A Stochastic Gradient Tracking Algorithm for Decentralized Optimization With Inexact Communication

Suhail M. Shah Raghu Bollapragada

***Abstract*—Decentralized optimization is typically studied under the assumption of noise-free transmission. However, real-world scenarios often involve the presence of noise due to factors such as additive white Gaussian noise channels or probabilistic quantization of transmitted data. These sources of noise have the potential to degrade the performance of decentralized optimization algorithms if not effectively addressed. In this paper, we focus on the noisy communication setting and propose an algorithm that bridges the performance gap caused by communication noise while also mitigating other challenges like data heterogeneity. We establish theoretical results of the proposed algorithm that quantify the effect of communication noise and gradient noise on the performance of the algorithm. Notably, our algorithm achieves the optimal convergence rate for minimizing strongly convex, smooth functions in the context of inexact communication and stochastic gradients. Finally, we illustrate the superior performance of the proposed algorithm compared to its state-of-the-art counterparts on machine learning problems using MNIST and CIFAR-10 datasets.**

***Index Terms*—Distributed Optimization, Network Optimization, Optimization Algorithms**

I. Introduction

The seminal works [1], [2] were one of the earliest works to formally study the problem of decentralized decision making and optimization. These works helped launch the field of decentralized optimization, where a connected network of multi agents collectively optimize an objective function by only exchanging information between neighboring agents in the network. Formally, the problem of decentralized optimization in its most succinct form can be stated as:

$$\min_{x_i \in \mathbb{R}^d} \mathbf{f}(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)$$
$$\text{s.t. } x_i = x_j,\ \forall i,j \in \{1,2,\cdots,n\} \tag{1}$$

where $\mathbf{x} := (x_1,\cdots,x_n) \in \mathbb{R}^{nd}$ with x_i being the copy of the optimization variable held by the ith node (agent) of a network and $f_i : \mathbb{R}^d \to \mathbb{R}$ is the expected value $f_i(.) = \mathbb{E}_{\xi_i}[F_i(.,\xi_i)]$ of the stochastic function $F_i(.,\xi_i) : \mathbb{R}^d \to \mathbb{R}$ private to node i. Problems of this nature arise in several applications with a prominent example being machine learning, where the f_i is a function of the data held at node i. The formulation also subsumes deterministic optimization for which $F_i(x,\xi) = f_i(x),\ \forall\,\xi$.

A key aspect of decentralized algorithms is the need for communication between nodes to achieve consensus ($x_i = x_j,\ \forall\, i,j \in [n] := \{1,2,\cdots,n\}$). However, this communication is typically not noise-free. For instance, in large-scale distributed systems used for machine learning applications, communicated vectors are often quantized to reduce overall communication costs, resulting in inexact communication [3]. Such inexactness, if not properly addressed, can degrade the algorithm's performance. Even fundamental algorithms like decentralized gradient descent (DGD) do not possess convergence guarantees or assured performance in the presence of inexact communication [4, Theorem III.8] or, Section IV, *ibid*. Therefore, it is essential to develop a framework that incorporates inexact communication to design algorithms that effectively mitigate its adverse effects.

Data heterogeneity poses another challenge in decentralized optimization. Here, data heterogeneity refers to the fact that the training data is decentralized over the nodes or generated on client devices so that each node has only access to $f_i(\cdot)$. Fundamental algorithms such as stochastic decentralized gradient descent (`S-DGD`), used to solve (1) are adversely affected by data heterogeneity [5]. To overcome these limitations, Gradient Tracking (`GT`) type methods [6], [7] have been developed which communicate an additional vector that tracks the gradient of the global objective function. However, any inexactness in the communication can again severely degrade the overall performance [4], [8]. In fact, with quantization, `GT` can empirically show divergent behavior [4, Section IV].

In this paper, we consider the question of whether the inadequacies in performance resulting from inexact communication in decentralized algorithms can be properly addressed while retaining the benefits such as achieving consensus or removing data heterogeneity dependence. Specifically, our focus is on designing and analyzing algorithms based on the `GT` strategy in the setting where the information, which could be the current iterate or the gradient tracking vector, is corrupted by additive zero-mean noise with finite variance.

A. Related Work

Several works explored the topic of inexact communication in the context of decentralized optimization, including [12]–

The authors are with the Operations Research and Industrial Engineering Program, University of Texas at Austin, TX 78705. *Email:* suhailshah2005@gmail.com, raghu.bollapragada@utexas.edu.

TABLE I: Comparison of convergence rates for strongly convex, smooth functions with stochastic gradients/communication noise for related works.

Reference	Grad. Noise	Comm. Noise	No. of iterations to ϵ-acc.
[5], [7], [9] - `Gradient Tracking (GT)`	✗	✗	$\mathcal{O}\left(\frac{L}{\mu\tau}\log\frac{1}{\epsilon}\right)$
[10] - `Stochastic DGD`	✓	✗	$\mathcal{O}\left(\frac{1}{n\mu}\frac{\sigma_g^2}{\epsilon}+\frac{\sqrt{L}}{\mu\tau}\frac{\sigma_g}{\sqrt{\epsilon}}+\frac{\sqrt{L}}{\mu\tau}\frac{\chi^2}{\sqrt{\epsilon}}+\frac{L}{\mu\tau}\log\frac{1}{\epsilon}\right)$
[5], [11] - `Stochastic GT`	✓	✗	$\mathcal{O}\left(\frac{1}{n\mu}\frac{\sigma_g^2}{\epsilon}+\frac{\sqrt{L}}{\mu\tau}\frac{\sigma_g}{\sqrt{\epsilon}}+\frac{L}{\mu\tau}\log\frac{1}{\epsilon}\right)$
[12] - `QDGD`	✗	✓	$\mathcal{O}\left(\frac{L^2}{\mu^2\tau}\frac{n\chi^2}{\epsilon^2}+\frac{L^2}{\mu^2\tau}\frac{n\sigma_c^2}{\epsilon^2}\right)$
[13] - `S-Near DGD`^{t}	✓	✓	Non convergent.
`This work, (IC-GT)`	✓	✓	$\mathcal{O}\left(\frac{1}{n\mu}\frac{\sigma_g^2}{\epsilon}+\frac{\sqrt{L}}{\mu\tau}\frac{\sigma_g}{\sqrt{\epsilon}}+\frac{1}{\mu\tau}\frac{\sigma_c^2}{\epsilon}+\frac{L}{\mu\tau}\log\frac{1}{\epsilon}\right)$

Notation: σ_g^2: Gradient noise Variance, σ_c^2: Communication noise variance, χ^2: Data heterogeneity constant satisfying $n^{-1}\sum_{i=1}^n \|\nabla f_i(x^*) - \nabla f(x^*)\|^2 \leq \chi^2$ for optimal point x^*.
L, μ, τ, n: Smoothness constant, strong convexity parameter, constant depending on network topology, total number of nodes.
For S-Near `DGD`, t denotes the number of consensus steps during each iteration and convergence is inexact even with $t \to \infty$. The convergence is to a neighbourhood of size $\mathcal{O}\left(\tau^{2t}\chi^2 + \frac{L^2}{\mu^2}\sigma_c^2\right)$.

[19]. Notably, one of the earliest and significant works in this setting is [20]. The current work extends them in several ways, including the utilization of `GT` to address data heterogeneity and the assumptions about the underlying functions. These differences allow us to achieve superior theoretical and empirical convergence properties compared to contemporary works, as documented in Table 1 and discussed in Section III.

Another related line of research to our work is that of decentralized optimization with randomized compressed communication [21]–[23]. These works focus on iterate quantization for smooth and strongly convex deterministic optimization problems using randomized compression operators. However, there are significant distinctions between our work and these prior works, including differences in the underlying assumptions. Specifically, the algorithms proposed in the aforementioned works assume access to the compression error vector, which is transmitted to the receiving node for error compensation over a noiseless channel. Furthermore, the error variance is assumed to be controllable ([21, Assumption 2]) with the convergence performance being intricately linked to it([21, Theorem 1]). In our setting, neither of these assumptions are applicable as they are violated in many practical scenarios, as discussed in Section II. On a related note, the work [24] explores a noisy iterate setting for accelerated algorithms while [25] solves the consensus problem in a continuous-time setting.

The benefits of using the `GT` strategy to address data heterogeneity have been extensively studied in numerous works [5], [7], [9].In the deterministic setting, algorithms such as `EXTRA` [6] achieve linear convergence for strongly convex, smooth functions. For the stochastic optimization setting (without communication noise), [5], [11] demonstrate that `GT` based `DGD` is agnostic to the data heterogeneity. Furthermore, variants of `GT` such as `NEXT` [26] or the `D`^{2} algorithm proposed in [27] have been shown to mitigate the effects of data heterogeneity. Other works exploring the `GT` strategy in various contexts include [7], [9], [28]–[31].

B. Contributions

The main contributions can be summarized as follows:

- We propose and analyze a novel variant of the Gradient Tracking algorithm called Inexact Communication based Gradient Tracking (`IC-GT`) to address the challenges posed by communication noise and data heterogeneity. Unlike previous approaches, our method not only retains the benefits of `GT` but also effectively eliminates the negative impact of inexact communication on algorithm performance through careful design interventions.
- We show `IC-GT` can recover (upto logarithmic factors) the optimal convergence rate requirements of $\mathcal{O}(1/\epsilon)$ iterations required to achieve ϵ-accuracy for stochastic optimization while removing the data heterogeneity dependence even in the presence of communication noise. By extending the theory for exact communication based decentralized optimization [5], [32], our results improve upon the existing works which consider communication and gradient noise under similar assumptions and achieve either a worse convergence rate or inexact convergence (cf. Table 1).
- To validate our theoretical results, we report experimental results that compare `IC-GT` with similar methods like `DGD` [1], `DIGGing` [7], and `EXTRA` [6]. Our experiments demonstrate the superior performance of `IC-GT` on logistic regression and image recognition problems on well known datasets.

The paper is organized as follows. We introduce the notation that is used through out the paper in the rest of this section. In Section II, we describe the problem formulation and in Section III, we present the proposed algorithm and its implementation. Section IV provides the convergence analysis while Section V presents the numerical evidence in its support. Future directions of research and conclusions are listed in Section 6.

