Distributed Zeroth Order Optimization Over Random Networks: A Kiefer-Wolfowitz Stochastic Approximation Approach

Anit Kumar Sahu, Dusan Jakovetic, Dragana Bajovic and Soummya Kar

***Abstract*— We study a standard distributed optimization framework where N networked nodes collaboratively minimize the sum of their local convex costs. The main body of existing work considers the described problem when the underling network is either static or deterministically varying, and the distributed optimization algorithm is of first or second order, i.e., it involves the local costs' gradients and possibly the local Hessians. In this paper, we consider the currently understudied but highly relevant scenarios when: 1) only noisy function values' estimates are available (no gradients nor Hessians can be evaluated); and 2) the underlying network is randomly varying (according to an independent, identically distributed process). For the described random networks-zeroth order optimization setting, we develop a distributed stochastic approximation method of the Kiefer-Wolfowitz type. Furthermore, under standard smoothness and strong convexity assumptions on the local costs, we establish the $O(1/k^{1/2})$ mean square convergence rate for the method – the rate that matches that of the method's centralized counterpart under equivalent conditions.**

1. INTRODUCTION

We consider a commonly studied distributed optimization setting where N nodes are interconnected in a generic, connected network, and they collectively minimize the sum of their local convex costs with respect to a global (vector-valued) variable of common interest. There has been a significant and increasing interest in the described distributed optimization problems, e.g., [1]–[4], which include algorithms which have access to a stochastic first order or a second order oracle. In that, every query made to the oracle gives an unbiased estimate of the gradient or the Hessian based on whether its a first order or second order oracle. Moreover, in the context of distributed optimization, most existing work in the literature consider static or deterministically varying networks.

In this paper, we consider the currently largely understudied, but highly relevant case of 1) zeroth order optimization (only noisy function values, and no gradients nor Hessians are available); and 2) randomly varying networks, more precisely, the networks modeled through a sequence of independent identically distributed (i.i.d.) symmetric Laplacian matrices, such that the network is connected on average. Regarding the former, zeroth order methods become highly desirable when the functions of interest are not given in analytical forms or evaluating the gradient or the Hessian is expensive. For the latter, random network models are more adequate than deterministically varying or static models when the networked nodes communicate through unreliable wireless links, like, e.g., with many emerging internet of things applications, including, for example, maintenance and monitoring of industrial manufacturing systems or large scale industrial plants where communication environments may be harsh.

Our main contributions are as follows. We propose a distributed stochastic approximation method of the Kiefer-Wolfowitz type (see, for example [5]). The method utilizes standard strategy in distributed (sub)gradient-like methods, where the iterations consist of 1)local estimates' weight averaging across nodes' neighborhoods (consensus); and 2) a negative step with respect to the Kiefer-Wolfowitz-type estimates of local functions gradients (innovations). We show that, by a careful design of the consensus and innovations time-varying weights, the distributed Kiefer-Wolfowitz method achieves the $O(1/k^{1/2})$ mean square convergence rate. This rate is achieved for twice continuously differentiable, convex local costs with bounded Hessians, assuming only the availability of noisy function values' estimates, with zero-mean and finite-second moment noises. The achieved $O(1/k^{1/2})$ rate of the distributed method is highest possible and matches that of the counterpart centralized Kiefer-Wolfowitz stochastic approximation method and the minimax rate for the aforementioned class of cost functions (see, [6]).

We now briefly review the literature to help us contrast this paper from prior work. Zeroth order optimization, where the stochastic oracle can be queried for only noisy function values has been applied to scenarios involving black-box based optimization and high-dimensional optimization (see, for example [7], [8]). Various approaches to zeroth order optimization have been adopted such as the classical Kiefer Wolfowitz stochastic approximation [5], [9], simultaneous perturbation stochastic approximation [10] and other random direction random smoothing based variants such as [6], [8], [11]. However, all the aforementioned references solve the minimization problem in a centralized framework or the setting where a single agent has all the data available to it. In the context of random networks, in [12], we consider a distributed stochastic gradient method and establish the method's $O(1/k^2)$ convergence rate. Reference [12] com-

This work is supported by the I-BiDaaS project, funded by the European Commission under Grant Agreement No. 780787. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. The work of DJ is also supported in part by the Serbian Ministry of Education, Science, and Technological Development, grant 174030. The work of AKS and SK was supported in part by National Science Foundation under grant CCF-1513936.

D. Bajovic is with the Faculty of Technical Sciences, University of Novi Sad 21000 Novi Sad, Serbia dbajovic@uns.ac.rs

D. Jakovetic is with the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad 21000 Novi Sad, Serbia djakovet@uns.ac.rs

A. K. Sahu and S. Kar are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 {anits,soummyak}@andrew.cmu.edu

plements the current paper by assuming that nodes have access to gradient estimates. Recently, in [13], a distributed zeroth optimization algorithm was proposed for non-convex minimization with a static graph where a random directions random smoothing approach was employed. In contrast, in this paper we solve a distributed zeroth optimization algorithm for a strongly convex minimization employing a Kiefer-Wolfowitz type stochastic approximation with a random sequence of graphs.

Paper organization.

The next paragraph introduces notation. Section 2 describes the model and the stochastic gradient method we consider. Section 3 states and proves the main result on the algorithm's MSE convergence rate. Section 4 provides a simulation example. Finally, we conclude in Section 5.

Notation.

We denote by $\mathbb{R}$ the set of real numbers and by $\mathbb{R}^m$ the m-dimensional Euclidean real coordinate space. We use normal lower-case letters for scalars, lower case boldface letters for vectors, and upper case boldface letters for matrices. Further, we denote by: $\mathbf{A}_{ij}$ the entry in the i-th row and j-th column of a matrix $\mathbf{A}$; $\mathbf{A}^\top$ the transpose of a matrix A; $\otimes$ the Kronecker product of matrices; I, 0, and $\mathbf{1}$, respectively, the identity matrix, the zero matrix, and the column vector with unit entries; $\mathbf{J}$ the $N \times N$ matrix $J := (1/N)\mathbf{1}\mathbf{1}^\top$. When necessary, we indicate the matrix or vector dimension as a subscript. Next, $A \succ 0\,(A \succeq 0)$ means that the symmetric matrix A is positive definite (respectively, positive semi-definite). We denote by: $\|\cdot\| = \|\cdot\|_2$ the Euclidean (respectively, spectral) norm of its vector (respectively, matrix) argument; $\lambda_i(\cdot)$ the i-th smallest eigenvalue; $\nabla h(w)$ and $\nabla^2 h(w)$ the gradient and Hessian, respectively, evaluated at w of a function $h : \mathbb{R}^m \to \mathbb{R}$, $m \geq 1$; $\mathbb{P}(\mathcal{A})$ and $\mathbb{E}[u]$ the probability of an event $\mathcal{A}$ and expectation of a random variable u, respectively. $\mathbf{e}_j$ denotes the j-th column on the identity matrix $\mathbf{I}$ where the dimension is made clear from the context. Finally, for two positive sequences η_n and χ_n, we have: $\eta_n = O(\chi_n)$ if $\limsup_{n\to\infty} \frac{\eta_n}{\chi_n} < \infty$.

2. Model, Algorithm, and Preliminaries

The network of N agents in our setup collaboratively aim to solve the following unconstrained problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{N} f_i(\mathbf{x}), \tag{1}$$

where $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex function available to node i, $i = 1, ..., N$. We make the following assumption on the functions $f_i(\cdot)$:

Assumption A1. For all $i = 1, ..., N$, function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is twice continuously differentiable with Lipschitz continuous gradients. In particular, there exist constants $L, \mu > 0$ such that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\mu\, \mathbf{I} \preceq \nabla^2 f_i(\mathbf{x}) \preceq L\mathbf{I}.$$

From Assumption A1 we have that each f_i, $i = 1, \cdots, N$, is strongly convex with modulus μ. Using standard properties of convex functions, we have for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2,$$
$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\,\|\mathbf{x} - \mathbf{y}\|.$$

We also have that from assumption A1, the optimization problem in (1) has a unique solution, which we denote by $\mathbf{x}^* \in \mathbb{R}^d$. Throughout the paper, we use the sum function which is defined as $f : \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x})$.

