}] = 0, \quad \forall\ i, \tag{43}$$

$$\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^K \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\Big\|^2\Big|\boldsymbol{\mathcal{F}}_{i-1}\Big] \le \frac{\beta^2}{K}\|\widetilde{\boldsymbol{\mathcal{W}}}_{i-1}\|^2 + \frac{\sigma^2}{K} \tag{44}$$

where $\beta^2 \triangleq \max_k\{\beta_k^2\}/K$ and $\sigma^2 \triangleq \sum_{k=1}^K \sigma_k^2/K$.

*Theorem 1 (*MEAN-SQUARE CONVERGENCE): Under Assumptions 1–3, if the step-size μ satisfies

$$\mu \le \frac{(1-\lambda)\nu}{(32 + 16c_1c_2 + 8\sqrt{c_1c_2})(\delta^2 + \beta^2)} = O\left(\frac{(1-\lambda)\nu}{\delta^2 + \beta^2}\right) \tag{45}$$

where $\lambda = \lambda_2(A)$, and c_1, c_2 are constants independent of λ, ν, δ, and β, then the exact diffusion recursion (14) converges exponentially fast to a neighborhood around $w^\star$. The convergence rate is $\rho = 1 - O(\mu\nu)$, and the size of the neighborhood can be characterized as follows:

$$\limsup_{i\to\infty} \frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2 = O\left(\frac{\mu\sigma^2}{K\nu} + \frac{\delta^2}{\nu^2}\cdot\frac{\mu^2\sigma^2}{1-\lambda}\right) \tag{46}$$

Proof. We omit the proof due to space limitations. The detail can be referred to Appendix A in the long report [24]. ■

Theorem 1 indicates that when μ is smaller than a specified upper bound, the exact diffusion over adaptive networks is stable. The theorem also provides a bound on the size of the steady-state mean-square error. To compare exact diffusion

with diffusion, we examine the mean-square convergence property of diffusion as well. The proof of the following result can be found in Appendix B of the long report [24].

*Lemma 5 (*MEAN-SQUARE STABILITY OF DIFFUSION): Under Assumptions 1–3, if μ satisfies

$$\mu \leq \frac{(1-\lambda)\nu}{(12+2e_1e_2+\sqrt{6e_1e_2})(\delta^2+\beta^2)} = O\left(\frac{(1-\lambda)\nu}{\delta^2+\beta^2}\right) \quad (47)$$

where $\lambda = \lambda_2(A)$, e_1 and e_2 are constants independent of λ, δ, ν and β, then the diffusion recursions (2)–(3) converge exponentially fast to a neighborhood around $w^\star$. The convergence rate is $1 - O(\mu\nu)$, and the size of the neighborhood can be characterized as follows

$$\limsup_{i\to\infty} \frac{1}{K}\sum_{k=1}^{K} \|\widetilde{\boldsymbol{w}}_{k,i}\|^2 = O\left(\frac{\mu\sigma^2}{K\nu} + \frac{\delta^2}{\nu^2}\cdot\frac{\mu^2\sigma^2}{1-\lambda} + \frac{\delta^2}{\nu^2}\cdot\frac{\mu^2 b^2}{(1-\lambda)^2}\right), \quad (48)$$

where $b^2 = (1/K)\sum_{k=1}^{K}\|\nabla J_k(w^\star)\|^2$ is a bias term. ■

Comparing (46) and (48), it is observed the expressions for both algorithms consist of two major terms – one $O(\mu)$ term and one $O(\mu^2)$ term. However, diffusion suffers from an additional bias term $O(\mu^2 b^2/(1-\lambda)^2)$. In the following, we compare diffusion and exact diffusion in two scenarios.

A. Bias term is significant

When b^2 is large, or the network is sparse, it is possible that the bias term $\mu^2 b^2/(1-\lambda)^2$ is significant. We assume such bias term in (48) is significant if

$$\frac{\delta^2}{\nu^2}\cdot\frac{\mu^2 b^2}{(1-\lambda)^2} \geq \frac{\mu\sigma^2}{\nu} \quad (49)$$

from which we get $\mu \geq (1-\lambda)^2\sigma^2\nu/\delta^2 b^2$. Combining with (45), we conclude that if step-size μ satisfies

$$\frac{d_1(1-\lambda)^2\sigma^2\nu}{\delta^2 b^2} \leq \mu \leq \frac{d_2(1-\lambda)\nu}{\delta^2+\beta^2}, \quad (50)$$

where d_1 and d_2 are some constants, then the bias term in (48) is significant and exact diffusion is expected to have better performance than diffusion in steady-state. To make the interval in (50) valid, it is enough to let

$$\frac{d_1(1-\lambda)^2\sigma^2\nu}{\delta^2 b^2} < \frac{d_2(1-\lambda)\nu}{\delta^2+\beta^2} \iff \frac{b^2}{1-\lambda} > \frac{d_1}{d_2}\sigma^2. \quad (51)$$

In the following example, we list several network topologies in which the inherent bias $\mu^2 b^2/(1-\lambda)^2$ dominates (5) easily.