Notation: We use $\mathbb{R}$ to denote the set of real numbers and $\mathbb{N}$ to denote the set of all strictly positive integers. We use $\mathbf{x}_k \in \mathbb{R}^{nd}$ to denote the stacked version of $\{x_{i,k}\}_{i\in[n]}$, where $x_{i,k} \in \mathbb{R}^d$ is a column vector which denotes the value of the objective variable held by node i at iteration k, i.e. $\mathbf{x}_k := (x_{1,k}, \cdots, x_{n,k})$. We define $\bar{\mathbf{x}}_k := \frac{1}{n}(1_n 1_n^T \otimes I_d)\mathbf{x}_k = \left(\frac{1}{n}\sum_{i=1}^n x_{i,k}, \cdots, \frac{1}{n}\sum_{i=1}^n x_{i,k}\right)$, where the column vector $1_n := (1, \cdots, 1) \in \mathbb{R}^n$ and $I_d \in \mathbb{R}^{d\times d}$ being the identity matrix. The symbol $\otimes$ is used to denote the Kronecker product between any two matrices while $\|\cdot\|$ is understood to be the ℓ_2-norm of a vector or a matrix depending upon the argument. The ℓ_2 inner product between any two vectors is denoted using $\langle \cdot, \cdot \rangle$. The following notation is used for the gradients, $\nabla \mathbf{f}(\mathbf{x}_k) := (\nabla f_1(x_{1,k}), \cdots, \nabla f_n(x_{n,k}))$ and $\nabla \mathbf{f}(\bar{\mathbf{x}}_k) := (\nabla f_1(\bar{x}_k), \cdots, \nabla f_n(\bar{x}_k))$. We also define the matrices,

$$\mathbf{I}_n = I_n \otimes I_d \quad \text{and} \quad \bar{\mathbf{I}}_n := \mathbf{I}_n - \frac{1_n 1_n^T \otimes I_d}{n}.$$

Finally, for any two real valued functions $f(\cdot)$ and $g(\cdot)$, $f(x) = \mathcal{O}(g(x))$ denotes the standard Big-O notation which implies that there exists a finite constant $C > 0$ and x_0 such that $|f(x)| \leq Cg(x)$ for all $x \geq x_0$. We use $\tilde{\mathcal{O}}(\cdot)$ when ignoring logarithmic factors.

II. Preliminaries

In this section, we provide preliminaries regarding the network and communication model, and also state the assumptions that are used in the paper.

The network is represented by a (undirected) graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{E}$ represents the set of edges. We use the matrix $Q = [q_{ij}]_{i,j\in[n]} \in \mathbb{R}^{n\times n}$ to denote the mixing matrix (or consensus matrix) that captures the connectivity of the network. By this, we mean that the entry $q_{ij} > 0$ (assumed to be equal to q_{ji}), if there is an edge between any two nodes i, $j \in \mathcal{V}$. We use $\mathcal{N}(i)$ to denote the set of neighbours of i, i.e., the set $j \in \mathcal{V}$ with $j \neq i$ for which $q_{ij} > 0$. We make the following assumption regarding the matrix Q.

Assumption 1 (Mixing matrix)**.** *The mixing matrix Q is symmetric and doubly stochastic. Furthermore, the eigenvalues $\{\lambda_i\}_{i\in[n]}$ of Q satisfy $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -1$.*

Remark 1. *The symmetric and double stochasticity assumption of Q is standard in decentralized optimization along with $\lambda_2 < 1$ which implies that the graph is connected. Therefore, it implies that $(Q \otimes I_d)\boldsymbol{x} = \boldsymbol{x}$ if and only if $x_i = x_j$ for all $i, j \in \mathcal{V}$. Moreover, it also ensures that the spectral gap $\delta(Q) := 1 - \max\{|\lambda_2|, |\lambda_n|\}$ is greater than zero which in turn ensures that the consensus error decreases linearly after each averaging step, i.e.,*

$$\left\|(Q \otimes I_d)\boldsymbol{x} - \left(\frac{1_n 1_n^T}{n} \otimes I_d\right)\boldsymbol{x}\right\|^2 \leq (1-\delta)^2 \left\|\boldsymbol{x} - \left(\frac{1_n 1_n^T}{n} \otimes I_d\right)\boldsymbol{x}\right\|^2 \quad (2)$$

for any $\boldsymbol{x} \in \mathbb{R}^{nd}$. For undirected graphs, this assumption can be guaranteed by using the Metropolis weights ([7, Section 3]).

We next describe the communication model considered in this work. We make the assumption that when any node $i \in [n]$ sends a signal vector $x_{i,k} \in \mathbb{R}^d$ to a neighboring node j at iteration $k \in \mathbb{N}$, node j receives the vector $\varphi_c(x_{i,k}) \in \mathbb{R}^d$ instead of the original vector $x_{i,k}$, where $\varphi_c(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ represents a random transformation given by

$$\varphi_c(x_{i,k}) := x_{i,k} + \epsilon_{i,k,c}\,,$$

where $\epsilon_{i,k,c} \in \mathbb{R}^d$ is a random vector. We emphasize that we do not assume access to the values of $\epsilon_{i,k,c}$. We make the following assumption concerning $\epsilon_{i,k,c}$.

Assumption 2 (Noisy signal transmission)**.** *The random noisy vector $\epsilon_{i,k,c}$ is assumed to be zero mean conditioned on x_k with bounded variance for all $i \in [n]$ and $k \in \mathbb{N}$, i.e.,*

$$\mathbb{E}\left[\epsilon_{i,k,c}|x_{i,k}\right] = 0, \quad \mathbb{E}\left[\|\epsilon_{i,k,c}\|^2\right] \leq \sigma_c^2,$$

for some finite $\sigma_c > 0$.

We note that inexact communication settings with noise satisfying Assumption 2 are widely studied in the context of decentralized optimization [13], [14], [16], [33], [34]. We emphasize that we define $\varphi_c(\cdot)$ instead of directly stating a zero-mean white noise assumption on the communication noise to highlight the fact that the noise in the communication process can arise from fundamentally different processes. To illustrate, we describe two important examples of $\varphi_c(\cdot)$ for which Assumption 2 is satisfied.

Additive White Gaussian Noise channel (AWGN): The most common approach to modeling an analog based communication channel between two nodes is through an AWGN channel [35]. In this scenario, when a node transmits a signal $y_{\text{tr}} \in \mathbb{R}$ to a neighboring node, the received signal at the receiving node, denoted as y_{rc}, can be represented as

$$y_{\text{rc}} = h y_{\text{tr}} + \epsilon_c,$$

where $h \in \mathbb{R}$ captures channel effects like fading [36], and ϵ_c represents zero-mean Gaussian noise with variance σ_c^2, independent of the transmitted signal y_{tr}. Assuming that the receiving node possesses a prior estimate of h [37, Chapter 4], it can construct an estimate of the true signal $\hat{y}_{\text{rc}}$ as,

$$\hat{y}_{\text{rc}} = \frac{1}{h} y_{\text{rc}} = y_{\text{tr}} + \frac{\epsilon_c}{h}.$$

Hence, in this scenario, we can express $\varphi_c(\cdot)$ as,

$$\varphi_c(y_{\text{tr}}) = y_{tr} + \frac{\epsilon_c}{h},$$

implying Assumption 2 is satisfied since $\mathbb{E}\left[\varphi_c(y_{\text{tr}})\right] = y_{\text{tr}}$ and $\mathbb{E}\left[\|\varphi_c(y_{\text{tr}}) - y_{\text{tr}}\|^2\right] \leq \sigma_c^2/h^2$.

Probabilistic Quantization: Another significant example of operator φ_c arises in the context of quantization with unbiased compression operators. Specifically, consider a scalar $x \in \mathbb{R}$. The quantized value $\varphi_c(x)$ can be determined based on the following rule:

$$\varphi_c(x) = \begin{cases} \lfloor x \rfloor_p \text{ with probability } (\lceil x \rceil_p - x)\Delta_p \\ \lceil x \rceil_p \text{ with probability } (x - \lfloor x \rfloor_p)\Delta_p \end{cases} \quad (3)$$

where $\lfloor x \rfloor_p$ and $\lceil x \rceil_p$ denote the operations of rounding down and up to the nearest integer multiple of $\frac{1}{\Delta_p}$ respectively, and Δ_p is a positive integer. The operator φ_c defined in (3) satisfies $\mathbb{E}\left[\varphi_c(x)\right] = x$ and $\mathbb{E}\left[|\varphi_c(x) - x|^2\right] \leq \frac{1}{4\Delta_p^2}$ as shown in [38] implying Assumption 2 is satisfied. However, it is important to note that [38] provides only a simplistic analysis and does not take into consideration the saturation errors ([39], [40]) that arise in quantization. Moreover, other quantization approaches, such as uniform quantization, natural compression, and LM (see [41] and references therein), while satisfying the unbiasedness assumption, exhibit varying variance rather than bounded variance.

We also make the following assumptions regarding the objective function.

Assumption 3 (Regularity and convexity)**.** *Each local function f_i is L-smooth and μ-strongly convex.*

Assumption 4 (Unbiased gradient samples)**.** *Each node i has access to conditionally unbiased, finite variance gradient samples $\nabla F_i(x_{i,k}, \xi_k)$ of $\nabla f_i(x_{i,k})$ for any given $x_{i,k} \in \mathbb{R}^d$, $k \in \mathbb{N}$. That is,*

$$\mathbb{E}_{\xi_{i,k}}\left[\nabla F_i(x_{i,k}, \xi_{i,k}) \,|\, x_{i,k}\right] = \nabla f_i(x_{i,k})$$
$$\mathbb{E}_{\xi_{i,k}}\left[\|\nabla F_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k})\|^2\right] \leq \sigma_g^2$$

for some finite $\sigma_g > 0$ with $\xi_{i,k}$ being assumed to be independent of $\epsilon_{i,k,c}$.

Remark 2. *The finite variance assumption in Assumption 4 can be relaxed along two possible lines with minor modifications to the convergence analysis. One relaxation would be to allow the noise to grow with the gradient norm (cf. Assumption 3b, [32]). The other possibility is to replace σ^2 with $\sigma_*^2 := \frac{1}{n}\sum_{i=1}^{n} \|\nabla F(x^*, \xi_i) - \nabla f(x^*)\|^2$, the noise at the optimal point x^*, as in [42].*

Remark 3. *The convergence analysis can also be extended to a non-convex setting by modifying the measure of stationary to be the ℓ_2-norm of the gradient.*

III. The IC-GT method

In this section, we describe the proposed method that accounts for inexact communication, referred to as Inexact Communication based Gradient Tracking (IC-GT) designed to solve the problem (1). Algorithm 1 presents the pseudo code of (IC-GT).