We consider distributed stochastic zeroth order optimization to solve (1) over random networks. Inter-agent communication is modeled by a sequence of independent and identically distributed (i.i.d.) undirected random networks: at each time instant $k = 0, 1, ...$, the underlying inter-agent communication network is denoted by $\mathcal{G}(k) = (V, \mathbf{E}(k))$, with $V = \{1, ..., N\}$ being the set of nodes and $\mathbf{E}(k)$ being the random set of undirected edges. The edge connecting node i and j is denoted as $\{i, j\}$. The time-varying random neighborhood of node i at time k (excluding node i) is represented as $\Omega_i(k) = \{j \in V : \{i, j\} \in \mathbf{E}(k)\}$. The graph Laplacian of the random graph $\mathcal{G}(k)$ at time k is given by $\mathbf{L}(k) \in \mathbb{R}^{N\times N}$, where $\mathbf{L}(k)$ is given by $\mathbf{L}_{ij}(k) = -1$, if $\{i, j\} \in \mathbf{E}(k)$, $i \neq j$; $\mathbf{L}_{ij}(k) = 0$, if $\{i, j\} \notin \mathbf{E}(k)$, $i \neq j$; and $\mathbf{L}_{ii}(k) = -\sum_{j\neq i} \mathbf{L}_{ij}(k)$. It is to be noted that the Laplacian at each time instant is symmetric and a positive semidefinite matrix. As the considered graph sequence is i.i.d., we have that $\mathbb{E}[\,\mathbf{L}(k)\,] = \overline{\mathbf{L}}$. Let the graph corresponding to $\overline{\mathbf{L}}$ be given by $\overline{\mathcal{G}} = (V, \overline{\mathbf{E}})$.

We make the following assumption on $\overline{\mathcal{G}}$.

Assumption A2. The inter-agent communication graph is connected on average, i.e., $\overline{\mathcal{G}}$ is connected. In other words, $\lambda_2(\overline{\mathcal{L}}) > 0$.

A. Distributed Kiefer Wolfowitz type Optimization

We employ a distributed Kiefer Wolfowitz stochastic approximation (KWSA) type method to solve (1). Each node i, $i = 1, ..., N$, in our setup maintains a local copy of its local estimate of the optimizer $\mathbf{x}_i(k) \in \mathbb{R}^d$ at all times. In order to carry out the optimization, each agent i makes queries to a stochastic zeroth order oracle at time k, from which the agent obtains noisy function values of $f_i(\mathbf{x}_i(k))$. Denote the noisy value of $f_i(\cdot)$ as $\widehat{f}_i(\cdot)$ where,

$$\widehat{f}_i(\mathbf{x}_i(k)) = f_i(\mathbf{x}_i(k)) + \widehat{v}_i(k). \tag{2}$$

Due to the unavailability of the analytic form of the functionals, the gradient can not be evaluated and hence, we resort to a gradient approximation. In order to approximate the gradient, each agent makes two calls to the stochastic zeroth order oracle corresponding to each dimension. For instance, for dimension $j \in \{1, \cdots, d\}$ agent i queries for $f_i(\mathbf{x}_i(k) + c_k\mathbf{e}_j)$ and $f_i(\mathbf{x}_i(k) - c_k\mathbf{e}_j)$ at time k and obtains $\widehat{f}_i(\mathbf{x}_i(k) + c_k\mathbf{e}_j)$ and $\widehat{f}_i(\mathbf{x}_i(k) - c_k\mathbf{e}_i)$ respectively, where c_k is a carefully chosen time-decaying potential (to be specified soon). Denote by $\mathbf{g}_i(\mathbf{x}_i(k))$ the approximated gradient, obtained as for each $j \in \{1, \cdots, d\}$:

$$\mathbf{e}_j^\top \mathbf{g}_i(\mathbf{x}_i(k)) = \frac{\widehat{f}_i\left(\mathbf{x}_i(k) + c_k\mathbf{e}_j\right) - \widehat{f}_i\left(\mathbf{x}_i(k) - c_k\mathbf{e}_j\right)}{2c_k}$$
$$\Rightarrow \mathbf{e}_j^\top \mathbf{g}_i(\mathbf{x}_i(k)) = \frac{f_i\left(\mathbf{x}_i(k) + c_k\mathbf{e}_j\right)}{2c_k}$$
$$- \frac{f_i\left(\mathbf{x}_i(k) - c_k\mathbf{e}_j\right)}{2c_k} + \frac{\hat{v}_{i,j}^+(k) - \hat{v}_{i,j}^-(k)}{2c_k}, \tag{3}$$

where $\hat{v}_{i,j}^{+}(k)$ and $\hat{v}_{i,j}^{-}(k)$ denote the measurement noise corresponding to the measurements $\widehat{f}_i\left(\mathbf{x}_i(k)+c_k\mathbf{e}_j\right)$ and $\widehat{f}_i\left(\mathbf{x}_i(k)-c_k\mathbf{e}_j\right)$ respectively. The vectors $\widehat{\mathbf{v}}_i^{+}(k) \in \mathbb{R}^d$ and $\widehat{\mathbf{v}}_i^{-}(k) \in \mathbb{R}^d$ stack all the component wise measurement noise at a node i and are given by $\widehat{\mathbf{v}}_i^{+}(k) = \left[\hat{v}_{i,1}^{+}(k), \cdots, \hat{v}_{i,N}^{+}(k)\right]$ and $\widehat{\mathbf{v}}_i^{-}(k) = \left[\hat{v}_{i,1}^{-}(k), \cdots, \hat{v}_{i,N}^{-}(k)\right]$ respectively. For the rest of the paper, we define $\mathbf{v}_i(k) \doteq \left(\widehat{\mathbf{v}}_i^{+}(k) - \widehat{\mathbf{v}}_i^{-}(k)\right)/2$. Using the mean value theorem, we have,

$$\mathbf{g}_i(\mathbf{x}_i(k)) = \nabla f(\mathbf{x}_i(k)) + c_k\mathbf{P}_i(\mathbf{x}_i(k)) + \frac{\mathbf{v}_i(k)}{c_k}, \quad (4)$$

where

$$\mathbf{e}_j^{\top}\mathbf{P}_i(\mathbf{x}_i(k)) = \frac{\mathbf{e}_j^{\top}\nabla^2 f(\mathbf{x}_i(k)+c_k\alpha_{i,j}^{+}\mathbf{e}_j)\mathbf{e}_j}{2} - \frac{\mathbf{e}_j^{\top}\nabla^2 f(\mathbf{x}_i(k)-c_k\alpha_{i,j}^{-}\mathbf{e}_j)\mathbf{e}_j}{2},$$

where $0 \leq \alpha_{i,j}^{+}, \alpha_{i,j}^{-} \leq 1$. Finally, for arbitrary deterministic initializations $\mathbf{x}_i(0) \in \mathbb{R}^d$, $i = 1, ..., N$, the optimizer update rule at node i and $k = 0, 1, ...,$ is given as follows:

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) - \beta_k \sum_{j\in\Omega_i(k)} \left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right) - \alpha_k\mathbf{g}_i(\mathbf{x}_i(k)). \quad (5)$$

It is to be noted that unlike first order stochastic gradient methods, where the algorithm has access to unbiased estimates of the gradient. The local gradient estimates $\mathbf{g}_i(\cdot)$ used in (5) are biased (see (4)) due to the unavailability of the exact gradient functions and their approximations using the zeroth order scheme in (3). The update is carried on in all agents parallely in a synchronous fashion. The weight sequences $\{\alpha_k\}$, $\{c_k\}$ and $\{\beta_k\}$ are given by $\alpha_k = \alpha_0/(k+1)$, $c_k = c_0/(k+1)^{\delta}$ and $\beta_k = \beta_0/(k+1)^{\tau}$ respectively, where $\alpha_0, c_0, \beta_0 > 0$. We state an assumption on the weight sequences before proceeding further.

