Communication-Efficient Distributed Strongly Convex Stochastic Optimization: Non-Asymptotic Rates

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

Abstract

We examine fundamental tradeoffs in iterative distributed zeroth and first order stochastic optimization in multi-agent networks in terms of *communication cost* (number of per-node transmissions) and *computational cost*, measured by the number of per-node noisy function (respectively, gradient) evaluations with zeroth order (respectively, first order) methods. Specifically, we develop novel distributed stochastic optimization methods for zeroth and first order strongly convex optimization by utilizing a probabilistic inter-agent communication protocol that increasingly sparsifies communications among agents as time progresses. Under standard assumptions on the cost functions and the noise statistics, we establish with the proposed method the $O(1/(C_{\rm comm})^{4/3-\zeta})$ and $O(1/(C_{\rm comm})^{8/9-\zeta})$ mean square error convergence rates, for the first and zeroth order optimization, respectively, where $C_{\rm comm}$ is the expected number of network communications and $\zeta > 0$ is arbitrarily small. The methods are shown to achieve order-optimal convergence rates in terms of computational cost $C_{\rm comp}$, $O(1/C_{\rm comp})$ (first order optimization) and $O(1/(C_{\rm comp})^{2/3})$ (zeroth order optimization), while achieving the order-optimal convergence rates in terms of iterations. Experiments on real-life datasets illustrate the efficacy of the proposed algorithms.

1. Introduction

Stochastic optimization has taken a central role in problems of learning and inference making over large data sets. Many practical setups are inherently distributed in which, due to sheer data size, it may not be feasible to store data in a single machine or agent. Further, due to the complexity of the objective functions (often, loss functions in the context of learning and inference problems), explicit computation of gradients or exactly evaluating the objective at desired arguments could be computationally prohibitive. The class of stochastic optimization problems of interest can be formalized in the following way:

$$\min f(\mathbf{x}) = \min \mathbb{E}_{\boldsymbol{\xi}\sim P}\left[F(\mathbf{x};\boldsymbol{\xi})\right],$$

where the information available to implement an optimization scheme usually involves gradients, i.e., $\nabla F(\mathbf{x};\boldsymbol{\xi})$ or function values of $F(\mathbf{x};\boldsymbol{\xi})$ itself. However, both the gradients and the function values are only unbiased estimates of the gradients and the function values of the desired objective $f(\mathbf{x})$. Moreover, due to huge data sizes and distributed

The work of DJ and DB was supported in part by the European Union (EU) Horizon 2020 project I-BiDaaS, project number 780787. The work of D. Jakovetic was 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 University of Novi Sad, Faculty of Technical Sciences, Department of Power, Electronics and Communication Engineering 21000 Novi Sad, Serbia dbajovic@uns.ac.rs. D. Jakovetic is with University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics 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.

applications, the data is often split across different agents, in which case the (global) objective reduces to the sum of N local objectives, $F(\mathbf{x};\boldsymbol{\xi}) = \sum_{i=1}^{N} F_i(\mathbf{x};\boldsymbol{\xi})$, where N denotes the number of agents. Such kind of scenarios are frequently encountered in setups such as empirical risk minimization in statistical learning [1]. In order to address the aforementioned problem setup, we study zeroth and first order distributed stochastic strongly convex optimization over networks.

There are N networked nodes, interconnected through a preassigned possibly sparse communication graph, that collaboratively aim to minimize the sum of their locally known strongly convex costs. We focus on zeroth and first order distributed stochastic optimization methods, where at each time instant (iteration) k, each node queries a stochastic zeroth order oracle ($\mathcal{SZO}$) for a noisy estimate of its local function's value at the current iterate (zeroth order optimization), or a stochastic first order oracle ($\mathcal{SFO}$) for a noisy estimate of its local function's gradient (first order optimization). In both of the proposed stochastic optimization methods, an agent updates its iterate at each iteration by simultaneously assimilating information obtained from the neighborhood (consensus) and the queried information from the relevant oracle (innovations). In the light of the aforementioned distributed protocol, our focus is then on examining the tradeoffs between the *communication cost*, measured by the number of per-node transmissions to their neighboring nodes in the network; and *computational cost*, measured by the number of queries made to $\mathcal{SZO}$ (zeroth order optimization) or $\mathcal{SFO}$ (first order optimization).

Contributions. Our main contributions are as follows. We develop novel methods for zeroth and first order distributed stochastic optimization, based on a probabilistic inter-agent communication protocol that increasingly sparsifies agent communications over time. For the proposed zeroth order method, we establish the $O(1/(C_{\mathrm{comm}})^{8/9-\zeta})$ mean square error (MSE) convergence rate in terms of communication cost C_{comm}, where $\zeta > 0$ is arbitrarily small. At the same time, the method achieves the order-optimal $O(1/(C_{\mathrm{comp}})^{2/3})$ MSE rate in terms of computational cost C_{comp} in the context of strongly convex functions with second order smoothness. For the first order distributed stochastic optimization, we propose a novel method that is shown to achieve the $O(1/(C_{\mathrm{comm}})^{4/3-\zeta})$ MSE communication rate. At the same time, the proposed method retains the order-optimal $O(1/(C_{\mathrm{comp}}))$ MSE rate in terms of the computational cost, the best achievable rate in the corresponding centralized setting.

The achieved results reveal an interesting relation between the zeroth and first order distributed stochastic optimization. Namely, as we show here, the zeroth order method achieves a slower MSE communication rate than the first order method due to the (unavoidable) presence of bias in nodes' local functions' gradient estimation. Interestingly, increasing the degree of smoothness[1] p in cost functions coupled with a fine-tuned gradient estimation scheme, adapted to the smoothness degree, effectively reduces the bias and enables the zeroth order optimization mean square error to scale as $O(1/(C_{\mathrm{comp}})^{(p-1)/p})$. Thus, with increased smoothness and appropriate gradient estimation schemes, the zeroth order optimization scheme gets increasingly close in mean square error of its first order counterpart. In a sense, we demonstrate that the first order (bias-free) stochastic optimization corresponds to the limiting case of the zeroth order stochastic optimization when $p \to \infty$.

[1]Degree of smoothness p refers to the function under consideration being p-times continuously differentiable with the p-th order derivative being Lipschitz continuous.

In more detail, the proposed distributed communication efficient stochastic methods work as follows. They utilize an increasingly sparse communication protocol that we recently proposed in the context of distributed estimation problems [2]. Therein, at each time step (iteration) k, each node participates in the communication protocol with its immediate neighbors with a time-decreasing probability p_k. The probabilities of communicating are equal across all nodes, while the nodes' decisions whether to communicate or not are independent of the past and of the other nodes. Upon the transmission stage, if active, each node makes a weighted average of its own solution estimate and the solution estimates received from all of its communication-active (transmitting) neighbors, assigning to each neighbor a time-varying weight β_k. In conjunction with the averaging step, the nodes in parallel assimilate the obtained neighborhood information and the local information through a local gradient approximation step – based on the noisy functions estimates only – with step-size α_k.

By structure, the proposed distributed zeroth and first order stochastic methods are of a similar nature, expect for the fact that rather than approximating local gradients based on the noisy functions estimates in the zeroth order case, the first order setup assumes noisy gradient estimates are directly available.

Brief literature review. We now briefly review the literature to help us contrast this paper from prior work. In the context of the extensive literature on distributed optimization, the most relevant to our work are the references that fall within the following three classes of works: 1) distributed strongly convex stochastic optimization; 2) distributed optimization over random networks (both deterministic and stochastic methods); and 3) distributed optimization methods that aim to improve communication efficiency. While we pursue stochastic optimization in this paper, the case of deterministic noiseless distributed optimization has seen much progress ([3]–[6]) and more recently accelerated methods ([7], [8]). For the first class of works, several papers give explicit convergence rates in terms of the iteration counter k, that here translates into computational cost C_{comp} or equivalently number of queries to $\mathcal{SZO}$ or $\mathcal{SFO}$, under different assumptions. Regarding the underlying network, references [9], [10] consider static networks, while the works [11]–[13] consider deterministic time-varying networks. They all consider *first order* optimization.

References [9], [10] consider distributed first order strongly convex optimization for static networks, assuming that the data distributions that underlie each node's local cost function are equal (reference [9] considers empirical risks while reference [10] considers risk functions in the form of expectation); this essentially corresponds to each nodes' local function having the same minimizer. References [11]–[13] consider deterministically varying networks, assuming that the "union graph" over finite windows of iterations is connected. The papers [9]–[12] assume undirected networks, while [13] allows for directed networks and assumes a bounded support for the gradient noise. The works [9], [11]–[13] allow the local costs to be non-smooth, while [10] assumes smooth costs, as we do here. With respect to these works, we consider random networks (that are undirected and connected on average), smooth costs, and allow the noise to have unbounded support. The authors of [14] propose a distributed zeroth optimization algorithm for non-convex minimization with a static graph, where a random directions-random smoothing approach was employed.

For the second class of works, distributed optimization over random networks has been studied in [15]–[17]. References [15], [16] consider non-differentiable convex costs, first order methods, and no (sub)gradient noise,

while reference [17] considers differentiable costs with Lipschitz continuous and bounded gradients, first order methods, and it also does not allow for gradient noise, i.e., it considers methods with exact (deterministic) gradients. Reference [18] considers distributed stochastic first order methods and establishes the method's $O(1/k)$ convergence rate. References [19] considers a zeroth order distributed stochastic approximation method, which queries the $\mathcal{SZO}$ $2d$ times at each iteration where d is the dimension of the optimizer and establishes the method's $O(1/k^{1/2})$ convergence rate in terms of the number of iterations under first order smoothness.

In summary, each of the references in the two classes above is not primarily concerned with studying communication rates of distributed stochastic methods. Prior work achieves order-optimal rates in terms of computational cost (that translates here into the number of iterations k), both for the zeroth order, e.g., [19], and for the first order, e.g., [18], distributed strongly convex optimization.[2]In contrast, we establish here communication rates as well. This paper and our prior works [19], [20] distinguish further from other works on distributed zeroth order optimization, e.g., [14], [21], in that, not only the gradient is approximated through function values due to the absence of first order information, but also the function values themselves are subject to noise. Reference [20] considers a communication efficient zeroth order approximation scheme, where the convergence rate is established to be $O(1/k^{1/2})$ and the MSE-communication is improved to $O(1/(C_{\text{comm}})^{2/3-\zeta})$. In contrast to [20], with additional smoothness assumptions we improve the convergence rate to $O(1/k^{2/3})$ and the MSE-communication is further improved to $O(1/(C_{\text{comm}})^{8/9-\zeta})$.

Finally, we review the class of works that are concerned with designing distributed methods that achieve communication efficiency, e.g., [2], [22]–[27]. In [26], a data censoring method is employed in the context of distributed least squares estimation to reduce computational and communication costs. However, the communication savings in [26] are a constant proportion with respect to a method which utilizes all communications at all times, thereby not improving the order of the convergence rate. References [22]–[24] also consider a different setup than we do here, namely they study distributed optimization where the data is available a priori (i.e., it is not streamed). This corresponds to an intrinsically different setting with respect to the one studied here, where actually geometric MSE convergence rates are attainable with stochastic-type methods, e.g., [28]. In terms of the strategy to save communications, references [22]–[25] consider, respectively, deterministically increasingly sparse communication, an adaptive communication scheme, and selective activation of agents. These strategies are different from ours; we utilize randomized, increasingly sparse communications in general. In references [2], [27], we study distributed estimation problems and develop communication-efficient distributed estimators. The problems studied in [2], [27] have a major difference with respect to the current paper in that, in [2], [27], the assumed setting yields individual nodes' local gradients to evaluate to zero at the global solution. In contrast, the model assumed here does not feature such property, and hence it is more challenging.

Finally, we comment on the recent paper [25] that develops communication-efficient distributed methods for both non-stochastic and stochastic distributed first order optimization, both in the presence and in the absence of the

[2]The works in the first two classes above utilize a non-diminishing amount of communications across iterations, and hence they achieve at best the $O(1/(C_{\text{comm}}))$ (first order optimization) and $O(1/(C_{\text{comm}})^{1/2})$ communication rates.

strong convexity assumption. For the stochastic, strongly convex first order optimization, [25] shows that the method therein gets ϵ-close to the solution in $O(1/\sqrt{\epsilon})$ communications and with an $O(1/\epsilon)$ computational cost. The current paper has several differences with respect to [25]. First, reference [25] does not study zeroth order optimization. Second, this work assumes for the gradient noise to be independent of the algorithm iterates. This is a strong assumption that may be not satisfied, e.g., with many machine learning applications. Third, while we assume here twice differentiable costs, this assumption is not imposed in [25]. Finally, the method in [25] is considerably more complex than the one proposed here, with two layers of iterations (inner and outer iterations). In particular, the inner iterations involve solving an exact minimization problem which necessarily points to the usage of an off-the-shelf solver, the computation cost of which is not factored into the computation cost in [25].

Paper organization. The next paragraph introduces notation. Section 2 describes the model and the proposed algorithms for zeroth and first order distributed stochastic optimization. Section 3 states our convergence rates results for the two methods. Sections 5 and 6 provide proofs for the zeroth and first order methods, respectively. Section 4 demonstrates communication efficiency of the proposed methods through numerical examples. Finally, we conclude in Section 7.