Example (Sparse networks). Consider a linear or cyclic network with K agents where each node connects with its previous and next neighbors. It is shown in [25] that

$$1-\lambda = O(1/K^2). \quad (52)$$

Therefore, the bias term in diffusion becomes $O(\mu^2 b^2 K^4)$, which increases rapidly with the size of the network. For a grid network with K agents where each node connects with its neighbors from left, right, behind and front. It is shown in [26] that

$$1-\lambda = O(1/K) \quad (53)$$

for grid networks. Therefore, the bias term in diffusion is $O(\mu^2 b^2 K^2)$ which also increases with K. ■

B. Bias term is trivial

In theory, if we adjust μ to be sufficiently small, the $O(\mu)$ term in both expressions (46) and (48) will eventually dominate for any b^2 and λ. In such scenario, it holds that

$$\limsup_{i\to\infty} \frac{1}{K}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2_{\mathrm{ed}} = O\Big(\frac{\mu\sigma^2}{K\nu}\Big), \quad (54)$$

$$\limsup_{i\to\infty} \frac{1}{K}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2_{\mathrm{d}} = O\Big(\frac{\mu\sigma^2}{K\nu}\Big). \quad (55)$$

It is observed that both diffusion and exact diffusion will have the same mean-square error order, which implies that diffusion and exact diffusion will perform similarly in this scenario. Such "sufficiently" small step-size can be roughly characterized by the range

$$\mu \leq d_3(1-\lambda)^{2+x} \quad \text{where} \quad x > 0. \quad (56)$$

For example, we can substitute it into (46) to verify

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2_{\mathrm{ed}} = O\Big(\mu + \frac{\mu^2}{1-\lambda}\Big) = O\Big((1-\lambda)^{2+x}\Big) = O(\mu) \quad (57)$$

and we can also verify $\limsup_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2_{\mathrm{d}} = O(\mu)$ with the same technique.

V. NUMERICAL SIMULATION

In this section we compare the performance of exact diffusion and diffusion when solving the decentralized logistic regression problem:

$$\min_{w\in\mathbb{R}^M} \sum_{k=1}^{K} \mathbb{E}\left\{\ln\left(1+e^{-\boldsymbol{\gamma}_k \boldsymbol{h}_k^{\mathsf{T}} w}\right)\right\} + \frac{\rho}{2}\|w\|^2, \quad (58)$$

where $(\boldsymbol{h}_k, \boldsymbol{\gamma}_k)$ represent the streaming data received by agent k. Variable $\boldsymbol{h}_k \in \mathbb{R}^M$ is the feature vector and $\boldsymbol{\gamma}_k \in \{-1, +1\}$ is the label scalar. In all experiments, we set $M = 20$ and $\rho = 0.001$. To make the $J_k(w)$'s have different minimizers, we first generate K different local minimizers $\{w_k^\star\}$. All $w_k^\star$ are normalized so that $\|w_k^\star\|^2 = 1$. At agent k, we generate each feature vector $\boldsymbol{h}_{k,i} \sim \mathcal{N}(0, I_{20})$. To generate the corresponding label $\boldsymbol{\gamma}_k(i)$, we generate a random variable $z_{k,i} \in \mathcal{U}(0,1)$. If $z_{k,i} \leq 1/(1+\exp(-\boldsymbol{h}_{k,i}^{\mathsf{T}} w_k^\star))$, we set $\boldsymbol{\gamma}_k(i) = 1$; otherwise $\boldsymbol{\gamma}_k(i) = -1$. The MSD in y-axis indicates mean-square deviation $\sum_{k=1}^{K}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$.

In the following, we run two sets of simulations. In the first set, we test the performance of diffusion and exact diffusion over cyclic networks with different size K. For each simulation, we fix $\mu = 0.01$ and compare diffusion and exact diffusion for $K = 10$ and $K = 35$. When the size of the cyclic network becomes larger, we know from examples in Sec.IV-A that the inherent bias $O(\mu^2 b^2/(1-\lambda)^2)$ will increase drastically. In this scenario, we can expect exact diffusion to have better performance in steady-state. The left and middle plots in Figure 1 confirm this conclusion. It is observed that when the network is small, both diffusion and exact diffusion performs almost the same since the inherent bias is trivial. However, as the size K increases, the term