Assumption A3. The constants α_0, $\delta > 0$ and $\tau \in (0, 1)$ are chosen such that,

$$\sum_{k=1}^{\infty} \frac{\alpha_k^2}{c_k^2} < \infty. \quad (6)$$

Denote by $\mathbf{x}(k) = \left[\mathbf{x}_1^{\top}(k), \cdots, \mathbf{x}_N^{\top}(k)\right]^{\top} \in \mathbb{R}^{Nd}$, $\mathbf{P}(\mathbf{x}(k)) = \left[\mathbf{P}_1^{\top}\left(\mathbf{x}_1(k)\right), \cdots, \mathbf{P}_N^{\top}\left(\mathbf{x}_N(k)\right)\right]^{\top} \in \mathbb{R}^{Nd}$ the vectors that stacks the local optimizers and the gradient bias terms (see (4))of all nodes. Also, define function $F : \mathbb{R}^{Nd} \mapsto \mathbb{R}$, by $F(\mathbf{x}) = \sum_{i=1}^{N} f_i(\mathbf{x}_i)$, with $\mathbf{x} = \left[\mathbf{x}_1^{\top}, \cdots, \mathbf{x}_N^{\top}\right]^{\top} \in \mathbb{R}^{Nd}$. Finally, let $\mathbf{W}_k = (\mathbf{I} - \mathbf{L}_k)\otimes\mathbf{I}_d$, where $\mathbf{L}_k = \beta_k\,\mathbf{L}(k)$. Then the update in (5) can be written as:

$$\mathbf{x}(k+1) = \mathbf{W}_k\mathbf{x}(k) - \alpha_k\left(\nabla F(\mathbf{x}(k)) + c_k\mathbf{P}(\mathbf{x}(k)) + \frac{\mathbf{v}(k)}{c_k}\right). \quad (7)$$

Let $\mathcal{F}_k$ denote the history of the proposed algorithm up to time k. Given that the sources of randomness in our algorithm are the noise sequence $\{\mathbf{v}(k)\}$ and the random network sequence $\{\mathbf{L}_k\}$, $\mathcal{F}_k$ is given by the σ-algebra generated by the collection of random variables $\{\mathbf{L}(s), \mathbf{v}_i(s)\}$, $i = 1, ..., N$, $s = 0, ..., k-1$.

Assumption A4. For each $i = 1, ..., N$, the sequence of measurement noises $\{\widehat{\mathbf{v}}_i(k)\}$ satisfies for all $k = 0, 1, ...$:

$$\mathbb{E}[\,\widehat{\mathbf{v}}_i(k)\,|\,\mathcal{F}_k\,] = 0, \text{ almost surely (a.s.)}$$
$$\mathbb{E}[\,\|\widehat{\mathbf{v}}_i(k)\|^2\,|\,\mathcal{F}_k\,] \leq c_f\|\mathbf{x}_i(k)\|^2 + \sigma^2, \text{ a.s.}, \quad (8)$$

where c_f and σ^2 are nonnegative constants.

It is to be noted that assumption A4 is trivially satisfied, when $\{\mathbf{v}_i(k)\}$ is an i.i.d. zero-mean, finite second moment, noise sequence such that $\mathbf{v}_i(k)$ is also independent of the history $\mathcal{F}_k$. However, the assumption allows the noise to be dependent on the current iterate at all times.

3. Performance Analysis

A. Main Result and Auxiliary Lemmas

We state the main result concerning the mean square error at each agent i next.

Theorem 3.1. *1) Consider the optimizer estimate sequence $\{\mathbf{x}(k)\}$ generated by the algorithm* (5)*. Let assumptions A1-A4 hold. Then, for each node i's optimizer estimate $\mathbf{x}_i(k)$ and the solution $\mathbf{x}^{\star}$ of problem* (1)*, $\forall k \geq k_2$ there holds:*

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq 2R_k + \frac{64N\Delta_{1,\infty}\alpha_0^2}{\mu c_0^2 p_{\mathcal{L}}^2\beta_0^2(k+1)^{2-2\tau-2\delta}} \frac{4(L-\mu)^2N^2dc_0^2}{\mu(k+1)^{2\delta}} + \frac{8\Delta_{1,\infty}\alpha_0^2}{p_{\mathcal{L}}^2\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}} + \frac{4N\alpha_0\left(c_f q_{\infty}(N,d,\alpha_0,c_0) + N\sigma_1^2\right)}{c_0^2\mu(k+1)^{1-2\delta}}, \quad (9)$$

where, $k_2 = \max\left\{k_0, (4|E|\rho_0^2)^{1/\epsilon} - 1\right\}$, $\Delta_{1,\infty} = 6c_f q_{\infty}(N,d,\alpha_0,c_0) + 6N\sigma_1^2$ *and* $q_{\infty}(N,d,\alpha_0,c_0) = \mathbb{E}\left[\|\mathbf{x}(k_0) - \mathbf{x}^o\|^2\right] + \frac{\sqrt{Nd}(L-\mu)\alpha_0c_0}{\delta} + \frac{Nd(L-\mu)^2\alpha_0^2c_0^2}{1+2\delta} + \frac{\alpha_0^2\left(2c_fN\|\mathbf{x}^o\|^2+N\sigma^2\right)}{c_0^2(1-2\delta)} + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} + \frac{2\alpha_0^2c_0\sqrt{Nd}(L-\mu)\|\nabla F(\mathbf{x}^o)\|}{1+\delta}$*. In the latter k_0 is given by* $k_0 = \inf\left\{k : \frac{\mu}{2} > (L-\mu)\sqrt{Nd}c_k + \frac{2c_f\alpha_k}{c_k^2}\right\}$*. R_k is a term which decay faster than the rest of the terms.*
2) In particular, the rate of decay of the RHS of (9) *is given by $(k+1)^{-\delta_1}$, where $\delta_1 = \min\{1-2\delta, 2-2\tau-2\delta, 2\delta\}$. By, optimizing over τ and δ, we obtain that for $\tau = 1/2$ and $\delta = 1/4$ and hence,*

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq 2R_k + \frac{64N\Delta_{1,\infty}\alpha_0^2}{\mu c_0^2 p_{\mathcal{L}}^2\beta_0^2(k+1)^{0.5}} \frac{4(L-\mu)^2N^2dc_0^2}{\mu(k+1)^{0.5}} + \frac{8\Delta_{1,\infty}\alpha_0^2}{p_{\mathcal{L}}^2\beta_0^2c_0^2(k+1)^{0.5}} + \frac{4N\alpha_0\left(c_f q_{\infty}(N,d,\alpha_0,c_0) + N\sigma_1^2\right)}{c_0^2\mu(k+1)^{0.5}} = O\left(\frac{1}{k^{\frac{1}{2}}}\right), \ \forall i.$$

Theorem 3.1 establishes the $O(1/k^{1/2})$ MSE rate of convergence of the algorithm (5); due to the assumed f_i's strong convexity, the theorem also implies that $\mathbb{E}\left[f(\mathbf{x}_i(k)) - f(\mathbf{x}^{\star})\right] = O(1/k^{1/2})$. Note that the expectation in Theorem 3.1 is both with respect to randomness in gradient noises and with respect to the randomness in the

underlying network. The $O(1/k^{1/2})$ rate is independent of the statistics of the underlying random network, as long as the network is connected on average.

From (9), it might seem that the dependence of the upper bound is linear in terms of d. However, on tuning the constants $\alpha_0 \asymp d^{-1/5}$, $\beta_0 \asymp d^{-1/10}$ and $c_0 \asymp d^{-3/10}$, the dependence of $\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right]$ can be reduced to $d^{2/5}$. It is to be noted that the upper bound derived in (9) matches with that of the minimax bound for (centralized) zeroth order optimization with twice continuously differentiable cost functions as derived in [6]. The sublinear rate of convergence of zeroth order optimization algorithms in the context of KWSA can be attributed to the biased gradients. For better finite time convergence rates, bias-reduction techniques such as the "twicing trick" and finite difference interpolation techniques can be used.