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). For a set $\mathcal{X}$, $|\mathcal{X}|$ denotes the cardinality of set $\mathcal{X}$. We denote by: $\|\cdot\| = \|\cdot\|_2$ the Euclidean (respectively, induced) norm of its vector (respectively, matrix) argument; $\lambda_i(\cdot)$, the i-th smallest eigenvalue of its matrix argument; $\nabla h(w)$ 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. By $\mathbf{e}_j$ we denote the j-th column of 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 and the proposed algorithms

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 strongly convex function available to node i, $i = 1, ..., N$. We make the following assumption on the functions $f_i(\cdot)$:

Assumption 1. 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 1 we have that each f_i, $i = 1, \cdots, N$, is μ-strongly convex. Using standard properties of strongly 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 1, 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 gradient methods to solve (1). That is, we study algorithms of the following form:

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) - \sum_{j \in \Omega_i(k)} \gamma_{i,j}(k) \left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right)$$
$$- \alpha_k \widehat{\mathbf{g}}_i(\mathbf{x}_i(k)), \tag{2}$$

where the weight assigned to an incoming message $\gamma_{i,j}(k)$ and the neighborhood of an agent $\Omega_i(k)$ are determined by the specific instance of the designated communication protocol. The approximated gradient $\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))$ is specific to the optimization, i.e., whether it is a zeroth order optimization or a first order optimization scheme. Technically speaking, as we will see later, a zeroth order optimization scheme approximates the gradient as a biased estimate of the gradient while a first order optimization scheme approximates the gradient as an unbiased estimate of the gradient. The variation in the gradient approximation across first order and zeroth order methods can be attributed to the fact that the oracles from which the agents query for information pertaining to the loss function differ. For instance, in the case of the zeroth order optimization, the agents query a stochastic zeroth order oracle ($\mathcal{SZO}$) and in turn receive noisy function values (unbiased estimates) for the queried point. However, in the case of first order optimization, the agents query a stochastic first order oracle ($\mathcal{SFO}$) and receive unbiased estimates of the gradient. In subsequent sections, we will explore the gradient approximations in greater detail. Before stating the algorithms, we first discuss the communication scheme. Specifically, we adopt the following model.

1) Communication Scheme: The inter-node communication network to which the information exchange between nodes conforms to is modeled as an *undirected* simple connected graph $G = (V, E)$, with $V = [1 \cdots N]$ and E denoting the set of nodes and communication links. The neighborhood of node n is given by $\Omega_n = \{l \in V \mid (n, l) \in E\}$. The node n has degree $d_n = |\Omega_n|$. The structure of the graph is described by the $N \times N$ adjacency matrix, $\mathbf{A} = \mathbf{A}^\top = [\mathbf{A}_{nl}]$, $\mathbf{A}_{nl} = 1$, if $(n, l) \in E$, $\mathbf{A}_{nl} = 0$, otherwise. The graph Laplacian $\overline{\mathbf{R}} = \mathbf{D} - \mathbf{A}$ is positive semidefinite, with eigenvalues ordered as $0 = \lambda_1(\overline{\mathbf{R}}) \leq \lambda_2(\overline{\mathbf{R}}) \leq \cdots \leq \lambda_N(\overline{\mathbf{R}})$, where $\mathbf{D}$ is given by $\mathbf{D} = \text{diag}\,(d_1 \cdots d_N)$. We make the following assumption on $\overline{\mathbf{R}}$.

Assumption 2. The inter-agent communication graph is connected on average, i.e., $\overline{\mathbf{R}}$ is connected. In other words, $\lambda_2(\overline{\mathbf{R}}) > 0$.

Thus, $\overline{\mathbf{R}}$ corresponds to the maximal graph, i.e., the graph of all *allowable* communications. We now describe our randomized communication protocol that selects a (random) subset of the allowable links at each time instant for information exchange.

For each node i, at every time k, we introduce a binary random variable $\psi_{i,k}$, where

$$\psi_{i,k} = \begin{cases} \rho_k & \textit{with probability } \zeta_k \\ 0 & \textit{otherwise}, \end{cases} \tag{3}$$

where $\psi_{i,k}$'s are independent both across time and the nodes, i.e., across k and i respectively. The random variable $\psi_{i,k}$ abstracts out the decision of the node i at time k whether to participate in the neighborhood information exchange or not. We specifically take ρ_k and ζ_k of the form

$$\rho_k = \frac{\rho_0}{(k+1)^{\epsilon/2}},\ \zeta_k = \frac{\zeta_0}{(k+1)^{(\tau/2-\epsilon/2)}}, \tag{4}$$

where $0 < \tau \leq \frac{1}{2}$ and $0 < \epsilon < \tau$. Furthermore, define β_k to be

$$\beta_k = (\rho_k \zeta_k)^2 = \frac{\beta_0}{(k+1)^{\tau}}, \tag{5}$$

where $\beta_0 = \rho_0^2 \zeta_0^2$. With the above development in place, we define the random time-varying Laplacian $\mathbf{R}(k)$, where $\mathbf{R}(k) \in \mathbb{R}^{N \times N}$ abstracts the inter-node information exchange as follows:

$$\mathbf{R}_{i,j}(k) = \begin{cases} -\psi_{i,k}\psi_{j,k} & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ \sum_{l \neq i} \psi_{i,k}\psi_{l,k} & i = j. \end{cases} \tag{6}$$

The above communication protocol allows two nodes to communicate only when the link is established in a bi-directional fashion and hence avoids directed graphs. The design of the communication protocol as depicted in (3)-(6) not only decays the weight assigned to the links over time but also decays the probability of the existence of a link. Such a design is consistent with frameworks where the agents have finite power and hence not only the number of communications, but also, the quality of the communication decays over time. We have, for $\{i,j\} \in E$ and $i \neq j$:

$$\begin{aligned} \mathbb{E}\left[\mathbf{R}_{i,j}(k)\right] &= -\left(\rho_k \zeta_k\right)^2 = -\beta_k = -\frac{\beta_0}{(k+1)^{\tau}} \\ \mathbb{E}\left[\mathbf{R}_{i,j}^2(k)\right] &= \left(\rho_k^2 \zeta_k\right)^2 = \frac{\rho_0^2 \beta_0}{(k+1)^{\tau+\epsilon}}. \end{aligned} \tag{7}$$

Thus, we have that, the variance of $\mathbf{R}_{i,j}(k)$ is given by,

$$\textit{Var}\left(\mathbf{R}_{i,j}(k)\right) = \frac{\beta_0 \rho_0^2}{(k+1)^{\tau+\epsilon}} - \frac{\beta_0^2}{(k+1)^{2\tau}}. \tag{8}$$

Define, the mean of the random time-varying Laplacian sequence $\{\mathbf{R}(k)\}$ as $\overline{\mathbf{R}}_k = \mathbb{E}\left[\mathbf{R}(k)\right]$ and $\widetilde{\mathbf{R}}(k) = \mathbf{R}(k) - \overline{\mathbf{R}}_k$. Note that, $\mathbb{E}\left[\widetilde{\mathbf{R}}(k)\right] = \mathbf{0}$, and

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{R}}(k)\right\|^2\right] \leq 4N^2 \mathbb{E}\left[\widetilde{\mathbf{R}}_{i,j}^2(k)\right] = \frac{4N^2 \beta_0 \rho_0^2}{(k+1)^{\tau+\epsilon}} - \frac{4N^2 \beta_0^2}{(k+1)^{2\tau}}, \tag{9}$$

where $\|\cdot\|$ denotes the $\mathcal{L}_2$ norm. The above equation follows by relating the $\mathcal{L}_2$ and Frobenius norms.
We also have that, $\overline{\mathbf{R}}_k = \beta_k \overline{\mathbf{R}}$, where

$$\overline{\mathbf{R}}_{i,j} = \begin{cases} -1 & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l \neq i} \overline{\mathbf{R}}_{i,l} & i = j. \end{cases} \tag{10}$$

Technically speaking, the communication graph at each time k encapsulated as $\mathbf{R}(k)$ need not be connected at all times, although the graph of allowable links G is connected.. In fact, at any given time k, only a few of the possible

links could be active. However, since $\overline{\mathbf{R}}_k = \beta_k \overline{\mathbf{R}}$, we note that, by Assumption 2, the instantaneous Laplacian $\mathbf{R}(k)$ is connected on average.The connectedness in average basically ensures that over time, the information from each agent in the graph reaches other agents over time in a symmetric fashion and thus ensuring information flow, while providing the leeway for the instantaneous communication graphs at different times to be not connected.
We employ a primal algorithm for solving the optimization problem in (1). In particular, the update in (2) can then be written in a vector form as follows:

$$\mathbf{x}(k+1) = \mathbf{W}_k \mathbf{x}(k) - \alpha_k \widehat{\mathbf{G}}(\mathbf{x}(k)), \tag{11}$$

where $\mathbf{x}(k) = \left[\mathbf{x}_1^\top(k), \cdots, \mathbf{x}_N^\top(k)\right]^\top \in \mathbb{R}^{Nd}$, $F(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x}_i)$, $\mathbf{x} = \left[\mathbf{x}_1^\top, \cdots, \mathbf{x}_N^\top\right]^\top \in \mathbb{R}^{Nd}$, $\widehat{\mathbf{G}}(\mathbf{x}(k)) = [\widehat{\mathbf{g}}_i^\top(\mathbf{x}_i(k)), \cdots, \widehat{\mathbf{g}}^\top(\mathbf{x}_N(k))]^\top$ and $\mathbf{W}_k = (\mathbf{I} - \mathbf{R}(k)) \otimes \mathbf{I}_d$. We state an assumption on the weight sequences before proceeding further.

Assumption 3. The weight sequence α_k is given by $\alpha_0/(k+1)$, where $\alpha_0 > 1/\mu$. For the sequence ρ_k as defined in (4), it is chosen in such a way that,

$$\rho_0^2 \leq \frac{4N^2}{\lambda_2\left(\overline{\mathbf{R}}\right)}. \tag{12}$$

In the following sections, we propose two different approaches to solve the optimization problem in (1). The first approach involves zeroth order optimization, while the second approach involves a first order optimization. We first study the zeroth order approach to the problem in (1).

A. Zeroth Order Optimization

We employ a random directions stochastic approximation (RDSA) type method from [29] adapted to our distributed setup 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 addition to the smoothness assumption in 1, we define additional smoothness assumptions in the context of zeroth order optimization.

Assumption A1. For all $i = 1, ..., N$, the functions $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ have their Hessian to be M-Lipschitz, i.e.,

$$\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{y})\| \leq M \|\mathbf{x} - \mathbf{y}\|, \forall i = 1, \cdots, N.$$

In order to carry out the optimization, each agent i makes queries to the $\mathcal{SZO}$ 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,

$$\widehat{f}_i(\mathbf{x}_i(k)) = f_i(\mathbf{x}_i(k)) + \widehat{v}_i(k; \mathbf{x}_i(k)), \tag{13}$$

where the first argument in $\widehat{v}_i(k; \mathbf{x}_i(k))$ is the iteration number, and the second argument is the point at which the $\mathcal{SZO}$ oracle is queried. The properties of the noise $\widehat{v}_i(k; \mathbf{x}_i(k))$ are discussed further ahead. Typically due to the unavailability of the analytic form of the functionals in zeroth order methods, the gradient cannot be explicitly evaluated and hence, we resort to a gradient approximation. In order to approximate the gradient, each agent makes three calls to the stochastic zeroth order oracle. For instance, agent i queries for $f_i(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k})$, $f_i(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k}/2)$ and $f_i(\mathbf{x}_i(k))$ at time k and obtains $\widehat{f}_i(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k})$, $\widehat{f}_i(\mathbf{x}_i(k) + c_k \mathbf{z}_{i,k}/2)$ and $\widehat{f}_i(\mathbf{x}_i(k))$ respectively, where c_k is a carefully chosen time-decaying constant and $\mathbf{z}_{i,k}$ is a random vector (to be specified soon) such that $\mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$.
Denote by $\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))$ the approximated gradient which is given by:

$$\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)) \doteq 2\widetilde{\mathbf{g}}_i\left(\mathbf{x}_i(k), \frac{c_k}{2}\right) - \widetilde{\mathbf{g}}_i\left(\mathbf{x}_i(k), c_k\right)$$

$$= \frac{4\widehat{f}_i\left(\mathbf{x}_i(k) + \frac{c_k}{2}\mathbf{z}_{i,k}\right) - 4\widehat{f}_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k}$$
$$- \frac{\widehat{f}_i\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right) - \widehat{f}_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k}, \tag{14}$$

where $\widetilde{\mathbf{g}}_i(\cdot,\cdot)$ represents a first order finite difference operation and $\theta_1, \theta_2 \in [0,1]$. Note that, the gradient approximation derived in (14) involves the noise in the retrieved function value from the $\mathcal{SZO}$ differently from other RDSA approaches such as in [21], [29]. The finite difference technique used in (14) resembles, *the twicing trick* commonly used in Kernel density estimation which is employed so as to reduce bias and approximately eliminate the effect of the second degree term from the bias. It is also to be noted that the number of queries made to the $\mathcal{SZO}$ at every gradient approximation is 3. Thus, we can write,

$$\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)) = \nabla f_i\left(\mathbf{x}_i(k)\right) + \underbrace{\mathbb{E}\left[\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))|\mathcal{F}_k\right] - \nabla f_i\left(\mathbf{x}_i(k)\right)}_{c_k\mathbf{b}_i(\mathbf{x}_i(k))}$$
$$+ \underbrace{\mathbf{g}_i(\mathbf{x}_i(k)) - \mathbb{E}\left[\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))|\mathcal{F}_k\right] + \frac{v_i(k;\mathbf{x}_i(k))\mathbf{z}_{i,k}}{c_k}}_{\mathbf{h}_i(\mathbf{x}_i(k))}, \tag{15}$$

where

$$\mathbf{g}_i(\mathbf{x}_i(k)) = \frac{4f_i\left(\mathbf{x}_i(k) + \frac{c_k}{2}\mathbf{z}_{i,k}\right) - 4f_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k}$$
$$- \frac{f_i\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right) - f_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k}, \tag{16}$$

$$v_i(k;\mathbf{x}_i(k)) = 4\left(\widehat{f}_i\left(\mathbf{x}_i(k) + \frac{c_k}{2}\mathbf{z}_{i,k}\right) - f_i\left(\mathbf{x}_i(k) + \frac{c_k}{2}\mathbf{z}_{i,k}\right)\right)$$
$$- 3(\widehat{f}_i(\mathbf{x}_i(k)) - f_i(\mathbf{x}_i(k))) - (\widehat{f}_i\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right)$$
$$- f_i\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right)), \tag{17}$$

and, $\mathcal{F}_k$ denotes the history of the proposed algorithm up to time k. Given that the sources of randomness in our algorithm are the noise sequence $\{\mathbf{v}(k;\mathbf{x}(k))\}$, the random network sequence $\{\mathbf{R}(k)\}$ and the random vectors for directional derivatives $\{\mathbf{z}_k\}$, $\mathcal{F}_k$ is given by the σ-algebra generated by the collection of random variables $\{\mathbf{R}(s), \mathbf{v}(k;\mathbf{x}(k)), \mathbf{z}_{i,s}\}$, $i = 1, ..., N$, $s = 0, ..., k-1$.