Algorithm 1 INEXACT COMMUNICATION based GRADIENT TRACKING (IC-GT)

```
Input Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$; Matrix $Q = [q_{ij}]_{i,j \in [n]} \in \mathbb{R}^{n \times n}$ ; Operator $\varphi_c(\cdot)$; Noise attenuation parameter $\gamma > 0$; Step size parameter $\alpha > 0$.
Initialization $x_{i,0} \in \mathbb{R}^d$, $\forall i$; $y_{i,0} := \nabla F_i(x_{i,0}, \xi_{i,0})$, $\forall i$.
while $k \geq 1$ in parallel: do
    for all $i \in [n]$, do
        $v_{i,k} = (1-\gamma)x_{i,k} + \gamma q_{ii} x_{i,k} + \gamma \sum_{j \in \mathcal{N}(i)} q_{ij} \varphi_c(x_{j,k})$
        $x_{i,k+1} = v_{i,k} - \alpha y_{i,k}$
        $y_{i,k+1} = (1-\gamma)y_{i,k} + \gamma q_{ii} y_{i,k} + \gamma \sum_{j \in \mathcal{N}(i)} q_{ij} \varphi_c(y_{j,k}) + \nabla F_i(x_{i,k+1}, \xi_{i,k+1}) - \nabla F_i(x_{i,k}, \xi_{i,k})$
    end for
    $k \to k+1$
end while
```

To express IC-GT in matrix form, we introduce the matrices $Q' := [q'_{ij}]_{i,j \in [n]}$ and $\hat{Q} := [\hat{q}_{ij}]_{i,j \in [n]}$ defined as follows:

$$\mathbf{Q}' \overset{\text{def}}{=} (I_n - Q) \otimes I_d \qquad \hat{\mathbf{Q}} \overset{\text{def}}{=} (Q - \text{diag}(Q)) \otimes I_d, \tag{4}$$

where diag(Q) denotes the diagonal matrix with entries q_{ij} for $i = j$ and 0 otherwise. Using the communication model, $\varphi_c(x_{j,k}) = x_{j,k} + \epsilon_{j,k,c}$, we can express the iteration for $v_{i,k}$ as follows:

$$\begin{aligned} v_{i,k} &= \left(1 - \gamma(1 - q_{ii})\right) x_{i,k} + \gamma \sum_{j \in \mathcal{N}(i)} q_{ij} \varphi_c(x_{j,k}) \\ &= x_{i,k} - \gamma(1 - q_{ii}))x_{i,k} + \gamma \sum_{j \in \mathcal{N}(i)} q_{ij} x_{j,k} \\ &\qquad + \gamma \sum_{j \in \mathcal{N}(i)} q_{ij} \epsilon_{j,k,c}. \end{aligned}$$

Performing a similar manipulation for the y update, we can express IC-GT using (4) as follows:

$$\mathbf{v}_k = (\mathbf{I}_n - \gamma \mathbf{Q}')\mathbf{x}_k + \gamma \hat{\mathbf{Q}} \boldsymbol{\epsilon}_{k,c} \tag{5}$$
$$\mathbf{x}_{k+1} = \mathbf{v}_k - \alpha \mathbf{y}_k \tag{6}$$
$$\begin{aligned} \mathbf{y}_{k+1} = (\mathbf{I}_n - \gamma \mathbf{Q}')\mathbf{y}_k + \nabla \mathbf{F}(\mathbf{x}_{k+1}, \boldsymbol{\xi}_{k+1}) - \nabla \mathbf{F}(\mathbf{x}_k, \boldsymbol{\xi}_k) \\ + \gamma \hat{\mathbf{Q}} \hat{\boldsymbol{\epsilon}}_{k,c} \end{aligned} \tag{7}$$

where $\hat{\epsilon}_{i,k,c} := \varphi_c(y_{i,k}) - y_{i,k}$, $\boldsymbol{\epsilon}_{k,c} := (\epsilon_{1,k,c}, \cdots, \epsilon_{n,k,c})$ and $\nabla \mathbf{F}(\mathbf{x}_k, \boldsymbol{\xi}_k) := (\nabla F_1(x_{1,k}, \xi_{1,k}), \cdots, \nabla F_n(x_{n,k}, \xi_{n,k}))$.

We next discuss the main modification made to the standard DGD algorithm [1] utilized in IC-GT to better understand its communicating and computational capabilities.

(i) Use of $\mathbf{I}_n - \gamma \mathbf{Q}'$: In the context of IC-GT, the weight matrix $\mathbf{I}_n - \gamma \mathbf{Q}'$ is employed instead of the typical $\mathbf{Q}$ used in DGD [1]. To illustrate its effectiveness in mitigating communication noise, let us examine the sequence $\{\mathbf{x}_k\}_{k \geq 0}$ generated according to the recursion:

$$\mathbf{x}_k = (\mathbf{I}_n - \gamma \mathbf{Q}')\mathbf{x}_{k-1} + \gamma \hat{\mathbf{Q}} \boldsymbol{\epsilon}_{k-1,c}, \tag{8}$$

where the noise term $\boldsymbol{\epsilon}_{k-1,c}$ satisfies Assumption 2. The recursion in (8) can be interpreted as a distributed averaging algorithm using the weight matrix $\mathbf{I}_n - \gamma \mathbf{Q}'$. Specifically, when $\gamma = 1$ and $\boldsymbol{\epsilon}_{k-1,c} = 0$, (8) reduces to the standard distributed averaging algorithm [43]. Next, we consider the expression for the averaged iterates $\bar{\mathbf{x}}_k$ obtained by multiplying (8) by $\frac{1}{n}\left(1_n 1_n^T \otimes I_d\right)$:

$$\bar{\mathbf{x}}_k = \bar{\mathbf{x}}_{k-1} - \gamma \frac{1}{n}\left(1_n 1_n^T \otimes I_d\right) \hat{\mathbf{Q}} \boldsymbol{\epsilon}_{k-1,c}, \tag{9}$$

where we used $\left(1_n^T \otimes I_d\right)\left(\mathbf{I}_n - \gamma \mathbf{Q}'\right) = (1_n^T \otimes I_d)$ from Assumption 1. Subtracting (9) from (8) and defining $\tilde{\mathbf{Q}} := \left(\mathbf{I}_n - n^{-1} 1_n 1_n^T \otimes I_d\right) \hat{\mathbf{Q}}$ and recalling $\bar{\mathbf{I}}_n := \mathbf{I}_n - \frac{1_n 1_n^T \otimes I_d}{n}$, we get,

$$\begin{aligned} \mathbf{x}_k - \bar{\mathbf{x}}_k &= (\mathbf{I}_n - \gamma \mathbf{Q}')(\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}) + \gamma \tilde{\mathbf{Q}} \boldsymbol{\epsilon}_{k-1,c} \\ &= (\mathbf{I}_n - \gamma \mathbf{Q}')(\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}) + \gamma \tilde{\mathbf{Q}} \boldsymbol{\epsilon}_{k-1,c} \\ &\qquad - \frac{1_n 1_n^T \otimes I_d}{n}(\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}) \\ &= (\bar{\mathbf{I}}_n - \gamma \mathbf{Q}')(\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}) + \gamma \tilde{\mathbf{Q}} \boldsymbol{\epsilon}_{k-1,c}, \end{aligned}$$

where the second equality is due to $\frac{1_n 1_n^T \otimes I_d}{n}(\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}) = 0$. Applying norms and taking squares yields,

$$\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \le \|\bar{\mathbf{I}}_n - \gamma \mathbf{Q}'\|^2 \|\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}\|^2 + \gamma^2 \|\tilde{\mathbf{Q}} \boldsymbol{\epsilon}_{k-1,c}\|^2 \\ + 2\gamma \left\langle (\bar{\mathbf{I}}_n - \gamma \mathbf{Q}')(\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}), \tilde{\mathbf{Q}} \boldsymbol{\epsilon}_{k-1,c} \right\rangle . \tag{10}$$

Using the conditional zero mean and finite variance assumption for $\boldsymbol{\epsilon}_{k-1,c}$ (Assumption 2), we get,

$$\mathbb{E}[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] \\ \le (1 - \gamma(1 - \lambda_2))^2 \, \mathbb{E}[\|\mathbf{x}_{k-1} - \bar{\mathbf{x}}_{k-1}\|^2] + 2n\gamma^2 \sigma_c^2,$$

where we used $\|\bar{\mathbf{I}}_n - \gamma \mathbf{Q}'\| \le 1 - \gamma(1-\lambda_2)$ since eigenvalues of $\bar{\mathbf{I}}_n - \gamma \mathbf{Q}'$ are of the form 0 and $1 - \gamma(1-\lambda_i)$ for $i = 2, \cdots, n$, and $\|\tilde{\mathbf{Q}}\|^2 \le 2$. Applying the above inequality repeatedly through iteration $k = 0$ yields,

$$\mathbb{E}[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] \le (1 - \gamma(1-\lambda_2))^{2k} \|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|^2 + \frac{2n\gamma\sigma_c^2}{1 - \lambda_2}. \tag{11}$$

(11) unveils a fundamental trade-off between two crucial aspects: the rate of decay of the consensus error and the mitigation of the influence exerted by the communication noise variance. As the parameter γ decreases, a smaller final consensus error can be achieved. However, this improvement comes at the expense of a slower convergence rate in reducing the consensus error. In view of this trade-off, the parameter γ is referred to as the '*noise attenuation*' parameter. An important work [44] which studies adaptive filtering (specifically the LMS adaptation algorithm), relates to these features of our algorithm. However, our algorithm incorporates additional considerations such as function optimization and weighted averaging over neighboring nodes rendering the analysis and results presented in the paper not applicable to our case.

(ii) Use of Gradient Tracking: Another crucial feature of IC-GT is its ability to track gradients while accommodating inexact communication through gradient tracking. The inclusion of gradient tracking offers the advantage of making the algorithm agnostic to data heterogeneity. To elaborate, the number of iterations required to achieve ϵ-accuracy using stochastic DGD depends on $\mathcal{O}\left(\frac{\sqrt{L}\chi^2}{\sqrt{\epsilon}}\right)$ [32], where χ is a constant satisfies the inequality

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*) - \nabla f(x^*)\|^2 \le \chi^2,$$

with x^* denoting the optimal solution of (1). In contrast, IC-GT eliminates the dependence on χ entirely and, moreover, recovers the linear convergence rate in scenarios where the variances of both the gradient and communication noise are zero.

IV. Convergence Analysis

In this section, we establish theoretical convergence guarantees for the proposed IC-GT algorithm. We build up to our main result through a series of technical lemmas which we state next.

Preliminaries

For the sake of brevity, we assume $\boldsymbol{\epsilon}_{k,c} = \hat{\boldsymbol{\epsilon}}_{k,c}$ in (5)-(7) for all $k \in \mathbb{N}$ without loss of generality. We begin by expressing the algorithm in terms of the difference between the variables and their corresponding averages, which we refer to as the *consensus* error. To denote this, we adopt the notation $\Delta \mathbf{z} := \mathbf{z} - \bar{\mathbf{z}}$ for any variable $\mathbf{z} \in \mathbb{R}^{nd}$, where $\bar{\mathbf{z}}$ denotes the average, i.e., $\bar{\mathbf{z}} := \left(\frac{1_n 1_n^T}{n} \otimes I_d\right) z$. We first establish a recursive relation for the consensus error.