Proof strategy and auxiliary lemmas. Establishing the main result in Theorem 3.1 involves three crucial steps which are outlined in the subsections 3-B, 3-C and 3-D. Subsection 3-B concerns with the mean square boundedness of the iterates $\mathbf{x}_i(k)$, which also implies the mean square boundedness of the gradients $\nabla f_i(\mathbf{x}_i(k))$. In subsection 3-C, the mean square error of the disagreements of a node's optimizer estimate with respect to the network averaged optimizer estimate ,i.e., $\overline{\mathbf{x}}(k) := \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i(k)$, is characterized in terms of k and the algorithm parameters. Finally, subsection 3-D characterizes the optimality gap of the networked average optimizer estimate sequence with respect to the optimizer of (1) and on combining the result from subsection 3-C, the result follows.

B. Mean square boundedness of the iterate sequence

This subsection shows the mean square boundedness of the algorithm iterates.

Lemma 3.2. *Let the hypotheses of Theorem 3.1 hold. In addition assume that, $\|\nabla F(\mathbf{1}_N \otimes \mathbf{x}^*)\|$ is bounded. Then, we have,*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}(k) - \mathbf{x}^o\|^2\right] \leq q_{k_0}(N, d, \alpha_0, c_0) \\
&+ \frac{\sqrt{Nd}(L-\mu)\alpha_0 c_0}{\delta} + \frac{Nd(L-\mu)^2\alpha_0^2 c_0^2}{1+2\delta} \\
&+ \frac{\alpha_0^2\left(2c_f N\|\mathbf{x}^o\|^2 + N\sigma^2\right)}{c_0^2(1-2\delta)} + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} \\
&\doteq q_\infty(N, d, \alpha_0, c_0),
\end{aligned}$$

where $\mathbb{E}\left[\|\mathbf{x}(k_0) - \mathbf{x}^o\|^2\right] \leq q_{k_0}(N, d, \alpha_0, c_0)$ *and* $k_0 = \inf\left\{k \,:\, \frac{\mu}{2} > (L-\mu)\sqrt{Nd}c_k + \frac{2c_f\alpha_k}{c_k^2}\right\}$.

Proof.

$$\begin{aligned}
\mathbf{x}(k+1) &= \mathbf{W}_k\mathbf{x}(k) \\
&- \frac{\alpha_k}{c_k}\left(c_k\nabla F(\mathbf{x}(k)) + c_k^2\mathbf{P}(\mathbf{x}(k)) + \mathbf{v}(k)\right). \qquad (10)
\end{aligned}$$

Denote $\mathbf{x}^o = \mathbf{1}_N \otimes x^*$. Then, we have,

$$\begin{aligned}
\mathbf{x}(k+1) - \mathbf{x}^o &= \mathbf{W}_k(\mathbf{x}(k) - \mathbf{x}^o) \\
&- \alpha_k\left(\nabla F(\mathbf{x}(k)) - \nabla F(\mathbf{x}^o)\right) \\
&- \frac{\alpha_k}{c_k}\mathbf{v}(k) - \alpha_k\nabla F(\mathbf{x}^o) - \alpha_k c_k\mathbf{P}(\mathbf{x}(k)). \qquad (11)
\end{aligned}$$

By Leibnitz rule, we have,

$$\begin{aligned}
&\nabla F(\mathbf{x}(k)) - \nabla F(\mathbf{x}^o) \\
&= \left[\int_{s=0}^1 \nabla^2 F\left(\mathbf{x}^o + s(\mathbf{x}(k) - \mathbf{x}^o)\right) ds\right](\mathbf{x}(k) - \mathbf{x}^o) \\
&= \mathbf{H}_k\left(\mathbf{x}(k) - \mathbf{x}^o\right). \qquad (12)
\end{aligned}$$

By Lipschitz continuity of the gradients and strong convexity of $f(\cdot)$, we have that $L\mathbf{I} \succcurlyeq \mathbf{H}_k \succcurlyeq \mu\mathbf{I}$. Denote by $\boldsymbol{\zeta}(k) = \mathbf{x}(k) - \mathbf{x}^o$ and by $\boldsymbol{\xi}(k) = (\mathbf{W}_k - \alpha_k\mathbf{H}_k)(\mathbf{x}(k) - \mathbf{x}^o) - \alpha_k\nabla F(\mathbf{x}^o)$. Then, there holds:

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2\,|\,\mathcal{F}_k\,] \leq \mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2|\,\mathcal{F}_k\,\right] \\
&- 2\alpha_k\mathbb{E}\left[\boldsymbol{\xi}(k)^\top|\,\mathcal{F}_k\right]\mathbb{E}[\,\mathbf{v}(k)\,|\,\mathcal{F}_k\,] + \alpha_k^2\,\mathbb{E}[\,\|\mathbf{v}(k)\|^2\,|\,\mathcal{F}_k\,] \\
&+ \alpha_k^2c_k^2\mathbf{P}^\top(\mathbf{x}(k))\mathbf{P}(\mathbf{x}(k)) - 2\alpha_k c_k\mathbf{P}^\top(\mathbf{x}(k))\mathbb{E}\left[\boldsymbol{\xi}(k)|\mathcal{F}_k\right] \\
&+ \mathbf{P}\left(\mathbf{x}(k)\right)^\top\mathbb{E}\left[\mathbf{v}(k)|\mathcal{F}_k\right]. \qquad (13)
\end{aligned}$$

We use the following inequalities:

$$\begin{aligned}
&- 2\alpha_k c_k\mathbf{P}^\top(\mathbf{x}(k))\left(\mathbf{I} - \beta_k\overline{\mathbf{L}} - \alpha_k\mathbf{H}_k\right)(\mathbf{x}(k) - \mathbf{x}^o) \\
&\leq 2\alpha_k c_k\,\|\mathbf{P}(\mathbf{x}(k))\|\,\left\|\mathbf{I} - \beta_k\overline{\mathbf{L}} - \alpha_k\mathbf{H}_k\right\|\,\|\mathbf{x}(k) - \mathbf{x}^o\| \\
&\leq \sqrt{Nd}(L-\mu)\alpha_k c_k\,(1-\mu\alpha_k)\left(1 + \|\mathbf{x}(k) - \mathbf{x}^o\|^2\right) \\
&\leq \sqrt{Nd}(L-\mu)\alpha_k c_k + \sqrt{Nd}(L-\mu)\alpha_k c_k\,\|\mathbf{x}(k) - \mathbf{x}^o\|^2, \qquad (14)
\end{aligned}$$

$$\alpha_k^2c_k^2\mathbf{P}^\top(\mathbf{x}(k))\mathbf{P}(\mathbf{x}(k)) \leq Nd\,(L-\mu)^2\,\alpha_k^2c_k^2, \qquad (15)$$

and

$$\begin{aligned}
&\frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\|\mathbf{v}(k)\|^2\,|\mathcal{F}_k\right] \leq \frac{\alpha_k^2}{c_k^2}c_f N\,\|\mathbf{x}(k)\|^2 + \frac{\alpha_k^2}{c_k^2}N\sigma^2 \\
&\leq 2\frac{\alpha_k^2}{c_k^2}c_f\,\|\mathbf{x}(k) - \mathbf{x}^o\|^2 + \frac{\alpha_k^2}{c_k^2}\left(2c_f\,\|\mathbf{x}^o\|^2 + N\sigma^2\right). \qquad (16)
\end{aligned}$$

Then from (13), we have,

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2\,|\,\mathcal{F}_k\,] \leq \mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2|\,\mathcal{F}_k\,\right] \\
&+ \sqrt{Nd}(L-\mu)\alpha_k c_k\|\boldsymbol{\zeta}(k)\|^2 + 2\frac{\alpha_k^2}{c_k^2}c_f\|\boldsymbol{\zeta}(k)\|^2 \\
&+ \frac{\alpha_k^2}{c_k^2}\left(2c_f\,\|\mathbf{x}^o\|^2 + N\sigma^2\right) + \sqrt{Nd}(L-\mu)\alpha_k c_k \\
&+ Nd\,(L-\mu)^2\,\alpha_k^2c_k^2 + 2\alpha_k^2c_k\sqrt{Nd}(L-\mu)\,\|\nabla F(\mathbf{x}^o)\|. \qquad (17)
\end{aligned}$$