In general, the higher order smoothness imposed by Assumption 3 allows us to use a higher order finite difference approximation for estimating the gradient. Due to assumption 3, the bias in the gradient estimate by employing a second order finite difference approximation of the gradient is of the order $O(c_k^2)$. Instead, a first order finite difference approximation of the gradient would have yielded a bias of $O(c_k)$. More generally, an assumption involving p-th order smoothness of the loss functions would have enabled usage of a p-th degree finite difference approximation of the gradient thus leading to a bias of $O(c_k^p)$.

Assumption A2. The $z_{i,k}$'s are drawn from a distribution P such that $\mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$, $s_1(P) = \mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right]$ and $s_2(P) = \mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right]$ are finite.

We provide two examples of two such distributions. If $\mathbf{z}_{i,k}$'s are drawn from $\mathcal{N}(0,\mathbf{I}_d)$, then $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d(d+2)$ and $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right] = d(d+2)(d+4)$. If $\mathbf{z}_{i,k}$'s are drawn uniformly from the l_2-ball of radius $\sqrt{d}$, then we have, $\|\mathbf{z}_{i,k}\| = \sqrt{d}$, $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d^2$ and $\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^4\right] = d^3$. For the rest of the paper, we assume that $\mathbf{z}_{i,k}$'s are sampled from a normal distribution with $\mathbf{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$ or uniformly from the surface of the l_2-ball of radius $\sqrt{d}$.

Remark 2.1. *The RDSA scheme (see, for example [29]) used here is similar to the simultaneous perturbation stochastic approximation scheme (SPSA) as proposed in [30]. In SPSA, each dimension i of the optimization iterate is perturbed by a random variable Δ_i. However, instead of RDSA where the directional derivative is taken along the sampled vector* $\mathbf{z}$*, the directional derivative in case of SPSA is along the direction* $[1/\Delta_1, \cdots, 1/\Delta_d]$ *which thus needs boundedness of the inverse moments of the random variable* Δ_i*. The particular choice for* Δ_i*'s is taken to be the Bernoulli distribution with* Δ_i*'s taking values* 1 *and* -1 *with probability* 0.5*. It is to be noted that at each iteration, both RDSA and SPSA approximate the gradient by making two calls to the stochastic zeroth order oracle as opposed to* d *calls in the case of Kiefer Wolfowitz Stochastic Approximation (KWSA) (see, [31] for example).*

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:

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) - \sum_{j \in \Omega_i(k)} \psi_{i,k}\psi_{j,k} \left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right) - \alpha_k \widehat{\mathbf{g}}_i(\mathbf{x}_i(k)), \tag{18}$$

where $\widehat{\mathbf{g}}_i(\cdot)$ is as defined in (15). Comparing to the general update in (2), the time-varying weight $\gamma_{i,j}(k)$ at agent i to the incoming message from agent j is given by $\psi_{j,k}$.

Remark 2.2. *The main intuition behind the randomized activation albeit in a controlled manner for both the zeroth order and first order optimization methods is the fact that in expectation both the updates exactly reduce to the update where the communication graph between agents is realized by the expected Laplacian.*

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 (18) are biased (see (15)) due to the unavailability of the exact gradient functions and their approximations using the zeroth order scheme in (14). 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 sequence c_k is given by:

$$c_k = \frac{1}{s_1(P)(k+1)^\delta}, \tag{19}$$

where $\delta > 0$. The constant $\delta > 0$ is chosen in such a way that,

$$\sum_{k=1}^{\infty} \frac{\alpha_k^2}{c_k^2} < \infty \tag{20}$$

The update in (18) can be written as:

$$\mathbf{x}(k+1) = \mathbf{W}_k\mathbf{x}(k) - \alpha_k \nabla F(\mathbf{x}(k)) - \alpha_k c_k \mathbf{b}(\mathbf{x}(k)) - \alpha_k \mathbf{h}(\mathbf{x}(k)), \tag{21}$$

where $\mathbf{b}(\mathbf{x}(k)) = \left[\mathbf{b}_1^\top\left(\mathbf{x}_1(k)\right), \cdots, \mathbf{b}_N^\top\left(\mathbf{x}_N(k)\right)\right]^\top \in \mathbb{R}^{Nd}$ and $\mathbf{h}(\mathbf{x}(k)) = \left[\mathbf{h}_1^\top\left(\mathbf{x}_1(k)\right), \cdots, \mathbf{h}_N^\top\left(\mathbf{x}_N(k)\right)\right]^\top \in \mathbb{R}^{Nd}$. We state an assumption on the measurement noises next.

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

$$\begin{aligned} &\mathbb{E}[\, v_i(k; \mathbf{x}_i(k)) \,|\, \mathcal{F}_k, \mathbf{z}_{i,k}] = 0, \ \text{almost surely (a.s.)} \\ &\mathbb{E}[\, v_i(k; \mathbf{x}_i(k))^2 \,|\, \mathcal{F}_k, \mathbf{z}_{i,k}] \leq c_v \|\mathbf{x}_i(k)\|^2 + \sigma_v^2, \ \text{a.s.}, \end{aligned} \tag{22}$$

where c_v and σ_v^2 are nonnegative constants.

Assumption A4 is standard in the analysis of stochastic optimization methods, e.g., [10]. It is stated in terms of noise $\mathbf{v}_i(k; \mathbf{x}_i(k))$ in (17) rather then directly in terms of the $\mathcal{SZO}$ noises in equation (13), for notational simplicity. An equivalent statement can be made in terms of the noises in (13). The assumption about the conditional independence between the random directions $\mathbf{z}_{i,k}$ and the function noise $v_i(k; \mathbf{x}_i(k))$ is mild. It merely formalizes the model that we consider, namely that, given history $\mathcal{F}_k$, drawing a random direction sample $\mathbf{z}_{i,k}$ and querying function values from the $\mathcal{SZO}$ are performed in a statistically independent manner.
We remark that by Assumption A4,

$$\begin{aligned} &\mathbb{E}\left[v_i(k; \mathbf{x}_i(k))\mathbf{z}_{i,k} | \mathcal{F}_k\right] = \mathbb{E}\left[\mathbf{z}_{i,k} \mathbb{E}\left[v_i(k; \mathbf{x}_i(k)) | \mathcal{F}_k, \mathbf{z}_{i,k}\right] \mid \mathcal{F}_k\right] \\ &\Rightarrow \mathbb{E}\left[\mathbf{v}_{\mathbf{z}}(k; \mathbf{x}(k)) \mid \mathcal{F}_k\right] = \mathbf{0}. \end{aligned} \tag{23}$$

and,

$$\begin{aligned} &\mathbb{E}\left[\|v_i(k; \mathbf{x}_i(k))\mathbf{z}_{i,k}\|^2 | \mathcal{F}_k\right] \\ &= \mathbb{E}\left[\|\mathbf{z}_{i,k}\|^2 \mathbb{E}\left[v_i^2(k; \mathbf{x}_i(k)) | \mathcal{F}_k, \mathbf{z}_{i,k}\right] | \mathcal{F}_k\right] \\ &\leq \mathbb{E}\left[\|\mathbf{z}_{i,k}\|^2\right] \left(c_v \|\mathbf{x}_i(k)\|^2 + \sigma_v^2\right), \end{aligned} \tag{24}$$

where if $\mathbf{z}_{i,k}$'s are sampled from a normal distribution with $\mathbf{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$ or uniformly from the surface of the l_2-ball of radius $\sqrt{d}$, then we have,

$$\mathbb{E}\left[\|v_i(k; \mathbf{x}_i(k))\mathbf{z}_{i,k}\|^2 | \mathcal{F}_k\right] \leq d\left(c_v \|\mathbf{x}_i(k)\|^2 + \sigma_v^2\right). \tag{25}$$

B. First Order Optimization

Each node i, $i = 1, ..., N$, in the network maintains its own optimizer $\mathbf{x}_i(k) \in \mathbb{R}^d$ at each time step (iterations) $k = 0, 1, ...,$. Specifically, for arbitrary deterministic initial points $\mathbf{x}_i(0) \in \mathbb{R}^d$, $i = 1, ..., N$, the update rule at node i and $k = 0, 1, ...,$ is as follows:

$$\begin{aligned} \mathbf{x}_i(k+1) &= \mathbf{x}_i(k) - \sum_{j \in \Omega_i} \psi_{i,k}\psi_{j,k} \left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right) \\ &- \alpha_k \left(\nabla f_i(\mathbf{x}_i(k)) + \mathbf{u}_i(k)\right). \end{aligned} \tag{26}$$

In comparison to the generalized update in (2), the weights assigned to incoming messages is given by $\gamma_{i,j}(k) = \psi_{i,k}\psi_{j,k}$, while the approximated gradient is given by $\nabla f_i(\mathbf{x}_i(k)) + \mathbf{u}_i(k)$. The update (26) is realized in a parallel fashion at all nodes $i = 1, ..., N$. First, each node i, when activated, i.e., when $\psi_{i,k} \neq 0$, broadcasts $\mathbf{x}_i(k)$ to all its active neighbors $j \in \Omega_i$ which satisfy $\psi_{j,k} \neq 0$ and receives $\mathbf{x}_j(k)$ from all $j \in \Omega_i$ which are active. Subsequently,

each node i, $i = 1, ..., N$ makes update (26), which completes an iteration. Finally, $\mathbf{u}_i(k)$ is noise in the calculation of the f_i's gradient at iteration k. For $k = 0, 1, ...$, algorithm (26) can be compactly written as follows:

$$\mathbf{x}(k+1) = \mathbf{W}_k \mathbf{x}(k) - \alpha_k \left(\nabla F(\mathbf{x}(k)) + \mathbf{u}(k)\right), \tag{27}$$

where $\mathbf{x} = \left[\mathbf{x}_1^\top, \cdots, \mathbf{x}_N^\top\right]^\top \in \mathbb{R}^{Nd}$ and $\mathbf{u}(k) = \left[\mathbf{u}_1^\top(k), \cdots, \mathbf{u}_N^\top(k)\right]^\top \in \mathbb{R}^{Nd}$. We make the following standard assumption on the gradient noises. First, denote by $\mathcal{S}_k$ the history of algorithm (26) up to time k; that is, $\mathcal{S}_k$, $k = 1, 2, ...$, is an increasing sequence of σ-algebras, where $\mathcal{S}_k$ is the σ-algebra generated by the collection of random variables $\{\mathbf{R}(s),\ \mathbf{u}_i(t)\}$, $i = 1, ..., N$, $s = 0, ..., k-1$, $t = 0, ..., k-1$.

Assumption B2. For each $i = 1, ..., N$, the sequence of noises $\{\mathbf{u}_i(k)\}$ satisfies for all $k = 0, 1, ...$:

$$\mathbb{E}[\,\mathbf{u}_i(k)\,|\,\mathcal{S}_k\,] = 0, \text{ almost surely (a.s.)} \tag{28}$$

$$\mathbb{E}[\,\|\mathbf{u}_i(k)\|^2\,|\,\mathcal{S}_k\,] \leq c_u\|\mathbf{x}_i(k)\|^2 + \sigma_u^2, \text{ a.s.,} \tag{29}$$

where c_u is a nonnegative constant.

Communication Cost Define the communication cost $\mathcal{C}_k$ to be the expected per-node number of transmissions up to iteration k, i.e.,

$$\mathcal{C}_k = \mathbb{E}\left[\sum_{s=0}^{k-1} \mathbb{I}_{\{\textit{node } C \textit{ transmits at } s\}}\right], \tag{30}$$

where $\mathbb{I}_A$ represents the indicator of event A. Note that the per-node communication cost in (30) is the same as the network average of communication costs across all nodes, as the activation probabilities are homogeneous across nodes. We now proceed to the main results pertaining to the proposed zeroth order and first order optimization schemes.

3. Convergence rates: Statement of main results and interpretations

In this section, we state the results for both the zeroth order and the first order optimization algorithms.

A. Main Results: Zeroth Order Optimization

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* (18)*. Let assumptions 1-3 and A1-A4 hold. Then, for each node i's optimizer estimate $\mathbf{x}_i(k)$ and the solution $\mathbf{x}^\star$ of problem* (1)*, $\forall k \geq 0$ there holds:*

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq\ & 2M_k + \frac{64NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{R}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau-2\delta}} \\ & \frac{16NM^2d^2(P)c_0^4}{\mu^2(k+1)^{4\delta}} + 2Q_k + \frac{8\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}} \\ & + \frac{4N\alpha_0\left(dc_vq_\infty(N,d,\alpha_0,c_0) + dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}},\end{aligned} \tag{31}$$

where, $\Delta_{1,\infty} = 6dc_vq_\infty(N,d,\alpha_0,c_0) + 6dN\sigma_1^2$ *and* $q_\infty(N,d,\alpha_0,c_0) = \mathbb{E}\left[\|\mathbf{x}(k_2) - \mathbf{x}^o\|^2\right] + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} + \frac{\sqrt{N}s_1(P)M\alpha_0c_0^2}{8\delta} + \frac{Ns_1^2(P)M^2\alpha_0^2c_0^4}{16(1+4\delta)} + \frac{d\alpha_0^2\left(2c_vN\|\mathbf{x}^o\|^2 + N\sigma_v^2\right)}{c_0^2(1-2\delta)} + \frac{\alpha_0^2c_0^2\sqrt{N}s_1(P)M\|\nabla F(\mathbf{x}^o)\|}{1+2\delta} + \frac{2N\alpha_0^2c_0^4s_2(P)}{1+4\delta} + \frac{4\alpha_0^2c_0^2Ns_1(P)}{1+2\delta}\|\nabla F\left(\mathbf{x}^o\right)\|^2$, $k_2 = \max\{k_0, k_1\}$, $k_0 = \inf\{k|\mu^2\alpha_k^2 < 1\}$ *and* $k_1 = \inf\left\{k|\frac{\mu}{2} > \frac{\sqrt{N}}{4}s_1(P)Mc_k^2 + \frac{2dc_v\alpha_k}{c_k^2} + 4\alpha_kc_k^2Ns_1(P)L^2\right\}$*, with M_k and Q_k decaying*

faster than the rest of the terms.