Lemma 1. **[Recursive relation for consensus errors]** *Suppose $\boldsymbol{\epsilon}_{k,c} = \hat{\boldsymbol{\epsilon}}_{k,c}$ in (5)-(7) for all $k \in \mathbb{N}$. Then, the iterates generated by IC-GT satisfy the following recursive relation:*

$$\Psi_k = \boldsymbol{J}_\gamma \Psi_{k-1} + \alpha \boldsymbol{E}_{k-1}, \tag{12}$$

where

$$\Psi_k \overset{\text{def}}{=} \begin{bmatrix} \Delta \mathbf{v}_k \\ \Delta \mathbf{x}_k \\ \alpha \Delta \mathbf{y}_k \end{bmatrix}, \boldsymbol{J}_\gamma \overset{\text{def}}{=} \begin{bmatrix} \bar{\boldsymbol{I}}_n - \gamma \boldsymbol{Q}' & 0 & -(\bar{\boldsymbol{I}}_n - \gamma \boldsymbol{Q}') \\ 0 & \bar{\boldsymbol{I}}_n - \gamma \boldsymbol{Q}' & -\bar{\boldsymbol{I}}_n \\ 0 & 0 & \bar{\boldsymbol{I}}_n - \gamma \boldsymbol{Q}' \end{bmatrix} \tag{13}$$

and

$$\boldsymbol{E}_{k-1} \overset{\text{def}}{=} \frac{\gamma}{\alpha} \begin{bmatrix} \tilde{\boldsymbol{Q}} \boldsymbol{\epsilon}_{k,c} \\ \tilde{\boldsymbol{Q}} \boldsymbol{\epsilon}_{k-1,c} \\ \alpha \tilde{\boldsymbol{Q}} \boldsymbol{\epsilon}_{k-1,c} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \bar{\boldsymbol{I}}_n \left(\nabla \boldsymbol{F}(\mathbf{x}_k, \boldsymbol{\xi}_k) - \nabla \boldsymbol{F}(\mathbf{x}_{k-1}, \boldsymbol{\xi}_{k-1})\right) \end{bmatrix}$$

with $\bar{\boldsymbol{I}}_n := \left(I_n - \frac{1_n 1_n^T}{n}\right) \otimes I_d$, $\boldsymbol{Q}' \overset{\text{def}}{=} (I_n - Q) \otimes I_d$, $\hat{\boldsymbol{Q}} \overset{\text{def}}{=} (Q - diag(Q)) \otimes I_d$ and $\tilde{\boldsymbol{Q}} \overset{\text{def}}{=} \bar{\boldsymbol{I}}_n \hat{\boldsymbol{Q}}$.

The proof of this lemma is provided in Appendix I. One of the challenges in analyzing IC-GT is that the matrix $\mathbf{J}_\gamma$ defined in (13) is not necessarily a contractive matrix. In other words, the condition $\|\mathbf{J}_\gamma\| < 1$ is not guaranteed to hold. However, the following result demonstrates that despite this restriction, there exists a positive integer τ such that $\|\mathbf{J}_\gamma\|^\tau < 1$.

Lemma 2. **[Strict contractive property for $\mathbf{J}_\gamma$]** *Suppose Assumption 1 holds. For any given $\delta \in (0,1)$, $\gamma \in (0, 1/4)$ and λ_2 associated with the matrix Q, suppose $\tau \in \mathbb{N}$ satisfies*

$$\tau \ge \left\lceil \frac{2}{\gamma(1-\lambda_2)} \max\left\{ 4 \ln\left(\frac{2}{\gamma(1-\lambda_2)}\right), \\ \left(\gamma(1-\lambda_2) - \ln\frac{\sqrt{\delta}}{4}\right) \right\} \right\rceil, \tag{14}$$

where $\lceil \cdot \rceil$ denotes the ceiling function. Then, $\|\boldsymbol{J}_\gamma^\tau\|^2 \le \delta < 1$, where $\boldsymbol{J}_\gamma^\tau := \underbrace{\boldsymbol{J}_\gamma \times \cdots \times \boldsymbol{J}_\gamma}_{\tau \text{ times}}$.

The proof of this lemma is provided in Appendix II. The next result establishes a descent relation for the consensus error $\mathbb{E}\left[\|\Psi_{t+\tau}\|^2\right]$ in terms of $\mathbb{E}[\|\Psi_t\|^2]$.

Lemma 3. **[Descent relation for consensus error, $\mathbb{E}[\|\Psi_t\|^2]$]** *Suppose Assumptions 1-4 hold and $\boldsymbol{\epsilon}_{k,c} = \hat{\boldsymbol{\epsilon}}_{k,c}$ in (5)-(7) for all $k \in \mathbb{N}$. If γ and α satisfy (20), then, for a given $0 < \rho' \le$*

$1/4$, there exists a $\tau \in \mathbb{N}$ such that the following relations are satisfied for $t \geq \tau$:

$$\mathbb{E}[\|\Psi_t\|^2] \leq \rho' \mathbb{E}[\|\Psi_{t-\tau}\|^2] + 576\alpha^2\tau L^2 \sum_{i=t-\tau}^{t-1} \mathbb{E}[\|\Psi_i\|^2]$$
$$+ 64\gamma^2\big(2 + \alpha^2(\tau^2 + 1/2) + \alpha^2 t\big) n\sigma_c^2\tau + 196n(\tau+1)\alpha^2\sigma_g^2$$
$$+ 1344\alpha^2\tau \sum_{i=t-\tau}^{t-1} \mathbb{E}\left[\|\nabla \boldsymbol{f}(\bar{\boldsymbol{x}}_i) - \nabla \boldsymbol{f}(\boldsymbol{x}^*)\|^2\right] \tag{15}$$

and for any $\ell < \tau$:

$$\mathbb{E}[\|\Psi_\ell\|^2] \leq 2(1+\tau^2)\|\Psi_0\|^2 + 576\alpha^2\tau L^2 \sum_{i=0}^{\ell-1} \mathbb{E}[\|\Psi_i\|^2]$$
$$+ 64\gamma^2\big(2 + \alpha^2(\tau^2 + 1/2) + \alpha^2 t\big) n\sigma_c^2\tau + 196n(\tau+1)\alpha^2\sigma_g^2$$
$$+ 1344\alpha^2\tau \sum_{i=0}^{\ell} \mathbb{E}\left[\|\nabla \boldsymbol{f}(\bar{\boldsymbol{x}}_i) - \nabla \boldsymbol{f}(\boldsymbol{x}^*)\|^2\right] \tag{16}$$

The proof of this lemma is provided in Appendix III. We next prove an auxiliary result that will be useful for bounding the consensus error.

Lemma 4. *Suppose the non-negative scalar sequences $\{a_t\}_{t\geq 0}$ and $\{e_t\}_{t\geq 0}$ satisfy the following recursive relation for a fixed $\tau \in \mathbb{N}$:*

$$a_t \leq \begin{cases} \rho' a_{t-\tau} + \frac{b}{\tau}\sum_{i=t-\tau}^{t-1} a_i + c\sum_{i=t-\tau}^{t-1} e_i + r & \text{if } t \geq \tau \\ \rho'' a_0 + \frac{b}{\tau}\sum_{i=0}^{t-1} a_i + c\sum_{i=0}^{t-1} e_i + r & \text{if } t < \tau \end{cases} \tag{17}$$

where b, c, r, ρ'' are non-negative constants satisfying $b \leq \rho'/4$ and $\rho' \in (0, 1/4]$. Then, for any $t \in \mathbb{N}$,

$$a_t \leq 20\rho''\Big(1 - \frac{3\rho}{4\tau}\Big)^t a_0 + 60c\sum_{i=0}^{t-1}\Big(1 - \frac{3\rho}{4\tau}\Big)^{t-i} e_i + \frac{26r}{\rho}, \tag{18}$$

where $\rho := 1 - 2\rho'$.

The detailed proof of this lemma is provided in the online version of this paper [45]. We are ready to state and prove the main convergence result.

Main Result

For convenience, we define Δx_k^* as

$$\Delta x_k^* \overset{\text{def}}{=} \mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right], \qquad \forall k \in \mathbb{N}. \tag{19}$$

where x^* is the optimal solution of (1).

Theorem 5. **[Convergence rate of IC-GT]** *Suppose Assumptions 1-4 hold and $\boldsymbol{\epsilon}_{k,c} = \hat{\boldsymbol{\epsilon}}_{k,c}$ in (5)-(7) for all $k \in \mathbb{N}$. If*

$$\alpha \leq \min\left\{1, \frac{1}{161280\tau L}\right\} \quad \textit{and} \quad 0 < \gamma < 1/4, \tag{20}$$

where

$$\tau = \left\lceil \tfrac{2}{\gamma(1-\lambda_2)} \max\left\{4\ln\left(\tfrac{2}{\gamma(1-\lambda_2)}\right), \gamma(1-\lambda_2) + \ln 16\right\}\right\rceil. \tag{21}$$

Then, for any $T \in \mathbb{N}$, we have,

$$\Delta x_T^* \leq (1 - \alpha\mu/4)^T \left(\Delta x_0^* + \frac{800(1+\tau^2)L}{n(1-\alpha\mu/4)\mu}\|\Psi_0\|^2\right)$$
$$+ \left(\frac{4(1+2\mu^{-1}T\alpha)}{\mu}\frac{\gamma^2}{\alpha}\right.$$
$$\left. + \frac{33280(2 + \alpha^2(\tau^2 + \frac{1}{2}) + \alpha^2 T)L}{\mu} n\tau\gamma^2\right)\frac{\sigma_c^2}{n}$$
$$+ \left(\frac{4\alpha}{\mu} + \frac{101920\, nL(\tau+1)\alpha^2}{\mu}\right)\frac{\sigma_g^2}{n}. \tag{22}$$

We make the following remarks regrading Theorem 5.