We next bound $\mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2|\,\mathcal{F}_k\,\right]$. Note that $\|\mathbf{W}_k - \alpha_k\,\boldsymbol{H}_k\| \leq 1 - \mu\,\alpha_k$. Therefore, we have:

$$\|\boldsymbol{\xi}(k)\| \leq (1 - \mu\,\alpha_k)\,\|\boldsymbol{\zeta}(k)\| + \alpha_k\,\|\nabla F(\mathbf{x}^o)\|. \qquad (18)$$

We now use the following inequality:

$$(a+b)^2 \leq (1+\theta)\,a^2 + \left(1 + \frac{1}{\theta}\right)b^2, \qquad (19)$$

for any $a, b \in \mathbb{R}$ and $\theta > 0$. We set $\theta = \mu\alpha_k$. Using the inequality (19) in (18), we have:

$$\begin{aligned}
\|\boldsymbol{\xi}(k)\|^2 &\leq (1+\mu\alpha_k)\,(1-\alpha_k\mu)^2\,\|\boldsymbol{\zeta}(k)\|^2 \\
&+ \left(1 + \frac{1}{\mu\alpha_k}\right)\alpha_k^2\|\nabla F(\mathbf{x}^o)\|^2
\end{aligned}$$

$$\leq (1-\alpha_k\mu)\left\|\boldsymbol{\zeta}(k)\right\|^2 + 2\frac{\alpha_k}{\mu}\|\nabla F(\mathbf{x}^o)\|^2. \quad (20)$$

Using (20) in (17), we have,

$$\begin{aligned}
&\mathbb{E}[\left\|\boldsymbol{\zeta}(k+1)\right\|^2 | \mathcal{F}_k] \leq 2\frac{\alpha_k}{\mu}\|\nabla F(\mathbf{x}^o)\|^2 + \left\|\boldsymbol{\zeta}(k)\right\|^2 \\
&\times \left(1-\alpha_k\mu+\sqrt{Nd}(L-\mu)\alpha_k c_k + 2\frac{\alpha_k^2}{c_k^2}c_f\right) \\
&+ \frac{\alpha_k^2}{c_k^2}\left(2c_f\left\|\mathbf{x}^o\right\|^2 + N\sigma^2\right) + \sqrt{Nd}(L-\mu)\alpha_k c_k \\
&+ Nd\left(L-\mu\right)^2\alpha_k^2 c_k^2 + 2\alpha_k^2 c_k\sqrt{Nd}(L-\mu)\left\|\nabla F(\mathbf{x}^o)\right\|. \quad (21)
\end{aligned}$$

Define k_0 as follows:

$$k_0 = \inf\left\{k \ : \ \frac{\mu}{2} > (L-\mu)\sqrt{Nd}c_k + \frac{2c_f\alpha_k}{c_k^2}\right\}.$$

It is to be noted that k_0 is necessarily finite as $c_k \to 0$ and $\alpha_k c_k^{-2} \to 0$ as $k \to \infty$.

Proposition 3.3. *Let the hypotheses of Theorem 3.1 hold. Then, we have $\forall k \geq k_0$,*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{\zeta}(k+1)\right\|^2\right] \leq q_{k_0}(N,d,\alpha_0,c_0) + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} \\
&+ \frac{\sqrt{Nd}(L-\mu)\alpha_0 c_0}{\delta} + \frac{Nd(L-\mu)^2\alpha_0^2 c_0^2}{1+2\delta} \\
&+ \frac{\alpha_0^2\left(2c_f N\left\|\mathbf{x}^o\right\|^2 + N\sigma^2\right)}{c_0^2(1-2\delta)} + \frac{2\sqrt{Nd}(L-\mu)\left\|\nabla F(\mathbf{x}^o)\right\|}{1+\delta} \\
&\doteq q_\infty(N,d,\alpha_0,c_0)
\end{aligned}$$

From proposition 3.3, we have that $\mathbb{E}\left[\left\|\mathbf{x}(k+1)-\mathbf{x}^o\right\|^2\right]$ is finite and bounded from above, where $\mathbb{E}\left[\left\|\mathbf{x}(k_0)-\mathbf{x}^o\right\|^2\right] \leq q_{k_0}(N,d,\alpha_0,c_0)$. From the boundedness of $\mathbb{E}\left[\left\|\mathbf{x}(k)-\mathbf{x}^o\right\|^2\right]$, we have also established the boundedness of $\mathbb{E}\left[\left\|\nabla F(\mathbf{x}(k))\right\|^2\right]$ and $\mathbb{E}\left[\left\|\mathbf{x}(k)\right\|^2\right]$. □

With the above development in place, we can bound the variance of the noise process $\{\mathbf{v}(k)\}$ as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{v}(k)\right\|^2 | \mathcal{F}_k\right] \leq 0.5\mathbb{E}\left[\left\|\widehat{\mathbf{v}}^+(k)\right\|^2 | \mathcal{F}_k\right] \\
&+ 0.5\mathbb{E}\left[\left\|\widehat{\mathbf{v}}^-(k)\right\|^2 | \mathcal{F}_k\right] \\
&\leq 2c_f q_\infty(N,d,\alpha_0,c_0) + 2N\underbrace{\left(\sigma^2 + \left\|\mathbf{x}^*\right\|^2\right)}_{\sigma_1^2}. \quad (22)
\end{aligned}$$

C. Disagreement Bounds

We now study the disagreement of the optimizer sequence $\{\mathbf{x}_i(k)\}$ at a node i with respect to the (hypothetically available) network averaged optimizer sequence, i.e., $\overline{\mathbf{x}}(k) = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i(k)$. Define the disagreement at the i-th node as $\widetilde{\mathbf{x}}_i(k) = \mathbf{x}_i(k) - \overline{\mathbf{x}}(k)$. The vectorized version of the disagreements $\widetilde{\mathbf{x}}_i(k),\ i = 1, \cdots, N$, can then be written as $\widetilde{\mathbf{x}}(k) = (\mathbf{I}-\mathbf{J})\mathbf{x}(k)$, where $\mathbf{J} = \frac{1}{N}\left(\mathbf{1}_N \otimes \mathbf{I}_d\right)\left(\mathbf{1}_N \otimes \mathbf{I}_d\right)^\top = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \otimes \mathbf{I}_d$. We have the following Lemma:

Lemma 3.4. *Let the hypotheses of Theorem 3.1 hold. Then, we have $\forall k \geq k_1$*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(k+1)\right\|^2\right] \leq Q_k + \frac{4\Delta_{1,\infty}\alpha_0^2}{p_{\mathcal{L}}^2\beta_0^2 c_0^2(k+1)^{2-2\tau-2\delta}} \\
&= O\left(\frac{1}{k^{2-2\delta-2\tau}}\right),
\end{aligned}$$

where Q_k is a term which decays faster than $(k+1)^{-2+2\tau+2\delta}$ and $k_1 = (4|E|\rho_0^2)^{1/\epsilon} - 1$.

As detailed in the next Subsection, Lemma 3.4 plays a crucial role in providing a tight bound for the bias in the gradient estimates according to which the global average $\overline{\mathbf{x}}(k)$ evolves.