2) In particular, the rate of decay of the RHS of (31) *is given by* $(k+1)^{-\delta_1}$*, where* $\delta_1 = \min\{1-2\delta, 2-2\tau-2\delta, 4\delta\}$*. By, optimizing over* τ *and* δ*, we obtain that for* $\tau = 1/2$ *and* $\delta = 1/6$*,*

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq 2M_k + \frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{R}}\right)c_0^2\beta_0^2(k+1)^{2/3}}$$
$$\frac{16NM^2d^2(P)c_0^2}{\mu^2(k+1)^{2/3}} + 2Q_k + \frac{8\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2c_0^2(k+1)^{2/3}}$$
$$+ \frac{4N\alpha_0\left(dc_vq_\infty(N,d,\alpha_0,c_0) + dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{2/3}} = O\left(\frac{1}{k^{\frac{2}{3}}}\right),\ \forall i.$$

3) The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k}\zeta_t\right] = O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right).$$

and the MSE-communication rate is given by,

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O\left(\mathcal{C}_k^{-8/9+\zeta}\right), \tag{32}$$

where ζ *can be arbitrarily small.*

Theorem 3.1 asserts an $O\left(\mathcal{C}_k^{-8/9+\zeta}\right)$ MSE-communication rate can be achieved while keeping the MSE decay rate at $O\left(k^{-\frac{2}{3}}\right)$. The performance of the zeroth order optimization scheme depends explicitly on the connectivity of the expected Laplacian through the terms $\frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2(\overline{\mathbf{R}})c_0^2\beta_0^2(k+1)^{0.5}}$ and $\frac{8\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2(\overline{\mathbf{R}})\beta_0^2c_0^2(k+1)^{0.5}}$. In particular, communication graphs which are well connected, i.e., have higher values of $\lambda_2\left(\overline{\mathbf{R}}\right)$ will have lower MSE as compared to a counterpart with lower values of $\lambda_2\left(\overline{\mathbf{R}}\right)$.

If higher order smoothness assumptions are made, i.e., a p-th order smoothness assumption is made which is then exploited by means of a p-th degree finite difference gradient approximation, then by repeating the same proof arguments, the rate in terms of iteration count can be shown to improve to $O\left(k^{-\frac{p}{p+1}}\right)$. The improvement can be attributed to a better bias-variance tradeoff as illustrated by the terms $\frac{8M^2d^2(P)c_0^4}{\mu^2(k+1)^{2p\delta}}$ and $\frac{4N\alpha_0\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}}$. The corresponding MSE-communication rate improves to $O\left(\mathcal{C}_k^{-\frac{4p}{3(p+1)}+\zeta}\right)$.

B. Main Results: First Order Optimization

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

Theorem 3.2. *Consider algorithm* (26) *with step-sizes* $\alpha_k = \frac{\alpha_0}{k+1}$ *and* $\beta_k = \frac{\beta_0}{(k+1)^{1/2}}$*, where* $\beta_0 > 0$ *and* $\alpha_0 > 2/\mu$*. Further, let Assumptions 1-3 and B2 hold.*

1) Then, for each node i*'s solution estimate* $\mathbf{x}_i(k)$ *and the solution* $\mathbf{x}^\star$ *of problem* (1)*, ,* $\forall k \geq 0$ *there holds:*

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq 2M_k + \frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)}$$
$$+ 2Q_k + \frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)}, \tag{33}$$

where, $\Delta_{1,\infty} = 2\|\nabla F(\mathbf{x}(k))\|^2 + 4c_uq_\infty(N,\alpha_0) + 4N\sigma_1^2$ *and* $q_\infty(N,\alpha_0) = \mathbb{E}\left[\|\mathbf{x}(k_2) - \mathbf{x}^o\|^2\right] + \frac{\pi^2}{6}\alpha_0^2\left(2c_uN\|\mathbf{x}^o\|^2 + N\sigma_u^2\right) + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}$*,* $k_2 = \max\{k_0, k_1\}$*,* $k_0 = \inf\{k|\mu^2\alpha_k^2 < 1\}$ *and* $k_1 = \inf\left\{k|\frac{\mu}{2} > 2c_u\alpha_k\right\}$*, with* M_k *and* Q_k *decaying faster*

than the rest of the terms.

2) The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k}\zeta_t\right] = O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right),$$

leading to the following MSE-communication rate:

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^\star\|^2\right] = O\left(\mathcal{C}_k^{-\frac{4}{3}+\zeta}\right), \tag{34}$$

where ζ can be arbitrarily small.

We remark that the condition $\alpha_0 > 2/\mu$ can be relaxed to require only a positive α_0, in which case the rate becomes $O(\ln(k)/k)$, instead of $O(1/k)$. Also, to avoid large step-sizes at initial iterations for a large α_0, step-size α_k can be modified to $\alpha_k = \alpha_0/(k+k_0)$, for arbitrary positive constant k_0, and Theorem 3.2 continues to hold. Theorem 3.2 establishes the $O(1/k)$ MSE rate of convergence of algorithm (26); due to the assumed f_i's strong convexity, the theorem also implies that $\mathbb{E}\left[f(\mathbf{x}_i(k)) - f(\mathbf{x}^\star)\right] = O(1/k)$.

4. Simulations

In this section, we provide evaluations of the proposed algorithms on the Abalone dataset ([32]). To be specific, we consider ℓ_2-regularized empirical risk minimization for the Abalone dataset, where the regularization function is given by $\Psi_i(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. We consider a 10 node network for both the zeroth and first order optimization schemes. The Abalone dataset has 4177 data points out of which 577 data points are kept aside as the test set and the other 3600 is divided equally among the 10 nodes resulting in each node having 360 data points. For the zeroth order optimization, we compare the proposed undirected sequence of Laplacian constructions based optimization scheme and the static Laplacian (Benchmark) based optimization schemes. The benchmark scheme is characterized by the communication graph being static and thereby resulting agents connected through a link to exchange messages at all times. The data points at each node are sampled without replacement in a contiguous manner. The vectors $\mathbf{z}_{i,k}$s for evaluating directional derivatives were sampled from a normal distribution with identity covariance. Figure 1 compares the test error for the three aforementioned schemes, where it can be clearly observed that the test error is indistinguishable in terms of the number of iterations or equivalently in terms of the number of queries to the stochastic zeroth oracle. Figure 2 demonstrates the superiority the proposed algorithm in terms of the test error versus communication cost as compared to the benchmark as predicted by Theorem 3.1. For example, at the same relative test error level, the proposed algorithm uses up to 3x less number of transmissions as compared to the benchmark scheme. In Figure 3, the test error of the communication efficient first order optimization scheme is compared with the test error of the benchmark scheme which refers to the optimization scheme with the communication graph abstracted by a static Laplacian in terms of iterations or equivalently the number of queries per agent to the stochastic first order oracle, i.e., gradient evaluations. Figure 4 demonstrates the superiority of the proposed communication efficient first order optimization scheme in terms of the test error versus communication cost as compared to the benchmark as predicted by Theorem 3.2. For example, at the same relative test error level, the proposed algorithm uses up to 3x less number of transmissions as compared to the benchmark scheme.

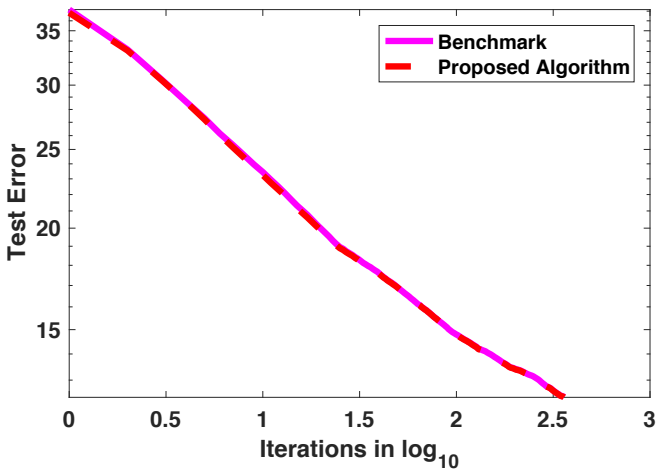

Fig. 1: Test Error vs Iterations

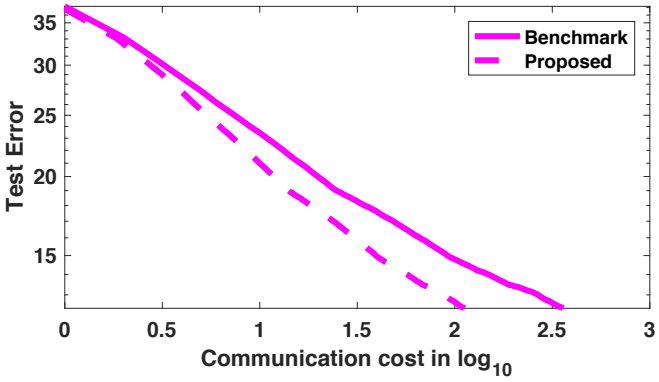

Fig. 2: Test Error vs Communication Cost

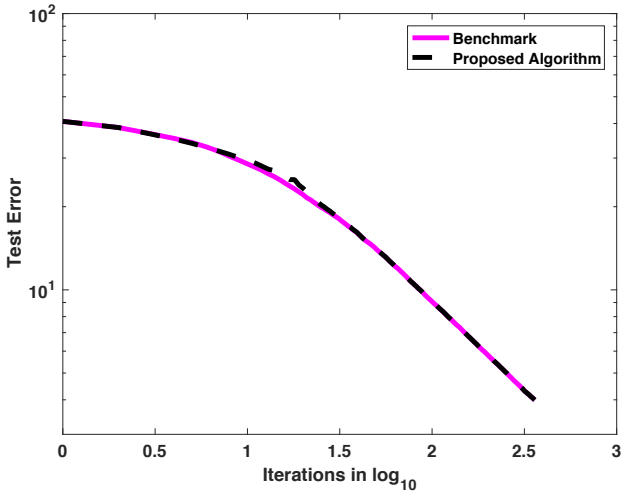

Fig. 3: Test Error vs Iteration

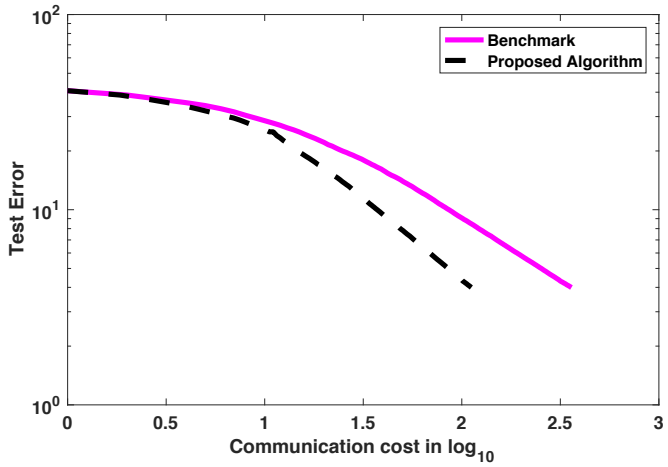

Fig. 4: Test Error vs Communication Cost

5. Proof of the main result: Zeroth order optimization

The proof of the main result proceeds through three main steps. The first step involves establishing the boundedness of the iterate sequence, while the second step involves establishing the convergence rate of the optimizer sequence at each agent to the network averaged optimizer sequence. The convergence of the network averaged optimizer is then analyzed as the final step and in combination with the second step results in the establishment of bounds on MSE of the optimizer sequence at each agent.

Lemma 5.1. *Let the hypotheses of Theorem 3.1 hold. Then, we have,*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}(k)-\mathbf{x}^o\|^2\right] \leq q_{k_2}(N,d,\alpha_0,c_0) + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}\\
&+\frac{\sqrt{N}s_1(P)M\alpha_0 c_0^2}{8\delta} + \frac{Ns_1^2(P)M^2\alpha_0^2c_0^4}{16(1+4\delta)}\\
&+\frac{d\alpha_0^2\left(2c_vN\|\mathbf{x}^o\|^2+N\sigma_v^2\right)}{c_0^2(1-2\delta)} + \frac{\alpha_0^2c_0^2\sqrt{N}s_1(P)L\|\nabla F(\mathbf{x}^o)\|}{1+2\delta}\\
&+\frac{N\alpha_0^2c_0^4s_2(P)}{1+4\delta} + \frac{4\alpha_0^2c_0^2Ns_1(P)}{1+2\delta}\|\nabla F(\mathbf{x}^o)\|^2\\
&\doteq q_\infty(N,d,\alpha_0,c_0),
\end{aligned}$$

where $\mathbb{E}\left[\|\mathbf{x}(k_2)-\mathbf{x}^o\|^2\right] \leq q_{k_2}(N,d,\alpha_0,c_0)$, $k_2=\max\{k_0,k_1\}$, $k_0=\inf\{k|\mu^2\alpha_k^2<1\}$ *and* $k_1=\inf\left\{k|\frac{\mu}{2} > \frac{\sqrt{N}}{4}s_1(P)Mc_k^2+\frac{2dc_v\alpha_k}{c_k^2}+4\alpha$

Proof.

$$\begin{aligned}
\mathbf{x}(k+1) &= \mathbf{W}_k\mathbf{x}(k)\\
&-\frac{\alpha_k}{c_k}\left(c_k\nabla F(\mathbf{x}(k)) + c_k^2\mathbf{b}(\mathbf{x}(k)) + c_k\mathbf{h}(\mathbf{x}(k))\right).
\end{aligned} \tag{35}$$

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

$$\begin{aligned}
\mathbf{x}(k+1)-\mathbf{x}^o &= \mathbf{W}_k(\mathbf{x}(k)-\mathbf{x}^o)\\
&-\alpha_k\left(\nabla F(\mathbf{x}(k))-\nabla F(\mathbf{x}^o)\right)
\end{aligned}$$

$$- \alpha_k \mathbf{h}(\mathbf{x}(k)) - \alpha_k \nabla F(\mathbf{x}^o) - \alpha_k c_k \mathbf{b}(\mathbf{x}(k)). \tag{36}$$