Remark 4. ***(Dependence of τ on network)*** *The parameter τ depends on the network connectivity (λ_2) and the noise attenuation parameter γ (cf. (21)) which highlights the role played by γ in shaping the consensus properties of `IC-GT` (cf. Lemma 3). From (21), we note that a smaller value of γ increases τ but reduces the impact of the communication noise variance σ_c^2 in (22) which is reminiscent of the trade-off discussed in Section III.*

Remark 5. ***(Iteration complexity of `IC-GT`)*** *(20) and (21) suggest that the choices of the step size α and the noise attenuation parameter γ are inherently connected. Using (21) in (20), we have the following relation:*

$$\frac{\alpha}{\gamma} = \tilde{\mathcal{O}}\left(\frac{1-\lambda_2}{L}\right) \tag{23}$$

To establish the iteration complexity, i.e. calculate the number of iterations T required to reach ϵ-accuracy (i.e., T such that $\Delta x_T^* \leq \epsilon$), we consider the contributions of the noise terms in the complexity bound of Theorem 5 individually. To keep the representation clear, we use the big-$\mathcal{O}$ notation which only considers the dependence on the free parameters (T, α, γ) and hides the dependence on constant factors such as μ, L, n, Δx_0^* and $\|\Psi_0\|^2$. Accordingly, we note that the contribution of the gradient noise terms in (22) is given by

$$\mathcal{O}\big((\alpha + \alpha^2 n)\sigma_g^2/n\big) = \mathcal{O}\big(\alpha\sigma_g^2/n\big) \qquad \textit{if } \alpha \leq 1/n \tag{24}$$

while the contribution of the communication noise terms in (22) is given by:

$$\mathcal{O}\left(\left(\frac{(1+T\alpha)\gamma^2}{\alpha} + (\alpha^2\tau^2 + \alpha^2 T)n\tau\gamma^2\right)\frac{\sigma_c^2}{n}\right)$$
$$= \tilde{\mathcal{O}}\left(\left(\frac{(1+T\alpha)\gamma^2}{\alpha} + \frac{n\alpha^2}{\gamma} + n\alpha^2\gamma T\right)\frac{\sigma_c^2}{n}\right), \tag{25}$$

where we used $\tau = \tilde{\mathcal{O}}(1/\gamma)$ and ignored the dependency on other problem parameters. If we set $\gamma = \tilde{\mathcal{O}}(\alpha)$ such that (23) is satisfied, (25) simplifies to

$$\tilde{\mathcal{O}}\left(\left((1+T\alpha)\alpha + n\alpha + n\alpha^3 T\right)\frac{\sigma_c^2}{n}\right).$$

For any given $\epsilon > 0$, we can set $\alpha = \epsilon$ implying that $T = \tilde{\mathcal{O}}(\epsilon^{-1})$ iterations are required to achieve the specified ϵ-accuracy.

Remark 6. *($\sigma_{\mathbf{c}}^{\mathbf{2}} = \mathbf{0}$, $\sigma_{\mathbf{g}}^{\mathbf{2}} = \mathbf{0}$ **and** $\sigma_{\mathbf{c}}^{\mathbf{2}} = \mathbf{0}$, $\sigma_{\mathbf{g}}^{\mathbf{2}} > \mathbf{0}$): In the absence of communication or gradient approximation errors ($\sigma_c^2 = 0$, $\sigma_g^2 = 0$), we can achieve the deterministic linear convergence rate of the gradient tracking algorithm [7]. Referring to equation (22), we obtain the following inequality:*

$$\Delta x_T^* \le (1-\alpha\mu/4)^T \left(\Delta x_0^* + \frac{800(1+\tau^2)L}{n(1-\alpha\mu/4)\mu}\|\Psi_0\|^2\right)$$

The case $\sigma_c^2 = 0$, $\sigma_g^2 > 0$ considers stochastic decentralized optimization with no communication noise. For this scenario, with a constant $\alpha > 0$, we have linear convergence to a neighbourhood of size $\mathcal{O}\big((\alpha^2 n + \alpha)\sigma_g^2/n\big)$ [11]. A point to be remarked here is that `IC-GT` *not only removes the data heterogeneity terms which arise in the convergence bound for* `DGD` *(cf. Table 1) but also makes sure that the variance scales linearly with the number of nodes provided $\alpha \le 1/n$ (cf.* (24)*).*

Remark 7. ***(Consensus Error):*** *We can establish convergence error bounds for the expected consensus error $\mathbb{E}[\|\Psi_k\|^2]$ by combining the results of Lemma 3 and Theorem 5. However, for brevity, we omit the explicit presentation of these results as they are of the same order as the results for Δx_T^*.*

Proof of Theorem 5: Using (5) and recalling that $\bar{\mathbf{x}} := \frac{(1_n1_n^T)\otimes I_d}{n}\mathbf{x}$, the recursion for $\bar{\mathbf{x}}_k$ can be expressed as

$$\begin{aligned}\bar{\mathbf{x}}_{k+1} &= \bar{\mathbf{v}}_k - \alpha\bar{\mathbf{y}}_k\\ &= \bar{\mathbf{x}}_k + \gamma\bar{\boldsymbol{\epsilon}}_{k,c} - \alpha\bar{\mathbf{y}}_k,\end{aligned} \tag{26}$$

where $\bar{\boldsymbol{\epsilon}}_{k,c} := \frac{1}{n}\big(1_n1_n^T \otimes I_d\big)\hat{\mathbf{Q}}\boldsymbol{\epsilon}_{k,c}$ and the last equality is due to $\bar{\mathbf{v}}_k = \bar{\mathbf{x}}_k + \gamma\bar{\boldsymbol{\epsilon}}_{k,c}$. Similarly, the recursion for $\bar{\mathbf{y}}_k := \frac{1}{n}\big(1_n1_n^T \otimes I_d\big)y_k$ can be given as,

$$\begin{aligned}\bar{\mathbf{y}}_k = \bar{\mathbf{y}}_{k-1} + \frac{1}{n}\big(1_n1_n^T \otimes I_d\big)\big(\nabla\mathbf{F}(\mathbf{x}_k,\boldsymbol{\xi}_k) - \nabla\mathbf{F}(\mathbf{x}_{k-1},\boldsymbol{\xi}_{k-1})\big)\\ + \gamma\bar{\boldsymbol{\epsilon}}_{k-1,c}.\end{aligned}$$

Taking telescopic sum from 0 to k leads to the following recursion:

$$\bar{\mathbf{y}}_k = \frac{1}{n}\big(1_n1_n^T \otimes I_d\big)\nabla\mathbf{F}(\mathbf{x}_k,\boldsymbol{\xi}_k) + \gamma\sum_{j=1}^{k}\bar{\boldsymbol{\epsilon}}_{j-1,c} \tag{27}$$

since $\bar{\mathbf{y}}_0 = \frac{1}{n}\big(1_n1_n^T \otimes I_d\big)\nabla\mathbf{F}(\mathbf{x}_0,\boldsymbol{\xi}_0)$. Plugging (27) in (26), we get,

$$\begin{aligned}\bar{\mathbf{x}}_{k+1} &= \bar{\mathbf{x}}_k + \gamma\bar{\boldsymbol{\epsilon}}_{k,c} - \frac{\alpha}{n}\big(1_n1_n^T \otimes I_d\big)\nabla\mathbf{F}(\mathbf{x}_k,\boldsymbol{\xi}_k) - \gamma\alpha\sum_{j=0}^{k-1}\bar{\boldsymbol{\epsilon}}_{j,c}\\ &= \bar{\mathbf{x}}_k - \frac{\alpha}{n}\big(1_n1_n^T \otimes I_d\big)\nabla\mathbf{f}(\mathbf{x}_k) + \gamma\left(\bar{\boldsymbol{\epsilon}}_{k,c} - \alpha\sum_{j=0}^{k-1}\bar{\boldsymbol{\epsilon}}_{j,c}\right)\\ &\quad + \frac{\alpha}{n}\big(1_n1_n^T \otimes I_d\big)\Big(\nabla\mathbf{f}(\mathbf{x}_k) - \nabla\ \mathbf{F}(\mathbf{x}_k,\boldsymbol{\xi}_k)\Big)\\ &= \bar{\mathbf{x}}_k - \frac{\alpha}{n}\big(1_n1_n^T \otimes I_d\big)\nabla\mathbf{f}(\mathbf{x}_k) + \underbrace{\alpha\boldsymbol{\epsilon}_{k,g} + \gamma\bar{\boldsymbol{\epsilon}}_{k,c}}_{\delta_k}\\ &\quad - \alpha\gamma\sum_{j=0}^{k-1}\bar{\boldsymbol{\epsilon}}_{j,c}\end{aligned} \tag{28}$$

where $\boldsymbol{\epsilon}_{k,g}$ is defined to be $\boldsymbol{\epsilon}_{k,g} := \frac{1}{n}\big(1_n1_n^T \otimes I_d\big)\Big(\nabla\mathbf{f}(\mathbf{x}_k) - \nabla\mathbf{F}(\mathbf{x}_k,\boldsymbol{\xi}_k)\Big)$ with $\mathbb{E}\left[\boldsymbol{\epsilon}_{k,g}|\mathbf{x}_k\right] = 0$ and $\mathbb{E}[\|\boldsymbol{\epsilon}_{k,g}\|^2] \le \sigma_g^2$ from Assumption 4. Now, let $\mathcal{F}_k \overset{\text{def}}{=} \sigma(\mathbf{x}_0, \boldsymbol{\xi}_0, \boldsymbol{\epsilon}_{0,c}, \cdots, \boldsymbol{\xi}_{k-1}, \boldsymbol{\epsilon}_{k-1,c})$ be the sigma algebra generated by the random variables up to iteration k. Then, for any constant $\beta > 0$, we have,

$$\begin{aligned}&\mathbb{E}[\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2|\mathcal{F}_k]\\ &\le (1+\beta)\mathbb{E}[\|\bar{\mathbf{x}}_k - \frac{\alpha}{n}\big(1_n1_n^T \otimes I_d\big)\nabla\mathbf{f}(\mathbf{x}_k) - \mathbf{x}^* + \delta_k\|^2|\mathcal{F}_k]\\ &\qquad + (1+\beta^{-1})\alpha^2\gamma^2\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\bar{\boldsymbol{\epsilon}}_{j,c}\right\|^2\middle|\mathcal{F}_k\right]\\ &= (1+\beta)\|\bar{\mathbf{x}}_k - \frac{\alpha}{n}\big(1_n1_n^T \otimes I_d\big)\nabla\mathbf{f}(\mathbf{x}_k) - \mathbf{x}^*\|^2\\ &\quad + (1+\beta)\mathbb{E}[\|\delta_k\|^2|\mathcal{F}_k] + (1+\beta^{-1})\alpha^2\gamma^2\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\bar{\boldsymbol{\epsilon}}_{j,c}\right\|^2\middle|\mathcal{F}_k\right],\end{aligned} \tag{29}$$

where the equality is due to $\mathbb{E}[\delta_k|\mathcal{F}_k] = 0$ from Assumption 3. From Assumptions 2 and 4, we have,

$$\mathbb{E}[\|\delta_k\|^2] = \mathbb{E}[\|\alpha\boldsymbol{\epsilon}_{k,g} + \gamma\bar{\boldsymbol{\epsilon}}_{k,c}\|^2] \le \alpha^2\sigma_g^2 + \gamma^2\sigma_c^2, \tag{30}$$

where we have used $\mathbb{E}[\langle\boldsymbol{\epsilon}_{k,g}, \bar{\boldsymbol{\epsilon}}_{k,c}\rangle] = 0$. Furthermore, we have,