Proof. The process $\{\widetilde{\mathbf{x}}(k)\}$ follows the recursion:

$$\begin{aligned}
&\widetilde{\mathbf{x}}(k+1) = \widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k) \\
&- \frac{\alpha_k}{c_k}\left(\mathbf{I}-\mathbf{J}\right)\underbrace{\left(c_k\nabla F(\mathbf{x}(k)) + \mathbf{v}(k) + c_k^2\mathbf{P}\left(\mathbf{x}(k)\right)\right)}_{\mathbf{w}(k)}, \quad (23)
\end{aligned}$$

where $\widetilde{\mathbf{W}}_k = \mathbf{W}_k - \mathbf{J} = (\mathbf{I}-\mathbf{L}_k)\otimes\mathbf{I}_d - \mathbf{J}$. Using (19) in (23), we have,

$$\begin{aligned}
\left\|\widetilde{\mathbf{x}}(k+1)\right\|^2 &\leq (1+\theta_k)\left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\|^2 \\
&+ \left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\left\|\widetilde{\mathbf{w}}(k)\right\|^2. \quad (24)
\end{aligned}$$

We, now bound the term $\mathbb{E}\left[\left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\|^2 | \mathcal{F}_k\right]$.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\mathbf{W}}(k)\widetilde{\mathbf{x}}(k)\right\|^2 | \mathcal{F}_k\right] = \widetilde{\mathbf{x}}^\top(k)\mathbb{E}\left[\widetilde{\mathbf{W}}^2(k) - \mathbf{J} | \mathcal{F}_k\right]\widetilde{\mathbf{x}}(k) \\
&= \widetilde{\mathbf{x}}^\top(k)\left(\mathbf{I} - 2\beta_k\overline{\mathbf{L}} + \beta_k^2\overline{\mathbf{L}}^2 + \widetilde{\mathbf{L}}(k)^2 - \mathbf{J}\right)\widetilde{\mathbf{x}}(k) \\
&\leq \left(1 - 2\beta_k\lambda_2\left(\overline{\mathbf{L}}\right) + \beta_k^2\lambda_N^2\left(\overline{\mathbf{L}}\right)\right. \\
&\left. + \frac{4|E|\beta_0\rho_0^2}{(k+1)^{1/2+\epsilon}} - 4\beta_k^2|E|\right)\left\|\widetilde{\mathbf{x}}(k)\right\|^2 \\
&\leq \left(1 - 2\beta_k\lambda_2\left(\overline{\mathbf{L}}\right) + \frac{4|E|\beta_0\rho_0^2}{(k+1)^{1/2+\epsilon}}\right)\left\|\widetilde{\mathbf{x}}(k)\right\|^2 \\
&\leq \left(1 - \beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\left\|\widetilde{\mathbf{x}}(k)\right\|^2, \quad (25)
\end{aligned}$$

where the last inequality holds for $k \geq (4|E|\rho_0^2)^{1/\epsilon} - 1 \doteq k_1$. Then, we have, $\forall k \geq k_1$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(k+1)\right\|^2 | \mathcal{F}_k\right] \leq (1+\theta_k)\left(1 - \beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\left\|\widetilde{\mathbf{x}}(k)\right\|^2 \\
&+ \left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\left\|\mathbf{w}(k)\right\|^2 | \mathcal{F}_k\right], \quad (26)
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{w}(k)\right\|^2 | \mathcal{F}_k\right] \leq 3c_k^2\left\|\nabla F(\mathbf{x}(k))\right\|^2 + 3\mathbb{E}\left[\left\|\mathbf{v}(k)\right\|^2 | \mathcal{F}_k\right] \\
&+ 3c_k^2\left\|P\left(\mathbf{x}(k)\right)\right\|^2 \\
&\leq 3c_k^2\left\|\nabla F(\mathbf{x}(k))\right\|^2 + 3c_k^2 Nd(L-\mu)^2 \\
&+ 6c_f q_\infty(N,d,\alpha_0,c_0) + 6N\sigma_1^2 \\
&\Rightarrow \mathbb{E}\left[\left\|\mathbf{w}(k)\right\|^2\right] \leq 3\left(2c_f + c_k^2 L^2\right)q_\infty(N,d,\alpha_0,c_0) \\
&+ 3c_k^2 Nd(L-\mu)^2 + 6N\sigma_1^2
\end{aligned}$$

$$= \underbrace{6c_f q_\infty(N,d,\alpha_0,c_0) + 6N\sigma_1^2}_{\Delta_{1,\infty}}$$
$$+ \underbrace{3c_k^2 Nd(L-\mu)^2 + 3c_k^2 L^2 q_\infty(N,d,\alpha_0,c_0)}_{c_k^2\Delta_{2,\infty}} \doteq \Delta_k$$
$$\Rightarrow \mathbb{E}\left[\|\mathbf{w}(k)\|^2\right] < \infty. \quad (27)$$

With the above development in place, we then have,

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] \leq (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{L}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2 + \left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\Delta_k. \quad (28)$$

In particular, we choose $\theta(k) = \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right)$.

Proposition 3.5. *Let the hypotheses of Theorem 3.1 hold. Then, we have $k \geq k_2 = \max\{k_0, k_1\}$ where $k_1 = (4|E|\rho_0^2)^{1/\epsilon} - 1$,*

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] \leq Q_k + \frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2 c_0^2 (k+1)^{2-2\tau-2\delta}}.$$

Hence, we have the disagreement given by,

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] = O\left(\frac{1}{k^{2-2\delta-2\tau}}\right).$$

□

D. Proof of Theorem 3.1

In this subsection, we complete the proof of Theorem 3.1 by characterizing the optimality gap of the network averaged optimizer estimate sequence and then combining it with the result obtained in Lemma 3.4.

Denote $\overline{\mathbf{x}}(k) = \frac{1}{N}\sum_{n=1}\mathbf{x}_i(k)$. From (23), we have,

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{c_k}\left[\frac{c_k}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i(k)) + \underbrace{\frac{c_k^2}{N}\sum_{i=1}^N \mathbf{P}_i(\mathbf{x}_i(k))}_{\overline{\mathbf{P}}(\mathbf{x}(k))} + \underbrace{\frac{1}{N}\sum_{i=1}^N \mathbf{v}_i(k)}_{\overline{\mathbf{v}}(k)}\right]$$
$$\Rightarrow \overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{Nc_k}\left[\sum_{i=1}^N \nabla f_i(\mathbf{x}_i(k)) - \nabla f_i(\overline{\mathbf{x}}(k)) + \nabla f_i(\overline{\mathbf{x}}(k))\right] - \frac{\alpha_k}{c_k}\left(\overline{\mathbf{v}}(k) + \overline{\mathbf{P}}(\mathbf{x}(k))\right). \quad (29)$$

Recall that $f(\cdot) = \sum_{i=1}^N f_i(\cdot)$. Then, we have,

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\nabla f(\overline{\mathbf{x}}(k)) - \frac{\alpha_k}{N}\left[\sum_{i=1}^N \nabla f_i(\mathbf{x}_i(k)) - \nabla f_i(\overline{\mathbf{x}}(k))\right] - \frac{\alpha_k}{c_k}\left(\overline{\mathbf{v}}(k) + \overline{\mathbf{P}}(\mathbf{x}(k))\right)$$
$$\Rightarrow \overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{Nc_k}\left[c_k\nabla f(\overline{\mathbf{x}}(k)) + \mathbf{e}(k)\right], \quad (30)$$

where

$$\mathbf{e}(k) = N\overline{\mathbf{v}}(k) + \underbrace{N\overline{P}(\mathbf{x}(k)) + c_k\sum_{i=1}^N\left(\nabla f_i(\mathbf{x}_i(k)) - \nabla f_i(\overline{\mathbf{x}}(k))\right)}_{\boldsymbol{\epsilon}(k)}. \quad (31)$$

Note that, $c_k\|\nabla f_i(\mathbf{x}_i(k)) - \nabla f_i(\overline{\mathbf{x}}(k))\| \leq c_k L\|\mathbf{x}_i(k) - \overline{\mathbf{x}}(k)\| = c_k L\|\widetilde{\mathbf{x}}_i(k)\|$. We also have that, $\left\|\overline{\mathbf{P}}(\mathbf{x}(k))\right\| \leq (L-\mu)\sqrt{d}c_k^2$. Thus, we can conclude that, $\forall k \geq k_2$

$$\boldsymbol{\epsilon}(k) = c_k\sum_{i=1}^N\left(\nabla f_i(\mathbf{x}_i(k)) - \nabla f_i(\overline{\mathbf{x}}(k))\right) + N\overline{P}(\mathbf{x}(k))$$
$$\Rightarrow \|\boldsymbol{\epsilon}(k)\|^2 \leq 2NL^2c_k^2\|\widetilde{\mathbf{x}}(k)\|^2 + 2(L-\mu)^2N^2dc_k^4$$
$$\Rightarrow \mathbb{E}\left[\|\boldsymbol{\epsilon}(k)\|^2\right] \leq \frac{8NL^2\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau}} + \frac{2(L-\mu)^2N^2dc_0^4}{(k+1)^{4\delta}} + \frac{4NL^2Q_kc_0^2}{(k+1)^{2\delta}}. \quad (32)$$