Moreover, note that, $\mathbb{E}\left[\mathbf{h}(\mathbf{x}(k)) \mid \mathcal{F}_k\right] = 0$. By Leibnitz rule, we have,

$$\begin{aligned}
&\nabla F(\mathbf{x}(k)) - \nabla F(\mathbf{x}^o) \\
&= \left[\int_{s=0}^{1} \nabla^2 F\left(\mathbf{x}^o + s(\mathbf{x}(k) - \mathbf{x}^o)\right) ds\right] (\mathbf{x}(k) - \mathbf{x}^o) \\
&= \mathbf{H}_k \left(\mathbf{x}(k) - \mathbf{x}^o\right). \qquad (37)
\end{aligned}$$

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 $\boldsymbol{\xi}(k) = \left(\mathbf{W}_k - \alpha_k \mathbf{H}_k\right)\left(\mathbf{x}(k) - \mathbf{x}^o\right) - \alpha_k \nabla F(\mathbf{x}^o)$. Then, there holds:

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2 \,|\, \mathcal{F}_k\,] \leq \mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2 | \mathcal{F}_k\right] \\
&- 2\alpha_k c_k \,\mathbb{E}\left[\boldsymbol{\xi}(k)^\top | \,\mathcal{F}_k\right] \mathbb{E}[\,\mathbf{h}(\mathbf{x}(k)) \,|\, \mathcal{F}_k\,] + \alpha_k^2 c_k^2 \,\mathbb{E}[\,\|\mathbf{h}(\mathbf{x}(k))\|^2 \,|\, \mathcal{F}_k\,] \\
&+ \alpha_k^2 c_k^2 \mathbf{b}^\top(\mathbf{x}(k))\mathbf{b}(\mathbf{x}(k)) - 2\alpha_k c_k \mathbf{b}^\top(\mathbf{x}(k)) \mathbb{E}\left[\boldsymbol{\xi}(k) | \mathcal{F}_k\right] \\
&+ \mathbf{b}\left(\mathbf{x}(k)\right)^\top \mathbb{E}\left[\mathbf{h}(\mathbf{x}(k)) | \mathcal{F}_k\right]. \qquad (38)
\end{aligned}$$

We use the following inequalities:

$$\begin{aligned}
&c_k \mathbf{b}(\mathbf{x}_i(k)) \\
&= \frac{c_k}{2} \mathbb{E}\left[\langle \mathbf{z}_{i,k}, \nabla^2 f_i \left(\mathbf{x}_i(k) + \frac{(1-\theta_1)}{2} c_k \mathbf{z}_{i,k}\right) \mathbf{z}_{i,k}\rangle \mathbf{z}_{i,k} | \mathcal{F}_k\right] \\
&- \frac{c_k}{2} \mathbb{E}\left[\langle \mathbf{z}_{i,k}, \nabla^2 f_i \left(\mathbf{x}_i(k) + (1-\theta_2)\, c_k \mathbf{z}_{i,k}\right) \mathbf{z}_{i,k}\rangle \mathbf{z}_{i,k} | \mathcal{F}_k\right] \\
&\Rightarrow c_k \left\|\mathbf{b}(\mathbf{x}_i(k))\right\| \leq \frac{c_k^2}{4} M s_1(P). \qquad (39)
\end{aligned}$$

$$\begin{aligned}
&- \mathbf{b}^\top(\mathbf{x}(k)) \mathbb{E}\left[\boldsymbol{\xi}(k) | \mathcal{F}_k\right] \\
&= -2\mathbf{b}^\top(\mathbf{x}(k)) \left(\mathbf{I} - \beta_k \overline{\mathbf{R}} - \alpha_k \mathbf{H}_k\right) \left(\mathbf{x}(k) - \mathbf{x}^o\right) \\
&+ 2\alpha_k \mathbf{b}^\top(\mathbf{x}(k)) \nabla F(\mathbf{x}^o) \\
&\leq 2 \left\|\mathbf{b}(\mathbf{x}(k))\right\| \left\|\mathbf{I} - \beta_k \overline{\mathbf{R}} - \alpha_k \mathbf{H}_k\right\| \left\|\mathbf{x}(k) - \mathbf{x}^o\right\| \\
&+ 2\alpha_k \left\|\mathbf{b}(\mathbf{x}(k))\right\| \left\|\nabla F(\mathbf{x}^o)\right\| \\
&\leq \frac{\sqrt{N}}{4} s_1(P) M c_k \left(1 - \mu\alpha_k\right) \left(1 + \left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2\right) \\
&+ \alpha_k c_k \frac{\sqrt{N}}{2} s_1(P) M \left\|\nabla F(\mathbf{x}^o)\right\| \\
&\leq \frac{\sqrt{N}}{4} s_1(P) M c_k + \frac{\sqrt{N}}{4} s_1(P) M c_k \left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2 \\
&+ \alpha_k c_k \frac{\sqrt{N}}{2} s_1(P) M \left\|\nabla F(\mathbf{x}^o)\right\|, \qquad (40)
\end{aligned}$$

$$\mathbf{b}^\top(\mathbf{x}(k))\mathbf{b}(\mathbf{x}(k)) \leq \frac{N}{16} s_1^2(P) M^2 c_k^2, \tag{41}$$

$$\begin{aligned}
&\mathbb{E}[\,\|\mathbf{h}(\mathbf{x}(k))\|^2 \,|\, \mathcal{F}_k\,] = \mathbb{E}\left[\|\mathbf{v}_{\mathbf{z}}(k; \mathbf{x}(k))\|^2 \,| \mathcal{F}_k\right] \\
&+ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}(k)) - \mathbb{E}\left[\widehat{\mathbf{g}}(\mathbf{x}(k)) \mid \mathcal{F}_k\right]\|^2 \mid \mathcal{F}_k\right], \qquad (42)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}(k)) - \mathbb{E}\left[\widehat{\mathbf{g}}(\mathbf{x}(k)) \mid \mathcal{F}_k\right]\|^2 \mid \mathcal{F}_k\right] \\
&\leq \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}(k))\|^2 \mid \mathcal{F}_k\right] \\
&\leq 4Ns_1(P)L^2\left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2 + 4Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2 + 2Nc_k^2 s_2(P), \qquad (43)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{v}_{\mathbf{z}}(k;\mathbf{x}(k))\|^2 |\mathcal{F}_k\right] \leq dc_v\left\|\mathbf{x}(k)\right\|^2 + dN\sigma_v^2 \\
&\leq 2dc_v\left\|\mathbf{x}(k) - \mathbf{x}^o\right\|^2 + \left(2dc_v\left\|\mathbf{x}^o\right\|^2 + N\sigma_v^2\right). \qquad (44)
\end{aligned}$$

Then from (38), we have,

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2 \,|\, \mathcal{F}_k\,] \leq \mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2 | \mathcal{F}_k\right] \\
&+ \frac{\sqrt{N}}{4}s_1(P)M\alpha_k c_k^2\|\boldsymbol{\zeta}(k)\|^2 + 2\frac{d\alpha_k^2}{c_k^2}c_v\|\boldsymbol{\zeta}(k)\|^2 \\
&+ \frac{d\alpha_k^2}{c_k^2}\left(2c_v\left\|\mathbf{x}^o\right\|^2 + N\sigma_v^2\right) + \frac{\sqrt{N}}{4}s_1(P)M\alpha_k c_k^2 \\
&+ \frac{N}{16}s_1^2(P)M^2\alpha_k^2 c_k^4 + \alpha_k^2 c_k^2\frac{\sqrt{N}}{2}s_1(P)M\left\|\nabla F(\mathbf{x}^o)\right\| \\
&+ 4\alpha_k^2 c_k^2 Ns_1(P)L^2\|\boldsymbol{\zeta}(k)\|^2 + 4\alpha_k^2 c_k^2 Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2 \\
&+ 2N\alpha_k^2 c_k^4 s_2(P). \qquad (45)
\end{aligned}$$

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:

$$\|\boldsymbol{\xi}(k)\| \leq (1 - \mu\,\alpha_k)\,\|\boldsymbol{\zeta}(k)\| + \alpha_k\,\|\nabla F(\mathbf{x}^o)\|. \qquad (46)$$

We now use the following inequality:

$$(a+b)^2 \leq (1+\theta)\,a^2 + \left(1 + \frac{1}{\theta}\right)b^2, \qquad (47)$$

for any $a, b \in \mathbb{R}$ and $\theta > 0$. We set $\theta = \mu\alpha_k$. Using the inequality (47) in (46) and we have $\forall k \geq k_0$, where $k_0 = \inf\{k|\mu^2\alpha_k^2 < 1\}$:

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{\xi}(k)\|^2 |\mathcal{F}_k\right] \leq (1+\mu\alpha_k)\,(1-\alpha_k\mu)^2\,\|\boldsymbol{\zeta}(k)\|^2 \\
&+ \left(1 + \frac{1}{\mu\alpha_k}\right)\alpha_k^2\|\nabla F(\mathbf{x}^o)\|^2 \\
&\leq (1-\alpha_k\mu)\,\|\boldsymbol{\zeta}(k)\|^2 + 2\frac{\alpha_k}{\mu}\|\nabla F(\mathbf{x}^o)\|^2. \qquad (48)
\end{aligned}$$

Using (48) in (45), we have for all $k \geq k_0$

$$\begin{aligned}
&\mathbb{E}[\,\|\boldsymbol{\zeta}(k+1)\|^2 \,|\, \mathcal{F}_k\,] \\
&\leq \Big(1 - \alpha_k\mu + \frac{\sqrt{N}}{4}s_1(P)M\alpha_k c_k^2 + 2\frac{d\alpha_k^2}{c_k^2}c_v \\
&+4\alpha_k^2 c_k^2 Ns_1(P)L^2\Big) \times \|\boldsymbol{\zeta}(k)\|^2 \\
&+ \frac{d\alpha_k^2}{c_k^2}\left(2c_v\left\|\mathbf{x}^o\right\|^2 + N\sigma_v^2\right) + \frac{\sqrt{N}}{4}s_1(P)L\alpha_k c_k^2
\end{aligned}$$

$$+\frac{N}{16}s_1^2(P)M^2\alpha_k^2c_k^4+2\frac{\alpha_k}{\mu}\|\nabla F(\mathbf{x}^o)\|^2+2N\alpha_k^2c_k^4s_2(P)$$
$$+\alpha_k^2c_k^2\frac{\sqrt{N}}{2}s_1(P)M\left\|\nabla F(\mathbf{x}^o)\right\|+4\alpha_k^2c_k^2Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2. \tag{49}$$

Define k_1 as follows:

$$k_1=\inf\left\{k|\frac{\mu}{2}>\frac{\sqrt{N}}{4}s_1(P)Mc_k^2+\frac{2dc_v\alpha_k}{c_k^2}+4\alpha_kc_k^2Ns_1(P)L^2\right\}.$$

It is to be noted that k_1 is necessarily finite as $c_k\to 0$ and $\alpha_kc_k^{-2}\to 0$ as $k\to\infty$. We proceed by using the following auxiliary lemma.

Lemma 5.2. *Let $a_k\in(0,1)$, $u\leq 0$ and $d_k\geq 0$, for all $k\geq 1$. If $q_{k_0}\geq 0$ and for all $k\geq k_0$ there holds $q_{k+1}\leq(1-a_k)q_k+a_ku+d_k$, then, for all $k\geq k_0$,*

$$q_{k+1}\leq q_{k_0}+u+\sum_{l=l_0}^{k}d_l. \tag{50}$$

Proof: Introduce $p(k,l)=(1-a_k)\cdots(1-a_l)$, for $l\leq k$ and also $p(k,k+1)=1$. It is easy to see that, for every $k\geq k_0$, $q_{k+1}\leq p(k,k_0)q_{k_0}+u\sum_{l=k_0}^{k}p(k,l+1)a_l+\sum_{l=k_0}^{k}p(k,l+1)d_l$. Note now that $p(k,l+1)a_l=p(k,l+1)-p(k,l)$, and hence $\sum_{l=k_0}^{k}p(k,l+1)a_l=1-p(k,k_0)\leq 1$. Using the latter together with the fact that $p(k,l+1)\leq 1$ proves the claim of the lemma. ■

Applying Lemma 5.2 to $q_k=\mathbb{E}\left[\|\boldsymbol{\zeta}(k)\|^2\right]$, $a_k=\frac{\mu\alpha_k}{2}$, $u=4\frac{\|\nabla F(x^o)\|^2}{\mu^2}$, and d_k defined as the remaining term in (49) we have, $\forall k\geq\max\{k_0,k_1\}\doteq k_2$,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\zeta}(k+1)\|^2\right]&\leq q_{k_2}(N,d,\alpha_0,c_0)+4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}\\&+\frac{\sqrt{N}s_1(P)M\alpha_0c_0^2}{8\delta}+\frac{Ns_1^2(P)M^2\alpha_0^2c_0^4}{16(1+4\delta)}\\&+\frac{d\alpha_0^2\left(2c_vN\left\|\mathbf{x}^o\right\|^2+N\sigma_v^2\right)}{c_0^2(1-2\delta)}+\frac{\alpha_0^2c_0^2\sqrt{N}s_1(P)L\left\|\nabla F(\mathbf{x}^o)\right\|}{1+2\delta}\\&+\frac{2N\alpha_0^2c_0^4s_2(P)}{1+4\delta}+\frac{4\alpha_0^2c_0^2Ns_1(P)}{1+2\delta}\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2\\&\doteq q_\infty(N,d,\alpha_0,c_0),\end{aligned} \tag{51}$$

□

From (51), we have that $\mathbb{E}\left[\left\|\mathbf{x}(k+1)-\mathbf{x}^o\right\|^2\right]$ is finite and bounded from above, where $\mathbb{E}\left[\left\|\mathbf{x}(k_2)-\mathbf{x}^o\right\|^2\right]\leq q_{k_2}(N,d,\alpha_0,c_0)$. 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 $\{\mathbf{v}_{\mathbf{z}}(k;\mathbf{x}(k))\}$ as follows:

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{v}_{\mathbf{z}}(k;\mathbf{x}(k))\right\|^2|\mathcal{F}_k\right]&\leq 2dc_vq_\infty(N,d,\alpha_0,c_0)\\&+2Nd\underbrace{\left(\sigma_v^2+\left\|\mathbf{x}^*\right\|^2\right)}_{\sigma_1^2}.\end{aligned} \tag{52}$$

We also have the following bound:

$$\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}(k))-\mathbb{E}\left[\widehat{\mathbf{g}}(\mathbf{x}(k))\mid\mathcal{F}_k\right]\right\|^2\mid\mathcal{F}_k\right]$$

$$\leq 4Ns_1(P)L^2q_\infty(N,d,\alpha_0,c_0) + 4Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2 + 2Nc_k^2 s_2(P).$$

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 5.3. *Let the hypotheses of Theorem 3.1 hold. Then, we have*

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

where Q_k is a term which decays faster than $(k+1)^{-2+2\tau+2\delta}$.