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\bar{\boldsymbol{\epsilon}}_{j,c}\right\|^2\right] = \mathbb{E}\left[\sum_{j=0}^{k-1}\|\bar{\boldsymbol{\epsilon}}_{j,c}\|^2\right]\\ &\qquad + \sum_{1\le p,p'\le k-1}\mathbb{E}\left[\langle\bar{\boldsymbol{\epsilon}}_{p,c}, \bar{\boldsymbol{\epsilon}}_{p',c}\rangle\right] \le \sum_{j=0}^{k-1}\sigma_c^2 = k\sigma_c^2,\end{aligned} \tag{31}$$

where we use $\mathbb{E}[\langle\bar{\boldsymbol{\epsilon}}_{p,c}, \bar{\boldsymbol{\epsilon}}_{p',c}\rangle] = \mathbb{E}[\mathbb{E}\left[\langle\bar{\boldsymbol{\epsilon}}_{p,c}, \bar{\boldsymbol{\epsilon}}_{p',c}\rangle|\mathcal{F}_{p'}\right]] = 0$ for $p < p'$. Taking full expectations in (29), it then follows that,

$$\begin{aligned}&\mathbb{E}[\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2]\\ &\quad\le (1+\beta)\mathbb{E}\left[\left\|\bar{\mathbf{x}}_k - \frac{\alpha}{n}\big(1_n1_n^T \otimes I_d\big)\nabla\mathbf{f}(\mathbf{x}_k) - \mathbf{x}^*\right\|^2\right]\\ &\qquad + \big((1+\beta)(\gamma^2\sigma_c^2 + \alpha^2\sigma_g^2) + k(1+\beta^{-1})\alpha^2\gamma^2\sigma_c^2\big).\end{aligned}$$

where we used (30) to get the inequality. We note that since $\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}^*\|^2 = \left\|\frac{(11^T)\otimes I_d}{n}(\mathbf{x}_{k+1} - \mathbf{x}^*)\right\|^2 = n\|\bar{x}_{k+1} - x^*\|^2$, the above inequality leads to,

$$\begin{aligned}&\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^2\right]\\ &\quad\le (1+\beta)\mathbb{E}\left[\left\|\bar{x}_k - \frac{\alpha}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k}) - x^*\right\|^2\right]\\ &\qquad + n^{-1}\big((1+\beta)(\gamma^2\sigma_c^2 + \alpha^2\sigma_g^2) + k(1+\beta^{-1})\alpha^2\gamma^2\sigma_c^2\big).\end{aligned} \tag{32}$$

Considering the first term on the right hand side of (32), we have,

$$\Big\|\bar{x}_k - \frac{\alpha}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k}) - x^*\Big\|^2 = \|\bar{x}_k - x^*\|^2$$
$$-\frac{2\alpha}{n}\Big\langle\sum_{i=1}^{n}\nabla f_i(x_{i,k}), \bar{x}_k - x^*\Big\rangle + \alpha^2\Big\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k})\Big\|^2. \tag{33}$$

The second term on the right hand side of (33) is bounded as

$$\begin{aligned}&\langle\sum_{i=1}^{n}\nabla f_i(x_{i,k}), \bar{x}_k - x^*\rangle\\ &= \langle\sum_{i=1}^{n}\nabla f_i(x_{i,k}), \bar{x}_k - x_{i,k}\rangle + \langle\sum_{i=1}^{n}\nabla f_i(x_{i,k}), x_{i,k} - x^*\rangle\\ &\geq \sum_{i=1}^{n}\Big[f_i(\bar{x}_k) - f_i(x_{i,k}) - \frac{L}{2}\|\bar{x}_k - x_{i,k}\|^2\\ &\qquad + f_i(x_{i,k}) - f_i(x^*) + \frac{\mu}{2}\|x_{i,k} - x^*\|^2\Big]\\ &\geq \sum_{i=1}^{n}\Big[f_i(\bar{x}_k) - f_i(x^*) - \frac{L+\mu}{2}\|\bar{x}_k - x_{i,k}\|^2\\ &\qquad + \frac{\mu}{4}\|\bar{x}_k - x^*\|^2\Big],\end{aligned} \tag{34}$$

where the second inequality is due to Assumption 3 and the last inequality is due to the inequality $\|\bar{x}_k - x^*\|^2 \leq 2\|\bar{x}_k - x_{i,k}\|^2 + 2\|x_{i,k} - x^*\|^2$. The last term on the right hand side of (33) can be bounded as

$$\begin{aligned}\Big\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k})\Big\|^2 &= \Big\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k}) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}_k)\\ &\qquad + \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}_k) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x^*)\Big\|^2\\ &\leq \frac{2L^2}{n}\sum_{i=1}^{n}\|\bar{x}_k - x_{i,k}\|^2 + \frac{4L}{n}\sum_{i=1}^{n}(f_i(\bar{x}_k) - f_i(x^*)),\end{aligned} \tag{35}$$

where in the second summation, we used the fact that $\|\nabla f_i(\bar{x}_k) - \nabla f_i(x^*)\|^2 \leq 2L(f_i(\bar{x}_k) - f_i(x^*))$ by Assumption 3 [46, Theorem 2.1.5]. Using (34) and (35) in (33) along with $\alpha < 1/4L$, we have,

$$\begin{aligned}&\|\bar{x}_{k+1} - x^*\|^2\\ &\leq (1 - \alpha\mu/2)\|\bar{x}_k - x^*\|^2 - \frac{\alpha}{n}\big(\sum_{i=1}^{n}f_i(\bar{x}_k) - f_i(x^*)\big)\\ &\qquad + \frac{(3L/2 + \mu)\alpha}{n}\sum_{i=1}^{n}\|\bar{x}_k - x_{i,k}\|^2\\ &\leq (1 - \alpha\mu/2)\|\bar{x}_k - x^*\|^2 - \alpha\left(\mathbf{f}(\bar{\mathbf{x}}_k) - \mathbf{f}(\mathbf{x}^*)\right) + \frac{5\alpha L}{2n}\|\Psi_k\|^2,\end{aligned} \tag{36}$$

where the last inequality is due to $\|\Delta\mathbf{x}_k\|^2 \leq \|\Psi_k\|^2$. Using (36) in (32), we get,

$$\begin{aligned}\mathbb{E}[\|\bar{x}_{k+1} - x^*\|^2] &\leq (1+\beta)\Big\{(1 - \alpha\mu/2)\mathbb{E}[\|\bar{x}_k - x^*\|^2]\\ &- \alpha\left(\mathbb{E}[\mathbf{f}(\bar{\mathbf{x}}_k)] - \mathbf{f}(\mathbf{x}^*)\right) + \frac{5\alpha L}{2n}\mathbb{E}[\|\Psi_k\|^2]\Big\}\\ &+ \frac{1}{n}\big((1+\beta)(\gamma^2\sigma_c^2 + \alpha^2\sigma_g^2) + k(1+\beta^{-1})\alpha^2\gamma^2\sigma_c^2\big).\end{aligned}$$

Set $\beta \overset{\text{def}}{=} \frac{\alpha\mu}{4(1-\frac{\alpha\mu}{2})}$. We note that $(1+\beta^{-1}) \leq 4/\alpha\mu$ and $(1+\beta) = \frac{(1-\alpha\mu/4)}{(1-\alpha\mu/2)}$ with $1 \leq (1+\beta) \leq 2$. Then, we have,

$$\begin{aligned}\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^2\right] &\leq \Big\{(1 - \alpha\mu/4)\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right]\\ &- \alpha\left(\mathbb{E}[\mathbf{f}(\bar{\mathbf{x}}_k)] - \mathbf{f}(\mathbf{x}^*)\right) + \frac{5\alpha L}{n}\mathbb{E}[\|\Psi_k\|^2]\Big\}\\ &+ \frac{1}{n}\big((2 + 4\mu^{-1}k\alpha)\gamma^2\sigma_c^2 + 2\alpha^2\sigma_g^2\big).\end{aligned} \tag{37}$$

Multiplying both sides of (37) by $w_{k+1} \overset{\text{def}}{=} (1 - \alpha\mu/4)^{-(k+1)}$, we have,

$$\begin{aligned}w_{k+1}\Delta x^*_{k+1} &\leq (1 - \alpha\mu/4)w_{k+1}\Delta x^*_k\\ &+ \frac{5\alpha L}{n}w_{k+1}\mathbb{E}[\|\Psi_k\|^2] - \alpha w_{k+1}(\mathbb{E}[\mathbf{f}(\bar{\mathbf{x}}_k)] - \mathbf{f}(\mathbf{x}^*))\\ &\qquad + \frac{w_{k+1}}{n}\big(\gamma^2(2 + 4k\mu^{-1}\alpha)\sigma_c^2 + 2\alpha^2\sigma_g^2\big).\end{aligned}$$

Rearranging the terms, we get,

$$\begin{aligned}0 &\leq w_k\Delta x^*_k - w_{k+1}\Delta x^*_{k+1} + \frac{5\alpha L}{n}w_{k+1}\mathbb{E}[\|\Psi_k\|^2] - \alpha w_{k+1}\\ &\times\left(\mathbb{E}\left[\mathbf{f}(\bar{\mathbf{x}}_k)\right] - \mathbf{f}(\mathbf{x}^*)\right) + \frac{w_{k+1}}{n}\big(\gamma^2(2 + 4k\mu^{-1}\alpha)\sigma_c^2 + 2\alpha^2\sigma_g^2\big).\end{aligned}$$

Summing the above inequality from $k = 0$ to $T - 1$, we get,

$$\begin{aligned}w_T\Delta x^*_T &\leq w_0\Delta x^*_0 + \frac{1}{n}\big(\gamma^2(2 + 4T\mu^{-1}\alpha)\sigma_c^2 + 2\alpha^2\sigma_g^2\big)\sum_{k=0}^{T-1}w_{k+1}\\ &+ \frac{5\alpha L}{n}\sum_{k=0}^{T-1}w_{k+1}\mathbb{E}[\|\Psi_k\|^2] - \alpha\sum_{k=0}^{T-1}w_{k+1}(\mathbb{E}[f(\bar{\mathbf{x}}_k)] - f(\mathbf{x}^*)).\end{aligned} \tag{38}$$