With the above development in place, we rewrite (30) as follows:

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\nabla f(\overline{\mathbf{x}}(k)) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{v}}(k)$$
$$\Rightarrow \overline{\mathbf{x}}(k+1) - \mathbf{x}^* = \overline{\mathbf{x}}(k) - \mathbf{x}^* - \frac{\alpha_k}{N}\left[\nabla f(\overline{\mathbf{x}}(k)) - \underbrace{\nabla f(\mathbf{x}^*)}_{=0}\right] - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{v}}(k). \quad (33)$$

By Leibnitz rule, we have,

$$\nabla f(\overline{\mathbf{x}}(k)) - \nabla f(\mathbf{x}^*) = \underbrace{\left[\int_{s=0}^1 \nabla^2 f(\mathbf{x}^* + s(\overline{\mathbf{x}}(k) - \mathbf{x}^*))\,ds\right]}_{\overline{\mathbf{H}}_k}(\overline{\mathbf{x}}(k) - \mathbf{x}^*), \quad (34)$$

where it is to be noted that $NL \succcurlyeq \overline{\mathbf{H}}_k \succcurlyeq N\mu$. Using (34) in (33), we have,

$$(\overline{\mathbf{x}}(k+1) - \mathbf{x}^*) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right](\overline{\mathbf{x}}(k) - \mathbf{x}^*) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{v}}(k). \quad (35)$$

Denote by $\mathbf{m}(k) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right](\overline{\mathbf{x}}(k) - \mathbf{x}^*) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k)$ and note that $\mathbf{m}(k)$ is conditionally independent from $\overline{\mathbf{v}}(k)$ given the history $\mathcal{F}_k$. Then (35) can be rewritten as:

$$(\overline{\mathbf{x}}(k+1) - \mathbf{x}^*) = \mathbf{m}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{v}}(k)$$
$$\Rightarrow \|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\|^2 \leq \|\mathbf{m}(k)\|^2 - 2\frac{\alpha_k}{c_k}\mathbf{m}(k)^\top\overline{\mathbf{v}}(k) + \frac{\alpha_k^2}{c_k^2}\|\overline{\mathbf{v}}(k)\|^2. \quad (36)$$

Using the properties of conditional expectation and noting that $\mathbb{E}[\mathbf{v}(k)|\mathcal{F}_k] = \mathbf{0}$, we have,

$$\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\|^2|\mathcal{F}_k\right] \leq \|\mathbf{m}(k)\|^2 + \frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\|v(k)\|^2|\mathcal{F}_k\right]$$

$$\Rightarrow \mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right] \leq \mathbb{E}\left[\|\mathbf{m}(k)\|^2\right] + \frac{2\alpha_k^2\left(c_f q_\infty(N,d,\alpha_0,c_0)+N\sigma_1^2\right)}{c_k^2}. \quad (37)$$

Using (19), we have for $\mathbf{m}(k)$,

$$\begin{aligned}\|\mathbf{m}(k)\|^2 &\leq (1+\theta_k)\left\|\mathbf{I}-\frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right\|^2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2 \\ &+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2c_k^2}\|\boldsymbol{\epsilon}(k)\|^2 \\ &\leq (1+\theta_k)\left(1-\frac{\mu\alpha_0}{k+1}\right)^2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2 \\ &+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2c_k^2}\|\boldsymbol{\epsilon}(k)\|^2. \quad (38)\end{aligned}$$

On choosing $\theta_k=\frac{\mu\alpha_0}{k+1}$, we have for all $k\geq k_2$,

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{m}(k)\|^2\right] \leq \left(1-\frac{\mu\alpha_0}{N(k+1)}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right] \\ &+\frac{16L^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}+\frac{4(L-\mu)^2Nd\alpha_0c_0^2}{\mu(k+1)^{1+2\delta}} \\ &+\frac{8Q_kc_0^2}{\mu(k+1)^{2\delta}} \\ &\Rightarrow \mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right] \leq \left(1-\frac{\mu\alpha_0}{N(k+1)}\right) \\ &\times\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right] \\ &+\frac{16L^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}+\frac{4(L-\mu)^2Nd\alpha_0c_0^2}{\mu(k+1)^{1+2\delta}} \\ &+\frac{8Q_kc_0^2}{\mu(k+1)^{2\delta}}+\frac{2\alpha_0^2\left(c_f q_\infty(N,d,\alpha_0,c_0)+N\sigma_1^2\right)}{c_0^2(k+1)^{2-2\delta}}. \quad (39)\end{aligned}$$

Then, we have $\forall k\geq k_2$

$$\begin{aligned}&\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right] \\ &\leq \underbrace{\exp\left(-\frac{\mu}{N}\sum_{l=k_5}^{k}\alpha_l\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]}_{t_6} \\ &+\underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{16L^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}}_{t_7} \\ &+\underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{4(L-\mu)^2Nd\alpha_0c_0^2}{\mu(k+1)^{1+2\delta}}}_{t_8} \\ &+\underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{8Q_kc_0^2}{\mu(k+1)^{2\delta}}}_{t_9} \\ &+\underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}-1}\frac{2\alpha_0^2c_f q_\infty(N,d,\alpha_0,c_0)}{c_0^2(k+1)^{2-2\delta}}}_{t_{10}} \\ &+\underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}-1}\frac{2\alpha_0^2N\sigma_1^2}{c_0^2(k+1)^{2-2\delta}}}_{t_{11}} \\ &+\underbrace{\frac{32L^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}}_{t_{12}} \\ &+\underbrace{\frac{4(L-\mu)^2N^2dc_0^2}{\mu(k+1)^{2\delta}}}_{t_{13}}+\underbrace{\frac{2Nc_0^2Q_k}{\mu\alpha_0(k+1)^{2\delta-1}}}_{t_{14}} \\ &+\underbrace{\frac{4N\alpha_0\left(c_f q_\infty(N,d,\alpha_0,c_0)+N\sigma_1^2\right)}{c_0^2\mu(k+1)^{1-2\delta}}}_{t_{15}}. \quad (40)\end{aligned}$$

It is to be noted that the term t_6 decays exponentially. The terms t_7, t_8, t_9, t_{10} and t_{11} decay faster than its counterparts in the terms t_{12}, t_{13}, t_{14} and t_{15} respectively. We note that Q_l also decays faster. Hence, the rate of decay of $\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right]$ is determined by the terms t_{12}, t_{13} and t_{15}. Thus, we have that, $\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right] = O\left(k^{-\delta_1}\right)$, where $\delta_1=\min\{1-2\delta, 2-2\tau-2\delta, 2\delta\}$. For notational ease, we refer to $t_6+t_7+t_8+t_9+t_{10}+t_{11}+t_{14}=M_k$ from now on. Finally, we note that,

$$\begin{aligned}&\|\mathbf{x}_i(k)-\mathbf{x}^*\| \leq \|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|+\left\|\underbrace{\mathbf{x}_i(k)-\overline{\mathbf{x}}(k)}_{\widetilde{\mathbf{x}}_i(k)}\right\| \\ &\Rightarrow \|\mathbf{x}_i(k)-\mathbf{x}^*\|^2 \leq 2\|\widetilde{\mathbf{x}}_i(k)\|^2+2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2 \\ &\Rightarrow \mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right] \leq 2R_k+\frac{64N\Delta_{1,\infty}\alpha_0^2}{\mu c_0^2p_{\mathcal{L}}^2\beta_0^2(k+1)^{2-2\tau-2\delta}} \\ &\frac{4(L-\mu)^2N^2dc_0^2}{\mu(k+1)^{2\delta}}+\frac{8\Delta_{1,\infty}\alpha_0^2}{p_{\mathcal{L}}^2\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}} \\ &+\frac{4N\alpha_0\left(c_f q_\infty(N,d,\alpha_0,c_0)+N\sigma_1^2\right)}{c_0^2\mu(k+1)^{1-2\delta}} \\ &\Rightarrow \mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]=O\left(\frac{1}{k^{\delta_1}}\right),\ \forall i, \quad (41)\end{aligned}$$

where $\delta_1=\min\{1-2\delta, 2-2\tau-2\delta, 2\delta\}$ and $R_k=M_k+Q_k$. By, optimizing over τ and δ, we obtain that for $\tau=1/2$ and $\delta=1/4$,