Lemma 5.3 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:

$$\widetilde{\mathbf{x}}(k+1) = \widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)$$
$$-\frac{\alpha_k}{c_k}(\mathbf{I}-\mathbf{J})\underbrace{\left(c_k\nabla F(\mathbf{x}(k)) + c_k\mathbf{h}(\mathbf{x}(k)) + c_k^2\mathbf{b}\left(\mathbf{x}(k)\right)\right)}_{\mathbf{w}(k)}, \tag{53}$$

where $\widetilde{\mathbf{W}}_k = \mathbf{W}_k - \mathbf{J}$. Then, we have,

$$\|\widetilde{\mathbf{x}}(k+1)\| \leq \left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\| + \frac{\alpha_k}{c_k}\|\mathbf{w}(k)\|. \tag{54}$$

Using (47) in (53), we have,

$$\|\widetilde{\mathbf{x}}(k+1)\|^2 \leq (1+\theta_k)\left\|\widetilde{\mathbf{W}}_k\widetilde{\mathbf{x}}(k)\right\|^2$$
$$+\left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\|\widetilde{\mathbf{w}}(k)\|^2. \tag{55}$$

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

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{W}}(k)\widetilde{\mathbf{x}}(k)\right\|^2|\mathcal{F}_k\right] = \widetilde{\mathbf{x}}^\top(k)\mathbb{E}\left[\widetilde{\mathbf{W}}^2(k) - \mathbf{J}|\mathcal{F}_k\right]\widetilde{\mathbf{x}}(k)$$
$$= \widetilde{\mathbf{x}}^\top(k)\left(\mathbf{I} - 2\beta_k\overline{\mathbf{R}} + \beta_k^2\overline{\mathbf{R}}^2 + \widetilde{\mathbf{R}}(k)^2 - \mathbf{J}\right)\widetilde{\mathbf{x}}(k)$$
$$\leq \left(1 - 2\beta_k\lambda_2\left(\overline{\mathbf{R}}\right) + \beta_k^2\lambda_N^2\left(\overline{\mathbf{R}}\right)\right.$$
$$\left.+\frac{4N^2\beta_0\rho_0^2}{(k+1)^{\tau+\epsilon}} - 4\beta_k^2N^2\right)\|\widetilde{\mathbf{x}}(k)\|^2$$
$$\leq \left(1 - 2\beta_k\lambda_2\left(\overline{\mathbf{R}}\right) + \frac{4N^2\beta_0\rho_0^2}{(k+1)^{\tau+\epsilon}}\right)\|\widetilde{\mathbf{x}}(k)\|^2$$
$$\leq \left(1 - \beta_k\lambda_2\left(\overline{\mathbf{R}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2, \tag{56}$$

where the last inequality follows from assumption A3. Then, we have,

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2|\mathcal{F}_k\right] \leq (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{R}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2 + \left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right], \tag{57}$$

where

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right] \leq 3c_k^2\|\nabla F(\mathbf{x}(k))\|^2 + 3c_k^2\mathbb{E}\left[\|\mathbf{h}(\mathbf{x}(k))\|^2|\mathcal{F}_k\right] \\
&+ 3c_k^2\|\mathbf{b}(\mathbf{x}(k))\|^2 \\
&\leq 3c_k^2\|\nabla F(\mathbf{x}(k))\|^2 + \frac{3}{16}c_k^4Ns_1^2(P)M^2 \\
&+ 6dc_vq_\infty(N,d,\alpha_0,c_0) + 6dN\sigma_1^2 + 6Nc_k^4s_2(P) \\
&+ 12c_k^2Ns_1(P)L^2q_\infty(N,d,\alpha_0,c_0) + 12c_k^2Ns_1(P)\|\nabla F(\mathbf{x}^o)\|^2 \\
&\Rightarrow \mathbb{E}\left[\|\mathbf{w}(k)\|^2\right] \leq 3\left(2dc_v + c_k^2L^2(1+4Ns_1(P))\right) \\
&\times q_\infty(N,d,\alpha_0,c_0) \\
&+ \frac{3}{16}c_k^4Ns_1^2(P)M^2 + 6Nc_k^4s_2(P) \\
&+ 6dN\sigma_1^2 + 12c_k^2Ns_1(P)\|\nabla F(\mathbf{x}^o)\|^2 \\
&= \Delta_{1,\infty} + c_k^2\Delta_{2,\infty} \doteq \Delta_k \\
&\Rightarrow \mathbb{E}\left[\|\mathbf{w}(k)\|^2\right] < \infty,
\end{aligned} \tag{58}$$

where $\Delta_{1,\infty} = 6dc_vq_\infty(N,d,\alpha_0,c_0) + 6dN\sigma_1^2$ and $c_k^2\Delta_{2,\infty} = \frac{3}{16}c_k^4Ns_1^2(P)M^2 + 3c_k^2L^2(1+4Ns_1(P))q_\infty(N,d,\alpha_0,c_0) + 12c_k^2Ns_1(P)\|\nabla F(\mathbf{x}^o)\|^2 + 6Nc_k^4s_2(P)$. With the above development in place, we then have,

$$\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] \leq (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{R}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2 + \left(1+\frac{1}{\theta_k}\right)\frac{\alpha_k^2}{c_k^2}\Delta_k. \tag{59}$$

In particular, we choose $\theta(k) = \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)$. From (59), we have,

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq \left(1-\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right] \\
&+ \left(1+\frac{2}{\beta_k\lambda_2\left(\overline{\mathbf{R}}\right)}\right)\frac{\alpha_k^2}{c_k^2}\Delta_k \\
&= \left(1-\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right] + \frac{2\alpha_k^2}{\lambda_2\left(\overline{\mathbf{R}}\right)c_k^2\beta_k}\Delta_k + \frac{\alpha_k^2}{c_k^2}\Delta_k.
\end{aligned} \tag{60}$$

For ease of analysis, define $s(k) = \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)$. We proceed by using the following technical lemma.

Lemma 5.4. *If for all $k \geq k_0$ there holds*

$$q_{k+1} \leq (1-s_k)q_k + \left(1+\frac{1}{s_k}\right)b_kd_k, \tag{61}$$

where $q_{k_0} \geq 0$, $s_k \in (0,1)$, d_k, $b_k \geq 0$ are monotonously decreasing, then, for any $k \geq m(k) \geq k_0$

$$\begin{aligned}
q_{k+1} &\leq e^{-\sum_{l=k_0}^k s_l}q_{k_0} + d_{k_0}e^{-\sum_{l=m(k)}^k s_l}\sum_{l=k_0}^{m(k)-1}\left(1+\frac{1}{s_l}\right)b_l \\
&+ d_{m(k)}b_{m(k)}\frac{s_k+1}{s_k^2}.
\end{aligned} \tag{62}$$

Proof: Similarly as before, define $p(k,l) = (1-s_k)\cdots(1-s_l)$ for $k_0 \le l \le k$, and let also $p(k,k+1) = 1$. Recall that $p(k,l+1)s_l$ can be expressed as $p(k,l+1)s_l = p(k,l+1) - p(k,l)$. Then, we have:

$$\begin{aligned} q_{k+1} &\le p(k,k_0)q_{k_0} + \sum_{l=k_0}^{k} p(k,l)\left(1 + \frac{1}{s_l} b_l d_l\right) \\ &\le p(k,k_0)q_{k_0} + d_{k_0} p(k,m(k)) \sum_{l=k_0}^{m(k)} \left(1 + \frac{1}{s_l}\right) b_l \\ &\quad + b_{m(k)} d_{m(k)} \frac{s_k + 1}{s_k^2} \sum_{m(k)}^{k} \left(p(k,l+1) - p(k,l)\right), \end{aligned} \tag{63}$$

where we break the sum in (63) at $l = m(k)$, and use the fact that $p\left(k, m(k)-1\right) \ge p(k,l)$ for every $l \le m(k)-1$, together with the fact that $1/s_l \le 1/s_k$, for every $l \le k$. Finally, noting that, for every $l \le k$, $p(k,l) \le e^{-\sum_{m=1}^{k} s_l}$, and also recalling that $\sum_{m(k)}^{k} \left(p(k,l+1) - p(k,l)\right) \le 1$, proves the claim of the lemma. ■

Applying the preceding lemma to $q_k = \mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(k)\right\|^2\right]$, $d_k = \Delta_k$, $b_k = \frac{\alpha_k^2}{c_k^2}$, and $s_k = \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)$ we have,

$$\begin{aligned} &\mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(k+1)\right\|^2\right] \\ &\le \underbrace{\exp\left(-\sum_{l=0}^{k} s(l)\right) \mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(0)\right\|^2\right]}_{t_1} \\ &+ \underbrace{\Delta_0 \exp\left(-\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k} s(m)\right) \sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor - 1} \left(\frac{2\alpha_l^2}{\lambda_2\left(\overline{\mathbf{R}}\right) c_l^2 \beta_l} + \frac{\alpha_l^2}{c_l^2}\right)}_{t_2} \\ &+ \underbrace{\frac{4\Delta_{\lfloor\frac{k-1}{2}\rfloor}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2 c_0^2 (k+1)^{2-2\tau-2\delta}}}_{t_3} + \underbrace{\frac{2\Delta_{\lfloor\frac{k-1}{2}\rfloor}\alpha_0^2}{\lambda_2\left(\overline{\mathbf{R}}\right)\beta_0 c_0^2 (k+1)^{2-\tau-2\delta}}}_{t_4}. \end{aligned} \tag{64}$$

In the above proof, the splitting in the interval $[0,k]$ was done at $\lfloor\frac{k-1}{2}\rfloor$ for ease of book keeping. The division can be done at an arbitrary point. It is to be noted that the sequence $\{s(k)\}$ is not summable and hence terms t_1 and t_2 decay faster than $(k+1)^{2-2\tau-2\delta}$. Also, note that term t_4 decays faster than t_3. Furthermore, t_3 can be written as

$$\begin{aligned} \frac{4\Delta_{\lfloor\frac{k-1}{2}\rfloor}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2 c_0^2 (k+1)^{2-2\tau-2\delta}} &= \underbrace{\frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2 c_0^2 (k+1)^{2-2\tau-2\delta}}}_{t_{31}} \\ &+ \underbrace{\frac{4c^2_{\lfloor\frac{k-1}{2}\rfloor}\Delta_{2,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2 c_0^2 (k+1)^{2-2\tau-2\delta}}}_{t_{32}}, \end{aligned}$$

from which we have that t_{32} decays faster than t_{31}. For notational ease, henceforth we refer to $t_1 + t_2 + t_{32} + t_4 = Q_k$, while keeping in mind that Q_k decays faster than $(k+1)^{2-2\tau-2\delta}$. Hence, we have the disagreement given by,

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

□

We now proceed to the proof of Theorem 3.1. Denote $\overline{\mathbf{x}}(k) = \frac{1}{N}\sum_{n=1} \mathbf{x}_i(k)$. Then, we have,

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k)$$

$$
\begin{aligned}
&- \frac{\alpha_k}{c_k}\Bigg[\frac{c_k}{N}\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right) + \underbrace{\frac{c_k^2}{N}\sum_{i=1}^{N}\mathbf{b}_i\left(\mathbf{x}_i(k)\right)}_{\overline{\mathbf{b}}\left(\mathbf{x}(k)\right)} \\
&+ \underbrace{\frac{c_k}{N}\sum_{i=1}^{N}\mathbf{h}_i(\mathbf{x}_i(k))}_{\overline{\mathbf{h}}(\mathbf{x}(k))}\Bigg] \\
&\Rightarrow \overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{c_k}\left(\overline{\mathbf{h}}(\mathbf{x}(k)) + \overline{\mathbf{b}}\left(\mathbf{x}(k)\right)\right) \\
&- \frac{\alpha_k}{Nc_k}\left[c_k\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right) + \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right]. \qquad (65)
\end{aligned}
$$

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

$$
\begin{aligned}
&\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{c_k}\left(\overline{\mathbf{h}}(\mathbf{x}(k)) + \overline{\mathbf{b}}\left(\mathbf{x}(k)\right)\right) \\
&- \frac{\alpha_k}{N}\nabla f\left(\overline{\mathbf{x}}(k)\right) - \frac{\alpha_k}{N}\left[\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right] \\
&\Rightarrow \overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{Nc_k}\left[c_k\nabla f\left(\overline{\mathbf{x}}(k)\right) + \mathbf{e}(k)\right], \qquad (66)
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathbf{e}(k) = N\overline{\mathbf{h}}(\mathbf{x}(k)) \\
&\underbrace{+ N\overline{\mathbf{b}}\left(\mathbf{x}(k)\right) + c_k\sum_{i=1}^{N}\left(\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right)}_{\boldsymbol{\epsilon}(k)}. \qquad (67)
\end{aligned}
$$

Note that, $c_k\left\|\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right\| \leq c_k L\left\|\mathbf{x}_i(k) - \overline{\mathbf{x}}(k)\right\| = c_k L\left\|\widetilde{\mathbf{x}}_i(k)\right\|$. We also have that, $\left\|\overline{\mathbf{b}}\left(\mathbf{x}(k)\right)\right\| \leq \frac{M}{4}s_1(P)c_k^3$. Thus, we can conclude that, $\forall k \geq k_3$

$$
\begin{aligned}
&\boldsymbol{\epsilon}(k) = c_k\sum_{i=1}^{N}\left(\nabla f_i\left(\mathbf{x}_i(k)\right) - \nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right) + N\overline{\mathbf{b}}\left(\mathbf{x}(k)\right) \\
&\Rightarrow \left\|\boldsymbol{\epsilon}(k)\right\|^2 \leq 2NL^2c_k^2\left\|\widetilde{\mathbf{x}}(k)\right\|^2 + \frac{N}{8}M^2d^2(P)c_k^6 \\
&\Rightarrow \mathbb{E}\left[\left\|\boldsymbol{\epsilon}(k)\right\|^2\right] \leq \frac{8NL^2\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)^{2-2\tau}} + \frac{NM^2d^2(P)c_0^6}{8(k+1)^{6\delta}} \\
&+ \frac{2NL^2Q_kc_0^2}{(k+1)^{2\delta}}. \qquad (68)
\end{aligned}
$$