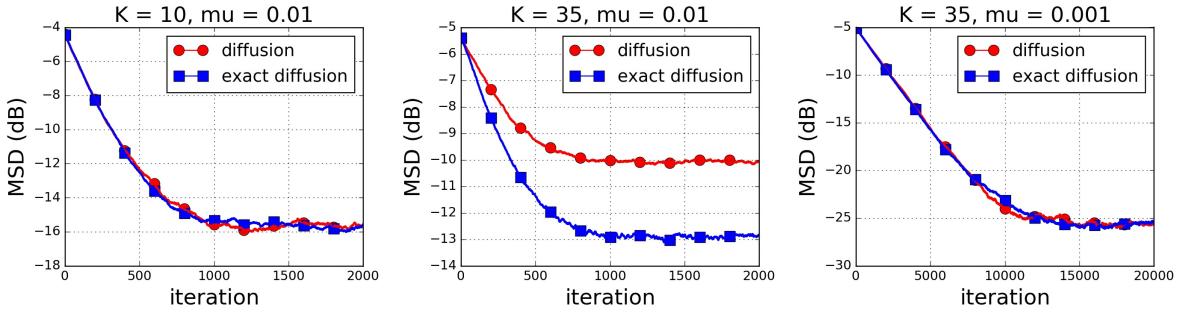

Fig. 1. Diffusion v.s. exact diffusion over cyclic networks.

$O(\mu^2 b^2/(1-\lambda)^2)$ becomes dominant and exact diffusion is significantly better than diffusion.

In the second set of simulations, we fix the cyclic network size $K = 35$ and compare diffusion and exact diffusion at $\mu = 0.01$ and $\mu = 0.001$. As we discussed in Sec. IV-B, the $O(\mu)$ term will gradually dominate all other higher-order terms as $\mu \to 0$. As a result, we can expect diffusion and exact diffusion to match with each other as μ becomes sufficiently small. The middle and right plots in Figure 1 confirm this conclusion. When $\mu = 0.001$, both methods perform almost the same.

REFERENCES

[1] L. A. Rossi, B. Krishnamachari, and C. C.J. Kuo, "Distributed parameter estimation for monitoring diffusion phenomena using physical models," in *Proc. IEEE Conference on Sensor and Ad Hoc Communications and Networks (SECON)*, Santa Clara, CA, 2004, pp. 460–469.

[2] D. Li, K. D. Wong, Y. Hu, and A. M. Sayeed, "Detection, classification, and tracking of targets," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 17–29, 2002.

[3] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion strategies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 205–220, 2013.

[4] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[5] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, April 2014.

[6] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2012.

[7] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[8] S. Kar and J. M. Moura, "Consensus+ innovations distributed inference over networks: cooperation and sensing in networked systems," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 99–109, 2013.

[9] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact dffusion for distributed optimization and learning – Part I: Algorithm development," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 708 – 723, 2019.

[10] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact dffusion for distributed optimization and learning – Part II: Convergence analysis," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 724 – 739, Feb. 2019.

[11] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[12] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[13] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "A decentralized second-order method with exact linear convergence rate for consensus optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 507–522, 2016.

[14] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *to appear in IEEE Transactions on Signal Processing*, 2019.

[15] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *IEEE Conference on Decision and Control (CDC)*, Osaka, Japan, 2015, pp. 2055–2060.

[16] P. Di Lorenzo and G. Scutari, "Next: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.

[17] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. 176, no. 1-2, pp. 497–544, 2019.

[18] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[19] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2018.

[20] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 315–320, 2018.

[21] S. Pu, W. Shi, J. Xu, and A. Nedić, "A push-pull gradient method for distributed optimization in networks," in *2018 IEEE Conference on Decision and Control (CDC)*, Miami Beach, FL, USA, Dec. 2018, pp. 3385–3390.

[22] S. Pu and A. Nedić, "A distributed stochastic gradient tracking method," in *IEEE Conference on Decision and Control (CDC)*, Miami, FL, Dec. 2018, pp. 963–968.

[23] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D2: Decentralized training over decentralized data," in *Proc. International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 1 – 8.

[24] K. Yuan, S. A. Alghunaim, B. Ying, and A. H. Sayed, "On the influence of bias-correction on distributed stochastic optimization," *arXiv preprint:1903.10956*, 2019.

[25] X. Mao, K. Yuan, Y. Hu, Y. Gu, A. H. Sayed, and W. Yin, "Walkman: A communication-efficient random-walk algorithm for decentralized optimization," *Submitted for publication. Also available at arXiv:1804.06568*, Apr. 2018.

[26] K. Seaman, F. Bach, S. Bubeck, Y.-T. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *Proc. International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017, vol. 70, pp. 3027–3036.