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]=O\left(\frac{1}{k^{\frac{1}{2}}}\right),\ \forall i.$$

4. Simulation Example

We provide a simulation example pertaining to ℓ_2-regularized logistic losses in random network characterized by link failures independent across iteration and links with probability p_{fail}. To be specific, we consider ℓ_2-regularized empirical risk minimization with logistic loss, where the regularization function is given by $\Psi_i(\mathbf{x})=\frac{\kappa}{2}\|\mathbf{x}\|^2$, $i=1,...,N$, with $\kappa=0.3$. In our simulation setup, each node has access to $n_i=10$ data points. The class labels and the classification vector given by $b_{ij}=\text{sign}\left((\mathbf{x}_1')^\top\mathbf{a}_{i,j}+x_0'+\epsilon_{ij}\right)$ and $x'=((\mathbf{x}_1')^\top,x_0')^\top$ respectively have ϵ_{ij}s and the entries of x' drawn independently from standard normal distribution.

The feature vectors $\mathbf{a}_{i,j}$, $j = 1, ..., n_i$, across different nodes $i = 1, \cdots, N$ and across different entries are drawn independently from different distributions. To be specific, at node i, $\mathbf{a}_{i,j}$, $j = 1, ..., n_i$ is generated by adding a standard normal random variable and an uniform random variable with support $[0,\, 5\,i]$.

We set $\beta_k = \frac{1}{\theta\,(k+1)^{1/2}}$, $\alpha_k = \frac{1}{k+1}$, $c_k = \frac{1}{(k+1)^{1/4}}$, $k = 0, 1, ...$, where $\theta = 7$ is the maximum degree across nodes. The optimizer estimate at each node is initialized as $\mathbf{x}_i(0) = 0,\ \forall\ i = 1, ..., N$.

We consider a connected network $\mathcal{G}$ with $N = 10$ nodes and 23 links, generated as an instance of a random geometric graph. The random network model assumes link failures independent across iterations and links with probability p_{fail}, where $p_{\text{fail}} \in \{0;\, 0.5;\, 0.7\}$. The case $p_{\text{fail}} = 0$ corresponds to the case where none of the links fail. We also include a comparison with the centralized zeroth order KWSA based optimization method:

$$\mathbf{y}(k+1) = \mathbf{y}(k) - \frac{1}{N(k+1)} \sum_{i=1}^{N} \nabla g_i\,(\,\mathbf{y}(k);\, \mathbf{a}_i(k), b_i(k)\,)\,, \tag{42}$$

where $(\mathbf{a}_i(k), b_i(k))$ is drawn uniformly from the set $(\mathbf{a}_{i,j}, b_{i,j})$, $j = 1, ..., n_i$. Algorithm (42) shows how (5) would be implemented if there existed a fusion node with access to all nodes' data. Hence, the comparison with (42) allows us to study the degradation of (5) due to lack of global model information. The step size for (42) is set to $1/N(k+1)$. As an error metric, we use the mean square error (MSE) estimate averaged across nodes: $\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2$.

Figure 1 plots the estimated MSE, averaged across 100 algorithm runs, versus iteration number k for $p_{\text{fail}} \in \{0;\, 0.5;\, 0.7\}$ in $\log_{10}$-$\log_{10}$ scale. The slope of the plot curve corresponds to the sublinear rate of the method; e.g., the $-1/2$ slope corresponds to a $1/k^{0.5}$ rate. It is to be noted that for all values of p_{fail}, the algorithm (5) achieves on this example (at least) the $1/k^{0.5}$ rate, thus corroborating our theory. The increase of the link failure probability only increases the constant in the MSE but does not affect the rate but the curves are only vertically shifted. Interestingly, the loss due to the increase of p_{fail} is small; e.g., the curves that correspond to $p_{\text{fail}} = 0.5$ and $p_{\text{fail}} = 0$ (no link failures) practically match. Figure 1 also shows the performance of the centralized method (42). We can see that, the distributed method (5) is very close in performance to the centralized method.

5. Conclusion

We have considered a distributed stochastic zeroth order optimization method for smooth strongly convex optimization, where we have employed a Kiefer Wolfowitz stochastic approximation type algorithm. Through the analysis of the considered method, we have established the $O(1/k^{1/2})$ MSE convergence rate for the assumed optimization setting when the underlying network is randomly varying. In particular, we have also quantified the mean square error of the generated optimizer estimate sequence in terms of the algorithm parameters. Future work includes extending the current approach to general class of convex and non-convex functions.

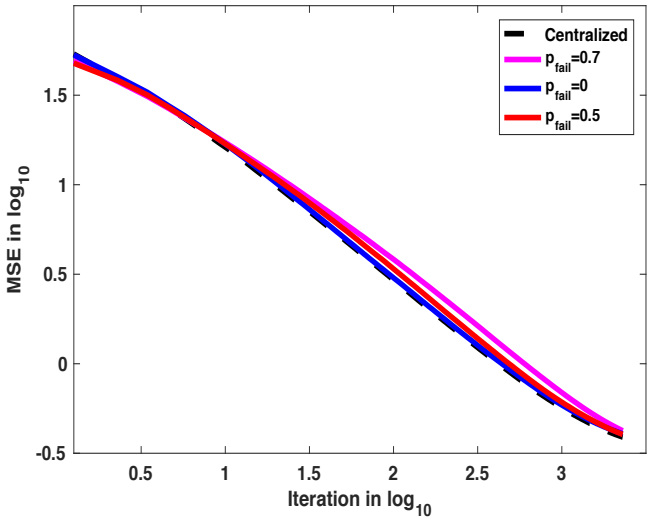

Fig. 1: Estimated MSE versus iteration number k for algorithm (5) with link failure probability $p_{\text{fail}} = 0$ (blue, solid line); 0.5 (red, solid line); and 0.7 (pink, solid line). The Figure also shows the performance of the centralized stochastic gradient method in (42) (black, dashed line).

References

[1] D. Yuan, Y. Hong, D. W. C. Ho, and G. Jiang, "Optimal distributed stochastic mirror descent for strongly convex optimization," *Automatica*, vol. 90, pp. 196–203, April 2018.

[2] K. Tsianos and M. Rabbat, "Distributed strongly convex optimization," *50th Annual Allerton Conference onCommunication, Control, and Computing*, Oct. 2012.

[3] N. D. Vanli, M. O. Sayin, and S. S. Kozat, "Stochastic subgradient algorithms for strongly convex optimization over distributed networks," *IEEE Transactions on network science and engineering*, vol. 4, no. 4, pp. 248–260, Oct.-Dec. 2017.

[4] A. Nedic and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, Dec. 2016.

[5] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

[6] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[7] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 15–26.

[8] Y. Wang, S. Du, S. Balakrishnan, and A. Singh, "Stochastic zeroth-order optimization in high dimensions," *arXiv preprint arXiv:1710.10551*, 2017.

[9] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.

[10] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, pp. 332–341, 1992.

[11] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Tech. Rep., 2011.

[12] D. Jakovetic, D. Bajovic, A. K. Sahu, and S. Kar, "Convergence rates for distributed stochastic optimization over random networks," in *57th IEEE Conference on Decision and Control (CDC)*, Miami, 2018, available at https://www.dropbox.com/s/zylonzrhypy29zj/MainCDC2018.pdf.

[13] D. Hajinezhad, M. Hong, and A. Garcia, "Zeroth order nonconvex multi-agent optimization over networks," *arXiv preprint arXiv:1710.09997*, 2017.