With the above development in place, we rewrite (66) as follows:

$$
\begin{aligned}
&\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\alpha_k}{N}\nabla f\left(\overline{\mathbf{x}}(k)\right) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{h}}(\mathbf{x}(k)) \\
&\Rightarrow \overline{\mathbf{x}}(k+1) - \mathbf{x}^* = \overline{\mathbf{x}}(k) - \mathbf{x}^* - \frac{\alpha_k}{N}\left[\nabla f\left(\overline{\mathbf{x}}(k)\right) - \underbrace{\nabla f\left(\mathbf{x}^*\right)}_{=0}\right] \\
&- \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{h}}(\mathbf{x}(k)). \qquad (69)
\end{aligned}
$$

By Leibnitz rule, we have,

$$
\nabla f\left(\overline{\mathbf{x}}(k)\right) - \nabla f\left(\mathbf{x}^*\right)
$$

$$= \underbrace{\left[\int_{s=0}^{1} \nabla^2 f\left(\mathbf{x}^* + s\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right)\right) ds\right]}_{\overline{\mathbf{H}}_k} \left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right), \tag{70}$$

where it is to be noted that $NL \succcurlyeq \overline{\mathbf{H}}_k \succcurlyeq N\mu$. Using (70) in (69), we have,

$$\begin{aligned}\left(\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right) &= \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right]\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right) \\ &- \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{h}}(\mathbf{x}(k)).\end{aligned} \tag{71}$$

Denote by $\mathbf{m}(k) = \left[\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right]\left(\overline{\mathbf{x}}(k) - \mathbf{x}^*\right) - \frac{\alpha_k}{Nc_k}\boldsymbol{\epsilon}(k)$ and note that $\mathbf{m}(k)$ is conditionally independent from $\overline{\mathbf{h}}(\mathbf{x}(k))$ given the history $\mathcal{F}_k$. Then (71) can be rewritten as:

$$\begin{aligned}&\left(\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right) = \mathbf{m}(k) - \frac{\alpha_k}{c_k}\overline{\mathbf{h}}(\mathbf{x}(k)) \\ &\Rightarrow \left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2 \leq \|\mathbf{m}(k)\|^2 - 2\frac{\alpha_k}{c_k}\mathbf{m}(k)^\top\overline{\mathbf{h}}(\mathbf{x}(k)) \\ &+ \frac{\alpha_k^2}{c_k^2}\left\|\overline{\mathbf{h}}(\mathbf{x}(k))\right\|^2.\end{aligned} \tag{72}$$

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

$$\begin{aligned}&\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2|\mathcal{F}_k\right] \leq \|\mathbf{m}(k)\|^2 + \frac{\alpha_k^2}{c_k^2}\mathbb{E}\left[\left\|\overline{\mathbf{h}}(\mathbf{x}(k))\right\|^2|\mathcal{F}_k\right] \\ &\Rightarrow \mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2\right] \leq \mathbb{E}\left[\|\mathbf{m}(k)\|^2\right] + 2N\alpha_k^2c_k^2s_2(P) \\ &+ \frac{2\alpha_k^2\left(dc_vq_\infty(N,d,\alpha_0,c_0) + dN\sigma_1^2\right)}{c_k^2} \\ &+ 4\alpha_k^2Ns_1(P)L^2q_\infty(N,d,\alpha_0,c_0) + 4\alpha_k^2Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2.\end{aligned} \tag{73}$$

For notational simplicity, we denote $\alpha_k^2\sigma_h^2 = 2N\alpha_k^2c_k^2s_2(P) + 4\alpha_k^2Ns_1(P)L^2q_\infty(N,d,\alpha_0,c_0) + 4\alpha_k^2Ns_1(P)\left\|\nabla F\left(\mathbf{x}^o\right)\right\|^2$. Using (47), we have for $\mathbf{m}(k)$,

$$\begin{aligned}\|\mathbf{m}(k)\|^2 &\leq (1+\theta_k)\left\|\mathbf{I} - \frac{\alpha_k}{N}\overline{\mathbf{H}}_k\right\|^2\left\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\right\|^2 \\ &+ \left(1 + \frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2c_k^2}\|\boldsymbol{\epsilon}(k)\|^2 \\ &\leq (1+\theta_k)\left(1 - \frac{\mu\alpha_0}{k+1}\right)^2\left\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\right\|^2 \\ &+ \left(1 + \frac{1}{\theta_k}\right)\frac{\alpha_k^2}{N^2c_k^2}\|\boldsymbol{\epsilon}(k)\|^2.\end{aligned} \tag{74}$$

On choosing $\theta_k = \frac{\mu\alpha_0}{k+1}$, where $\alpha_0 > \frac{1}{\mu}$, we have,

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{m}(k)\|^2\right] \leq \left(1 - \frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\right\|^2\right] \\ &+ \frac{16L^2\Delta_{1,\infty}N\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{R}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}} + \frac{4M^2Nd^2(P)c_0^4\alpha_0}{\mu(k+1)^{1+4\delta}} + \frac{4L^2NQ_k}{\mu(k+1)} \\ &\Rightarrow \mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2\right] \leq \left(1 - \frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\right\|^2\right] \\ &+ \frac{16NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{R}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}} + \frac{4NM^2d^2(P)c_0^4\alpha_0}{\mu(k+1)^{1+4\delta}} + \frac{4NL^2Q_k}{\mu(k+1)} \\ &+ \frac{2\alpha_k^2\left(dc_vq_\infty(N,d,\alpha_0,c_0) + dN\sigma_1^2\right)}{c_k^2} + \alpha_k^2\sigma_h^2 \\ &\Rightarrow \mathbb{E}\left[\left\|\overline{\mathbf{x}}(k+1) - \mathbf{x}^*\right\|^2\right] \leq \left(1 - \frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\left\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\right\|^2\right]\end{aligned}$$

$$+\frac{16NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{R}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}$$
$$+\frac{4M^2Nd^2(P)c_0^4\alpha_0}{\mu(k+1)^{1+4\delta}}+\frac{2\alpha_0^2\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{c_0^2(k+1)^{2-2\delta}}+P_k, \tag{75}$$

where $P_k=\frac{4NL^2Q_k}{\mu(k+1)}+\frac{\alpha_0^2\sigma_h^2}{(k+1)^2}$ decays faster as compared to the other terms. Proceeding as in (64), we have

$$\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right]$$
$$\leq\underbrace{\exp\left(-\mu\sum_{l=0}^{k}\alpha_l\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]}_{t_6}$$
$$+\underbrace{\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{16NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{R}}\right)c_0^2\beta_0^2(k+1)^{3-2\tau-2\delta}}}_{t_7}$$
$$+\underbrace{\exp\left(-\frac{\mu}{N}\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=k_5}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{4M^2Nd^2(P)c_0^4\alpha_0}{\mu(k+1)^{1+4\delta}}}_{t_8}$$
$$+\underbrace{\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}-1}P_l+\frac{2\alpha_0^2dc_vq_\infty(N,d,\alpha_0,c_0)}{c_0^2(l+1)^{2-2\delta}}}_{t_{10}}$$
$$+\underbrace{\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}-1}\frac{2\alpha_0^2dN\sigma_1^2}{c_0^2(l+1)^{2-2\delta}}}_{t_{11}}$$
$$+\underbrace{\frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{R}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau-2\delta}}}_{t_{12}}$$
$$+\underbrace{\frac{8NM^2d^2(P)c_0^4}{\mu^2(k+1)^{4\delta}}}_{t_{13}}+\underbrace{\frac{N(k+1)P_k}{\mu\alpha_0}}_{t_{14}}$$
$$+\underbrace{\frac{4N\alpha_0\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}}}_{t_{15}}. \tag{76}$$

It is to be noted that the term t_6 decays as $1/k$. The terms t_7, t_8, t_{10}, t_{11} and t_{14} decay faster than its counterparts in the terms t_{12}, t_{13} 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 $\delta_1=\min\{1-2\delta,2-2\tau-2\delta,4\delta\}$. For notational ease, we refer to $t_6+t_7+t_8+t_{10}+t_{11}+t_{14}=M_k$ from now on. Finally, we note that,

$$\|\mathbf{x}_i(k)-\mathbf{x}^*\|\leq\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|+\left\|\underbrace{\mathbf{x}_i(k)-\overline{\mathbf{x}}(k)}_{\widetilde{\mathbf{x}}_i(k)}\right\|$$
$$\Rightarrow\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\leq2\|\widetilde{\mathbf{x}}_i(k)\|^2+2\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2$$
$$\Rightarrow\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]\leq2M_k+\frac{64NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{R}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau-2\delta}}$$

$$+\frac{16NM^2d^2(P)c_0^4}{\mu^2(k+1)^{4\delta}}+2Q_k+\frac{8\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}}$$
$$+\frac{4N\alpha_0\left(dc_vq_\infty(N,d,\alpha_0,c_0)+dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}}$$
$$\Rightarrow\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]=O\left(\frac{1}{k^{\delta_1}}\right),\ \forall i, \tag{77}$$

where $\delta_1=\min\{1-2\delta,2-2\tau-2\delta,4\delta\}$. By, optimizing over τ and δ, we obtain that for $\tau=1/2$ and $\delta=1/6$,

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^*\|^2\right]=O\left(\frac{1}{k^{\frac{2}{3}}}\right),\ \forall i.$$

The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k}\zeta_t\right]=O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right).$$

Thus, we achieve the communication rate to be,

$$\mathbb{E}\left[\|\mathbf{x}_i(k)-\mathbf{x}^\star\|^2\right]=O\left(\frac{1}{\mathcal{C}_k^{8/9-\zeta}}\right), \tag{78}$$

where ζ can be arbitrarily small.

6. Proof of the main result: First order optimization

Lemma 6.1. *Consider algorithm* (26)*, and let the hypotheses of Theorem 3.2 hold. Then, we have that for all* $k=0,1,...,$ *there holds:*

$$\mathbb{E}\left[\|\mathbf{x}(k)-\mathbf{x}^o\|^2\right]\leq q_{k_0}(N,\alpha_0)$$
$$+\frac{\pi^2}{6}\alpha_0^2\left(2c_uN\|\mathbf{x}^o\|^2+N\sigma_u^2\right)+4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}\doteq q_\infty(N,\alpha_0),$$

where $\mathbb{E}\left[\|\mathbf{x}(k_2)-\mathbf{x}^o\|^2\right]\leq q_{k_2}(N,\alpha_0)$*,* $k_2=\max\{k_0,k_1\}$*,* $k_0=\inf\{k|\mu^2\alpha_k^2<1\}$ *and* $k_1=\inf\left\{k|\frac{\mu}{2}>2c_u\alpha_k\right\}$.

Proof: Proceeding as in the proof of Lemma 5.1, with $c_k=1$ and $\mathbf{b}(\mathbf{x}(k))=0$, we have that, $\forall k\geq\max\{k_0,k_1\}$,

$$\mathbb{E}\left[\|\boldsymbol{\zeta}(k+1)\|^2\right]\leq\prod_{l=k_0}^{k}\left(1-\frac{\mu\alpha_l}{2}\right)\mathbb{E}\left[\|\boldsymbol{\zeta}(k_0)\|^2\right]$$
$$+\frac{\pi^2}{6}\alpha_0^2\left(2c_uN\|\mathbf{x}^o\|^2+N\sigma_u^2\right)$$
$$+4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}$$
$$\mathbb{E}\left[\|\boldsymbol{\zeta}(k+1)\|^2\right]\leq q_{k_2}(N,\alpha_0)+\frac{\pi^2}{6}\alpha_0^2\left(2c_uN\|\mathbf{x}^o\|^2+N\sigma_u^2\right)$$
$$+4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2}$$
$$\doteq q_\infty(N,\alpha_0), \tag{79}$$

where $k_0=\inf\{k|\mu^2\alpha_k^2<1\}$ and

$$k_1=\inf\left\{k|\frac{\mu}{2}>2c_u\alpha_k\right\}.$$

and $k_2=\max\{k_0,k_1\}$. It is to be noted that k_1 is necessarily finite as $\alpha_k\to0$ as $k\to\infty$. Hence, we have that $\mathbb{E}\left[\|\mathbf{x}(k+1)-\mathbf{x}^o\|^2\right]$ is finite and bounded from above, where $\mathbb{E}\left[\|\mathbf{x}(k_2)-\mathbf{x}^o\|^2\right]\leq q_{k_2}(N,\alpha_0)$.

From the boundedness of $\mathbb{E}\left[\|\mathbf{x}(k)-\mathbf{x}^o\|^2\right]$, we have also established the boundedness of $\mathbb{E}\left[\|\nabla F(\mathbf{x}(k))\|^2\right]$ and $\mathbb{E}\left[\|\mathbf{x}(k)\|^2\right]$.

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

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{u}(k)\|^2|\mathcal{S}_k\right] &\leq 2c_u q_\infty(N,\alpha_0)\\ &+2N\underbrace{\left(\sigma_u^2+\|\mathbf{x}^*\|^2\right)}_{\sigma_1^2}.\end{aligned} \tag{80}$$

The proof of Lemma 6.1 is now complete. ■

Recall the (hypothetically available) global average of nodes' estimates $\overline{\mathbf{x}}(k)=\frac{1}{N}\sum_{i=1}^N\mathbf{x}_i(k)$, and denote by $\widetilde{\mathbf{x}}_i(k)=\mathbf{x}_i(k)-\overline{\mathbf{x}}(k)$ the quantity that measures how far apart is node i's solution estimate from the global average. Introduce also vector $\widetilde{\mathbf{x}}(k)=(\,\widetilde{\mathbf{x}}_1(k),...,\widetilde{\mathbf{x}}_N(k)\,)^\top$, and note that it can be represented as $\widetilde{\mathbf{x}}(k)=(\mathbf{I}-\mathbf{J})\,\mathbf{x}(k)$, where we recall $\mathbf{J}=\frac{1}{N}\mathbf{1}\mathbf{1}^\top$. We have the following Lemma.