We note that we can write the relations (15)-(16) in Lemma 3 in the form of (17) with

$$\begin{aligned}&b := 576\alpha^2L^2\tau^2 \qquad c := 1344\alpha^2\tau\\ &r := 64\gamma^2\big(2 + \alpha^2(\tau^2 + 1/2) + \alpha^2T\big)n\sigma_c^2\tau + 196n(\tau+1)\alpha^2\sigma_g^2\\ &e_k := \mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_k) - \nabla\mathbf{f}(\mathbf{x}^*)\|^2\right]\end{aligned} \tag{39}$$

and we have taken $\rho' = 1/4$ which fixes τ in (21) according to the bound (14) (cf. (83)). Note that since $\alpha < \frac{\sqrt{\rho'}}{2\sqrt{576}L\tau}$, $b \leq \frac{\rho'}{4} = \frac{1}{16}$. Then, with $a_t \overset{\text{def}}{=} \mathbb{E}[\|\Psi_t\|^2]$ in Lemma 4, we get,

$$\begin{aligned}\mathbb{E}[\|\Psi_t\|^2] &\leq 40(1+\tau^2)\left(1 - \frac{3\rho}{4\tau}\right)^t\|\Psi_0\|^2\\ &+ 60c\sum_{j=0}^{t-1}\left(1 - \frac{3\rho}{4\tau}\right)^{t-j}\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_j) - \nabla\mathbf{f}(\mathbf{x}^*)\|^2\right] + 52r\end{aligned} \tag{40}$$

with $\rho'' = 2(1+\tau^2)\|\Psi_0\|^2$ and $\rho = 1 - 2\rho' = 1/2$. We next bound the summation $\sum_{k=0}^{T-1} w_{k+1}\mathbb{E}\|\Psi_k\|^2$ in (38). To do this, we multiply both sides of (40) by w_{k+1} and sum from $t=0$ to $T-1$:

$$\begin{aligned}&\sum_{k=0}^{T-1}(1-\alpha\mu/4)^{-(k+1)}\mathbb{E}\|\Psi_k\|^2\\ &\leq 40(1+\tau^2)\|\Psi_0\|^2\sum_{k=0}^{T-1}(1-\alpha\mu/4)^{-(k+1)}\left(1-\frac{3\rho}{4\tau}\right)^k\\ &+60c\sum_{k=0}^{T-1}(1-\alpha\mu/4)^{-(k+1)}\sum_{j=0}^{k-1}\left(1-\frac{3\rho}{4\tau}\right)^{t-j}e_j + 52rW_{T-1},\end{aligned} \tag{41}$$

where $W_{T-1} = \sum_{k=0}^{T-1} w_{k+1}$. From (20), we have,

$$\begin{aligned}\alpha\mu/2 \leq 3\rho/4\tau &\implies \alpha\mu/2(1-\alpha\mu/8) \leq 3\rho/4\tau\\ &\implies 1 - 3\rho/4\tau \leq (1-\alpha\mu/4)^2.\end{aligned} \tag{42}$$

We use (42) to bound the two summations on the right hand side of (41) as follows:

$$\sum_{k=0}^{T-1}(1-\alpha\mu/4)^{-(k+1)}\left(1-\frac{3\rho}{4\tau}\right)^k \leq \sum_{k=0}^{T-1}(1-\alpha\mu/4)^{k-1} \leq \frac{4w_1}{\alpha\mu}, \tag{43}$$

and

$$\begin{aligned}&\sum_{k=0}^{T-1}(1-\alpha\mu/4)^{-(k+1)}\sum_{j=0}^{k-1}\left(1-\frac{3\rho}{4\tau}\right)^{k-j}e_j\\ &=\sum_{k=0}^{T-1}\sum_{j=0}^{k-1}(1-\alpha\mu/4)^{-(k+1)+j+1}\left(1-\frac{3\rho}{4\tau}\right)^{k-j}w_{j+1}e_j\\ &=\sum_{k=0}^{T-1}\sum_{j=0}^{k-1}\left(\frac{1-3\rho/4\tau}{1-\alpha\mu/4}\right)^{k-j}w_{j+1}e_j\\ &\leq\sum_{k=0}^{T-1}\sum_{j=0}^{k-1}(1-\alpha\mu/4)^{k-j}w_{j+1}e_j\\ &\leq\sum_{k=0}^{T-1}(1-\alpha\mu/4)^k\sum_{k=0}^{T-1}w_{k+1}e_k \leq \frac{4}{\mu\alpha}\sum_{k=0}^{T-1}w_{k+1}e_k,\end{aligned} \tag{44}$$

where the first equality is due to (42) and the second inequality is obtained using the relation $\sum_{k=0}^{T-1}\sum_{j=0}^{k-1}a_{k-j}b_j \leq \sum_{k=0}^{T-1}a_k\sum_{k=0}^{T-1}b_k$ for any two non-negative scalar sequences a_k, b_k, $k \geq 0$. Plugging the previous two bounds in (41), we get,

$$\begin{aligned}\sum_{k=0}^{T-1}w_{k+1}\mathbb{E}[\|\Psi_k\|^2] &\leq \frac{160w_1(1+\tau^2)\|\Psi_0\|^2}{\alpha\mu}\\ &+\frac{240ncL}{\mu\alpha}\sum_{k=0}^{T-1}w_{k+1}\left(\mathbb{E}[\mathbf{f}(\bar{\mathbf{x}}_k)] - \mathbf{f}(\mathbf{x}^*)\right) + 52rW_{T-1},\end{aligned}$$

where we have additionally used the fact that $\|e_k\|^2 = \mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_k)) - \nabla\mathbf{f}(\mathbf{x}^*)\|^2\right] \leq 2nL(\mathbb{E}[\mathbf{f}(\bar{\mathbf{x}}_k))] - \mathbf{f}(\mathbf{x}^*))$ from Assumption 3 [46, Theorem 2.1.5]. Finally, using the above bound in (38), we get,

$$\begin{aligned}w_T\Delta x_T^* &\leq w_0\Delta x_0^* + \frac{1}{n}(\gamma^2(2+4T\mu^{-1}\alpha)\sigma_c^2 + 2\alpha^2\sigma_g^2)W_{T-1}\\ &+\frac{5\alpha L}{n}\left(\frac{160w_1(1+\tau^2)\|\Psi_0\|^2}{\mu\alpha} + \frac{240ncL}{\mu\alpha}\sum_{k=0}^{T-1}w_{k+1}(\mathbb{E}[\mathbf{f}(\bar{\mathbf{x}}_k)]\right.\\ &\left.-\mathbf{f}(\mathbf{x}^*)) + 52rW_{T-1}\right) - \alpha\sum_{k=0}^{T-1}w_{k+1}(\mathbb{E}[\mathbf{f}(\bar{\mathbf{x}}_k)] - \mathbf{f}(\mathbf{x}^*)).\end{aligned}$$

Rearranging the terms in the above inequality and recalling that $c = 1344\alpha^2\tau$, we get,

$$\begin{aligned}\Delta x_T^* &\leq \frac{1}{w_T}\left\{\Delta x_0^* + \frac{800w_1(1+\tau^2)L}{n\mu}\|\Psi_0\|^2\right\}\\ &+\frac{2}{\mu\alpha}\left\{\frac{\gamma^2(2+4T\mu^{-1}\alpha)\sigma_c^2}{n} + \frac{2\alpha^2\sigma_g^2}{n}\right\} + \frac{520Lr}{n\mu}\\ &+\alpha\underbrace{\left\{\frac{1612800\tau L^2}{\mu}\alpha - 1\right\}}_{\leq 0}\sum_{k=0}^{T-1}\frac{w_{k+1}}{w_T}\underbrace{(\mathbb{E}[\mathbf{f}(\bar{\mathbf{x}}_k)] - f(\mathbf{x}^*))}_{\geq 0},\end{aligned}$$

where we used $W_{T-1}/w_T \leq 2/\mu\alpha$. The last term on the right had side is less than zero due to the condition on α (see (20)). Plugging the value of r from (39) in the above inequality completes the proof. □

V. Numerical Experiments

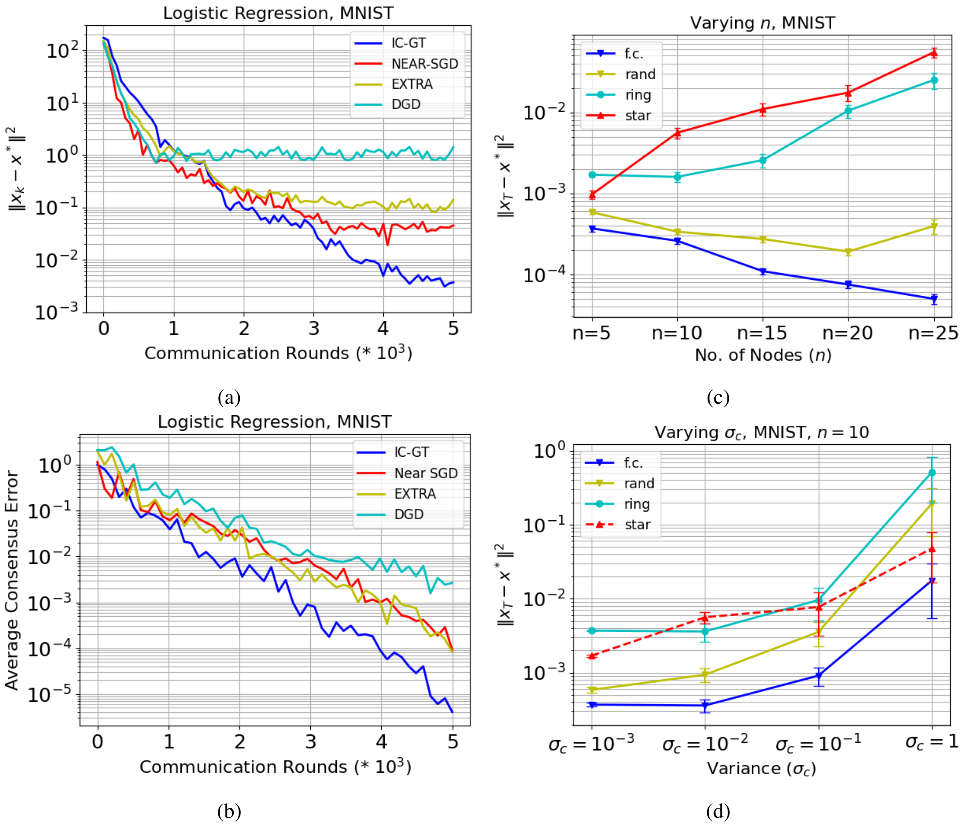
In this section, we present an empirical evaluation of the performance of IC-GT through two sets of numerical experiments. The first set focuses on logistic regression on the MNIST dataset, while the second set explores the effect of different noise variances in a deep learning setting. All experiments were implemented using PyTorch, with a dedicated CPU core functioning as a node.