Lemma 6.2. *Let the hypotheses of Theorem 3.2 hold. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq Q_k+\frac{2\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)}\\ &=O\left(\frac{1}{k}\right),\end{aligned}$$

where Q_k is a term which decays faster than $(k+1)^{-1}$.

Lemma 6.2 is important as it allows to sufficiently tightly bound the bias in the gradient estimates according to which the global average $\overline{\mathbf{x}}(k)$ evolves.

Proof: Proceeding as in the proof of Lemma 5.3 in (53)-(56), we have,

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2|\mathcal{S}_k\right] &\leq (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{R}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2\\ &+\left(1+\frac{1}{\theta_k}\right)\alpha_k^2\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{F}_k\right],\end{aligned} \tag{81}$$

where

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{w}(k)\|^2|\mathcal{S}_k\right]\leq 2\|\nabla F(\mathbf{x}(k))\|^2+2\mathbb{E}\left[\|\mathbf{v}(k)\|^2|\mathcal{F}_k\right]\\ &\leq\underbrace{2\|\nabla F(\mathbf{x}(k))\|^2+4c_u q_\infty(N,\alpha_0)+4N\sigma_1^2}_{\Delta_{1,\infty}}\\ &\Rightarrow\mathbb{E}\left[\|\mathbf{w}(k)\|^2\right]<\infty.\end{aligned} \tag{82}$$

With the above development in place, we then have,

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq (1+\theta_k)\left(1-\beta_k\lambda_2\left(\overline{\mathbf{R}}\right)\right)\|\widetilde{\mathbf{x}}(k)\|^2\\ &+\left(1+\frac{1}{\theta_k}\right)\alpha_k^2\Delta_{1,\infty}.\end{aligned} \tag{83}$$

In particular, we choose $\theta(k)=\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)$. From (59), we have,

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k+1)\|^2\right] &\leq\left(1-\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right]\\ &+\left(1+\frac{2}{\beta_k\lambda_2\left(\overline{\mathbf{R}}\right)}\right)\alpha_k^2\Delta_{1,\infty}\\ &=\left(1-\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)\right)\mathbb{E}\left[\|\widetilde{\mathbf{x}}(k)\|^2\right]\end{aligned}$$

$$+\frac{2\alpha_k^2}{\lambda_2\left(\overline{\mathbf{R}}\right)\beta_k}\Delta_{1,\infty}+\alpha_k^2\Delta_{1,\infty}. \tag{84}$$

Applying lemma 5.4 to $q_k=\mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(k)\right\|^2\right]$, $d_k=\Delta_k$, $b_k=\alpha_k^2$, and $s_k=\frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{R}}\right)$, we obtain for $m(k)=\lfloor\frac{k-1}{2}\rfloor$:

$$\begin{aligned}\mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(k+1)\right\|^2\right]&\leq\underbrace{\exp\left(-\sum_{l=0}^{k}s(l)\right)\mathbb{E}\left[\left\|\widetilde{\mathbf{x}}(0)\right\|^2\right]}_{t_1}\\&+\underbrace{\Delta_{1,\infty}\exp\left(-\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}s(m)\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\left(\frac{2\alpha_l^2}{\lambda_2\left(\overline{\mathbf{R}}\right)\beta_l}+\alpha_l^2\right)}_{t_2}\\&+\underbrace{\frac{2\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)}}_{t_3}+\underbrace{\frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2\left(\overline{\mathbf{R}}\right)\beta_0(k+1)^{3/2}}}_{t_4}.\end{aligned} \tag{85}$$

In the proof of Lemma 5.4, the splitting in the interval $[0,k]$ was done at $\lfloor\frac{k-1}{2}\rfloor$ for ease of book keeping. The division can be done at an arbitrary point. It is to be noted that the sequence $\{s(k)\}$ is not summable and hence terms t_1 and t_2 decay faster than $(k+1)$. Also, note that term t_4 decays faster than t_3. For notational ease, henceforth we refer to $t_1+t_2+t_4=Q_k$, while keeping in mind that Q_k decays faster than $(k+1)$. Hence, we have the disagreement given by,

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

■

Lemma 6.3. *Consider algorithm* (26) *and let the hypotheses of Theorem 3.2 hold. Then, there holds:*

$$\mathbb{E}[\,\|\mathbf{x}_i(k)-\mathbf{x}^\star\|^2\,]=O(1/k).$$

and

$$\mathbb{E}[\,\|\mathbf{x}_i(k)-\mathbf{x}^\star\|^2\,]=O\left(\frac{1}{\mathcal{C}_k^{4/3-\zeta}}\right),$$

where $\zeta>0$ can be arbitrarily small, for all $i=1,\cdots,N$.

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

$$\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k)-\alpha_k\left[\frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right)+\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i(k)}_{\overline{\mathbf{u}}(k)}\right] \tag{86}$$

which implies:

$$\begin{aligned}\overline{\mathbf{x}}(k+1)=\overline{\mathbf{x}}(k)-\frac{\alpha_k}{N}\Bigg[&\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_i(k)\right)\\&-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)+\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\Bigg]-\alpha_k\overline{\mathbf{u}}(k).\end{aligned}$$

where

$$\mathbf{e}(k)=N\overline{\mathbf{u}}(k)$$

$$+\underbrace{\sum_{i=1}^{N}\left(\nabla f_i\left(\mathbf{x}_i(k)\right)-\nabla f_i\left(\overline{\mathbf{x}}(k)\right)\right)}_{\boldsymbol{\epsilon}(k)}. \tag{87}$$

Proceeding as in (68)-(74), with $c_k = 1$ and $\mathbf{b}(\mathbf{x}_i(k)) = 0$, $\forall i = 1, \cdots, N$, we have on choosing $\theta_k = \frac{\mu\alpha_0}{k+1}$, where $\alpha_0 > \frac{1}{\mu}$,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{m}(k)\|^2\right] \leq \left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]\\
&+\frac{8NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)^2}+\frac{2NL^2Q_k}{\mu(k+1)}\\
&\Rightarrow \mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right] \leq \left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]\\
&+\frac{8NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)^2}+\frac{2NL^2Q_k}{\mu(k+1)}+2\alpha_k^2\left(c_u q_\infty(N,\alpha_0)+N\sigma_1^2\right)\\
&\Rightarrow \mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right] \leq \left(1-\frac{\mu\alpha_0}{k+1}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]\\
&+\frac{8NL^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)^2}+2\alpha_k^2\left(c_u q_\infty(N,\alpha_0)+N\sigma_1^2\right)+P_k,
\end{aligned} \tag{88}$$

where P_k decays faster as compared to the other terms. Proceeding as in (64), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right]\\
&\leq \underbrace{\exp\left(-\mu\sum_{l=0}^{k}\alpha_l\right)\mathbb{E}\left[\|\overline{\mathbf{x}}(k)-\mathbf{x}^*\|^2\right]}_{t_6}\\
&+\underbrace{\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{8L^2\Delta_{1,\infty}\alpha_0^3}{\mu\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(l+1)^2}}_{t_7}\\
&+\underbrace{\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}P_l}_{t_{10}}\\
&+\underbrace{\exp\left(-\mu\sum_{m=\lfloor\frac{k-1}{2}\rfloor}^{k}\alpha_m\right)\sum_{l=0}^{\lfloor\frac{k-1}{2}\rfloor-1}\frac{2\alpha_0^2\left(c_u q_\infty(N,\alpha_0)+N\sigma_1^2\right)}{(l+1)^2}}_{t_{11}}\\
&+\underbrace{\frac{16NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)}}_{t_{12}}\\
&+\underbrace{\frac{N(k+1)P_k}{\mu\alpha_0}}_{t_{14}}+\underbrace{\frac{4N\alpha_0\left(c_u q_\infty(N,\alpha_0)+N\sigma_1^2\right)}{\mu(k+1)}}_{t_{15}}.
\end{aligned} \tag{89}$$

It is to be noted that the term t_6 decays as $1/k$. The terms t_7, t_{10} and t_{11} decay faster than its counterparts in the terms t_{12} 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} and t_{15}. Thus, we have that, $\mathbb{E}\left[\|\overline{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2\right] = O\left(\frac{1}{k}\right)$. For notational ease,

we refer to $t_6 + t_7 + t_{10} + t_{11} + t_{14} = M_k$ from now on. Finally, we note that,

$$\begin{aligned}
&\|\mathbf{x}_i(k) - \mathbf{x}^*\| \leq \|\overline{\mathbf{x}}(k) - \mathbf{x}^*\| + \left\| \underbrace{\mathbf{x}_i(k) - \overline{\mathbf{x}}(k)}_{\widetilde{\mathbf{x}}_i(k)} \right\| \\
&\Rightarrow \|\mathbf{x}_i(k) - \mathbf{x}^*\|^2 \leq 2\|\widetilde{\mathbf{x}}_i(k)\|^2 + 2\|\overline{\mathbf{x}}(k) - \mathbf{x}^*\|^2 \\
&\Rightarrow \mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq 2M_k + \frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)} \\
&+ 2Q_k + \frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{R}}\right)\beta_0^2(k+1)} \\
&\Rightarrow \mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] = O\left(\frac{1}{k}\right),\ \forall i. \qquad (90)
\end{aligned}$$

The communication cost is given by,

$$\mathbb{E}\left[\sum_{t=1}^{k} \zeta_t\right] = O\left(k^{\frac{3}{4}+\frac{\epsilon}{2}}\right).$$

Thus, we achieve the communication rate to be,

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = O\left(\frac{1}{\mathcal{C}_k^{\frac{4}{3}-\zeta}}\right). \qquad (91)$$

■

7. Conclusion

In this paper, we have developed and analyzed a novel class of methods for distributed stochastic optimization of the zeroth and first order that are based on increasingly sparse randomized communication protocols. We have established for both the proposed zeroth and first order methods explicit mean square error (MSE) convergence rates with respect to (appropriately defined) computational cost C_{comp} and communication cost C_{comm}. Specifically, the proposed zeroth order method achieves the $O(1/(C_{\text{comm}})^{8/9-\zeta})$ MSE communication rate, which significantly improves over the rates of existing methods, while maintaining the order-optimal $O(1/(C_{\text{comp}})^{2/3})$ MSE computational rate. Similarly, the proposed first order method achieves the $O(1/(C_{\text{comm}})^{4/3-\zeta})$ MSE communication rate, while maintaining the order-optimal $O(1/(C_{\text{comp}}))$ MSE computational rate. Numerical examples on real data demonstrate the communication efficiency of the proposed methods.

References

[1] V. Vapnik, *Statistical learning theory. 1998*. Wiley, New York, 1998, vol. 3.

[2] A. K. Sahu, D. Jakovetic, and S. Kar, "Communication optimality trade-offs for distributed estimation," *arXiv preprint arXiv:1801.04050*, 2018.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning, Michael Jordan, Editor in Chief*, vol. 3, no. 1, pp. 1–122, 2011.

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

[5] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM J. Optim.*, vol. 26, no. 3, pp. 1835?–1854, 2016.

[6] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning — Part I: Algorithm development," 2017, arxiv preprint, arXiv:1702.05122.

[7] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Fast distributed gradient methods," *IEEE Trans. Autom. Contr.*, vol. 59, no. 5, pp. 1131–1146, May 2014.

[8] C. Xi and U. A. Khan, "DEXTRA: A fast algorithm for optimization over directed graphs," *IEEE Transactions on Automatic Control*, 2017, to appear, DOI: 10.1109/TAC.2017.2672698.

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

[10] Z. J. Towfic, J. Chen, and A. H. Sayed, "Excess-risk of distributed stochastic learners," *IEEE Transactions on Information Theory*, vol. 62, no. 10, Oct. 2016.

[11] 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.

[12] 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.

[13] 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.

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

[15] I. Lobel and A. E. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Trans. Automat. Contr.*, vol. 56, no. 6, pp. 1291–1306, Jan. 2011.

[16] I. Lobel, A. Ozdaglar, and D. Feijer, "Distributed multi-agent optimization with state-dependent communication," *Mathematical Programming*, vol. 129, no. 2, pp. 255–284, 2011.

[17] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Convergence rates of distributed Nesterov-like gradient methods on random networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 4, pp. 868–882, February 2014.

[18] 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.

[19] A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar, "Distributed zeroth order optimization over random networks: A Kiefer-Wolfowitz stochastic approximation approach," in *57th IEEE Conference on Decision and Control (CDC)*, Miami, 2018, available at https://www.dropbox.com/s/kfc2hgbfcx5yhr8/MainCDC2018KWSA.pdf.

[20] ——, "Non-asymptotic rates for communication efficient distributed zeroth order strongly convex optimization," 2018, available at https://www.dropbox.com/s/53rfp208rmysym3/globalsip2018.pdf.

[21] 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.

[22] K. Tsianos, S. Lawlor, and M. G. Rabbat, "Communication/computation tradeoffs in consensus-based distributed optimization," in *Advances in neural information processing systems*, 2012, pp. 1943–1951.

[23] K. I. Tsianos, S. F. Lawlor, J. Y. Yu, and M. G. Rabbat, "Networked optimization with adaptive communication," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. IEEE, 2013, pp. 579–582.

[24] D. Jakovetic, D. Bajovic, N. Krejic, and N. K. Jerinkic, "Distributed gradient methods with variable number of working nodes." *IEEE Trans. Signal Processing*, vol. 64, no. 15, pp. 4080–4095, 2016.

[25] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *arXiv preprint arXiv:1701.03961*, 2017.

[26] Z. Wang, Z. Yu, Q. Ling, D. Berberidis, and G. B. Giannakis, "Decentralized RLS with data-adaptive censoring for regressions over large-scale networks," *IEEE Transactions on Signal Processing*, vol. 66, no. 6, pp. 1634–1648, 2018.

[27] A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar, "Communication efficient distributed weighted non-linear least squares estimation," *arXiv preprint arXiv:1801.04050*, 2018.

[28] A. Mokhtari and A. Ribeiro, "Dsa: Decentralized double stochastic averaging gradient algorithm," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2165–2199, 2016.

[29] 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.

[30] 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.

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

[32] "Libsvm regression datasets," https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html.