Logistic regression

We first consider ℓ_2 regularized logistic regression problems of the form,

$$\begin{aligned}\min_{x\in\mathbb{R}^d}\Bigg\{L(x;y,z) := -\frac{1}{m}\sum_{i=1}^m\Big\{z^i\log\varphi(x^Ty^i)\\ +(1-z^i)\log(1-\varphi(x^Ty^i))\Big\} + \frac{\lambda}{2}\|x\|^2\Bigg\},\end{aligned} \tag{45}$$

where $x \in \mathbb{R}^d$ denote the learnable model parameters, $\{y^i, z^i\}_{i=1}^m$ denote the set of m data points, $\varphi(\cdot)$ denotes the sigmoid function, and $\lambda > 0$ is the regularization parameter. We use the MNIST dataset which consists of 60,000, 28×28 pixel grayscale images of handwritten single digits between 0 and 9. The data is partitioned in a disjoint manner amongst the nodes by assigning each node 10^3 data samples independently. To simulate the inexact communication setting, we incorporate zero-mean Gaussian noise with a variance of σ_c^2 into the transmitted model estimates independently. We adopt a star topology with $n = 10$ for the communication structure. In evaluating the performance, we employ the ℓ_2 distance



(a)

(c)

(b)

(d)

Fig. 1: **(a)-(b)** Optimality Error, Average consensus error vs. communication rounds for MNIST dataset with star topology ($n = 10$). **(c)** Final Optimal error $\|x_T - x^*\|^2$, $T = 5 \times 10^3$ for $n \in \{5, 10, 15, 20, 25\}$ for different topologies. **(d)** Final Optimal error $\|x_T - x^*\|^2$ for $\sigma_c \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$

between the averaged variable $\bar{x}_k$ and the optimal point x^*. The optimal point x^* is computed using the L-BFGS algorithm from the SciPy library in Python. We also include the average consensus error as a performance metric, which is computed as $\frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} \|x_i - x_j\|^2$, where $\mathcal{E}$ represents the edge set. We compare our proposed algorithm (IC-GT) with several baselines, including the NEAR-SGD algorithm from [13], the EXTRA algorithm proposed in [6], and the gradient tracking method [5]. Additionally, we include the performance of the DGD algorithm for comparison purposes.

In our experiments, we set the batch size to 32 and tune the step size α using a grid-search over the range $\alpha \in [10^{-4}, 1]$ to obtain the best performance for all the algorithms. The total number of communication rounds is set to $T = 5 \times 10^3$. We selected γ using the results of Theorem 5 (which suggests the noise attenuation parameter be set as $\gamma = \alpha \log T$), where α is the step size. While fine tuning, we found that, in general, the ratio $\frac{\alpha}{\gamma} \in [0.1, 0.01]$ tends to give the best performance. Outside this range, the performance began to steadily decrease with increase or decrease in the ratio. Accordingly, we set the attenuation noise parameter γ to $\gamma = \alpha \times \log T$. The performance results are reported in Figure 1(a)-(b). *For the comparative experiments (Fig 1 (a)-(b)), we incorporated zero-mean Gaussian noise with a variance of $\sigma_c^2 = 0.01$ into the transmitted model estimates for all the compared algorithms.* From the plots, it is evident that IC-GT outperforms all the other algorithms in terms of both the optimality error and the consensus error.

To assess the scalability of IC-GT and examine the impact of graph connectivity on its convergence accuracy, we conducted experiments with varying network sizes, specifically $n \in \{5, 10, 15, 20, 25\}$. We kept the noise variance fixed at $\sigma_c^2 = 0.01$ for the following graph topologies: (i) Fully connected (f.c.), (ii) Erdős-Rényi graph with an edge probability of 0.5 (rand), (iii) Ring topology, and (iv) Star topology. From Figure 1(c), we observe that as the graph connectivity deteriorates, the final performance of IC-GT also deteriorates. In the case of a fully connected graph, there is an improvement in performance with an increasing number of nodes due to a decrease in gradient variance resulting from an increased effective mini-batch size. Finally, we also investigate the effect of varying σ_c^2 on the performance of IC-GT, as depicted in Figure 1(d).

Neural network based experiments

In this subsection, we investigate a deep learning scenario that involves random compressed communication using probabilistic quantization (see (3)). We assume a star-based topology with $n = 10$ for both the MNIST and CIFAR datasets. For the MNIST dataset, we utilize a learning model with a total of 8.4K parameters. This model comprises two convolution layers, the first with 250 parameters and the second with 5K parameters, followed by a fully connected layer with 3.2K parameters. For the CIFAR-10 dataset, we adopt the standard LENET architecture, which consists of three convolution layers and two fully connected layers. This architecture has a total of

0.54M parameters. The configuration of the max-pooling and batch normalization layers follows the standard settings used in `LENET` models.

We compare `IC-GT` with two other strategies commonly employed to address noise in an inexact communication setting. The first strategy involves utilizing a decreasing noise variance policy, where the variance decreases as the number of communication rounds progresses. In this approach, we employ `GT` with quantization and adjust the quantization levels to become finer as the rounds increase. Specifically, in the case of (3), we increase the parameter Δ_p from $\Delta_p = 1$ to $\Delta_p = 5 \times 10^3$ as the rounds progress. This results in higher levels of noise variance in the initial rounds and lower levels in the final rounds. The second strategy maintains a uniform quantization level of $\Delta_p = 10^2$ throughout all communication rounds, leading to a fixed noise variance. We employ the same quantization level of $\Delta_p = 10^2$ for `IC-GT`.

The results of the comparison have been plotted in Figure 2(a)-(b). In both plots, the baseline represents the highest achievable accuracy that can be obtained in a centralized setting using the models employed. From the plots, we observe that for both the `CIFAR-10` and `MNIST` datasets, the performance of `IC-GT` is the closest to the baseline. The performance difference between `IC-GT` and the baseline appears to be more pronounced in the case of `CIFAR-10` compared to `MNIST`.

VI. Final Remarks

In this paper, we proposed a gradient tracking based algorithm for decentralized optimization in an inexact communication scenario. We established theoretical convergence guarantees and analyzed the impact of communication and gradient noise on performance. Our algorithm effectively mitigates the impact of communication noise and data heterogeneity, and achieves optimal iteration complexity for strongly convex, stochastic smooth functions. Experimental results on logistic regression and neural networks demonstrated the superiority of the proposed algorithm over existing methods. As future work, the algorithm can be extended to other settings, such as directed graphs and asynchronous updates, and incorporate variance reduction techniques to enhance convergence rate. Moreover, probabilistic quantization can be subject to saturation errors ([39], [40]) and other complex quantization approaches, which are commonly used in machine learning, may involve different noise models that do not satisfy Assumption 2 [41]. Investigating the impact of such errors on `IC-GT` and decentralized optimization algorithms, in general, could be another potential avenue for future research.

Appendix I: Proof of Lemma 1

Proof: From (5), we have,

$$\mathbf{v}_k = (\mathbf{I}_n - \gamma\mathbf{Q}')\mathbf{x}_k + \gamma\hat{\mathbf{Q}}\epsilon_{k,c}. \tag{46}$$

Multiplying both sides of (46) by $\frac{1}{n}1_n1_n^T \otimes I_d$, we get,

$$\begin{aligned}\bar{\mathbf{v}}_k &= \bar{\mathbf{x}}_k + \frac{\gamma}{n}\big(1_n1_n^T \otimes I_d\big)\hat{\mathbf{Q}}\epsilon_{k,c}\\ &= (\mathbf{I}_n - \gamma\mathbf{Q}')\bar{\mathbf{x}}_k + \frac{\gamma}{n}\big(1_n1_n^T \otimes I_d\big)\hat{\mathbf{Q}}\epsilon_{k,c}.\end{aligned} \tag{47}$$

where we used $\frac{1}{n}1_n^T\mathbf{Q}' = 0$ to get the first inequality and $(\mathbf{I}_n - \gamma\mathbf{Q}')\bar{\mathbf{x}}_k = \bar{\mathbf{x}}_k$ to get the last inequality. Subtracting (47) from (46) and adding $-\frac{1}{n}1_n1_n^T\Delta\mathbf{x}_k$, we get,

$$\Delta\mathbf{v}_k = (\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')\Delta\mathbf{x}_k + \gamma\tilde{\mathbf{Q}}\epsilon_{k,c}. \tag{48}$$

From (6), the expression for $\Delta\mathbf{x}_k$ can be written as,

$$\Delta\mathbf{x}_k = \Delta\mathbf{v}_{k-1} - \alpha\Delta\mathbf{y}_{k-1}. \tag{49}$$

Substituting for $\Delta\mathbf{x}_k$ in (48) using (49) yields the following recursive relation for $\Delta\mathbf{v}_k$ in terms of $\Delta\mathbf{v}_{k-1}$ and $\Delta\mathbf{y}_{k-1}$:

$$\Delta\mathbf{v}_k = (\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')\Delta\mathbf{v}_{k-1} - \alpha(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')\Delta\mathbf{y}_{k-1} + \gamma\tilde{\mathbf{Q}}\epsilon_{k,c}$$

Next, the recursive relation for $\Delta\mathbf{x}_k$ in terms of $\Delta\mathbf{x}_{k-1}$ and $\Delta\mathbf{y}_{k-1}$ is obtained by substituting for $\Delta\mathbf{v}_{k-1}$ in (49) using (48). That is,

$$\Delta\mathbf{x}_k = (\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')\Delta\mathbf{x}_{k-1} + \gamma\tilde{\mathbf{Q}}\epsilon_{k-1,c} - \alpha\Delta\mathbf{y}_{k-1}.$$

The recursive expression for $\Delta\mathbf{y}_k$ can be obtained similarly using the expression for $\bar{\mathbf{y}}_k$ and subtracting it from (7), concluding the proof.

Appendix II: Proof of Lemma 2

Proof: Using mathematical induction, we can show that $\mathbf{J}_\gamma^\tau$ for any $\tau \in \mathbb{N}$ is given as,

$$\mathbf{J}_\gamma^\tau = \begin{bmatrix}(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau & 0 & -\tau(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau\\ 0 & (\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau & -\tau(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^{(\tau-1)}\\ 0 & 0 & (\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau\end{bmatrix}. \tag{50}$$

Taking norms in (50) and using triangle inequality, we get,

$$\|\mathbf{J}_\gamma^\tau\| \le \|(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau\| + \tau\|(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^{(\tau-1)}\| + \tau\|(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau\|. \tag{51}$$

We will next bound the terms on the right hand side of (51). Note that the smallest eigenvalue of the matrix $(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau$ is zero and the remaining eigenvalues are of the form $(1 - \gamma(1 - \lambda_i))^\tau$ for $i = 2, \ldots, n$, where λ_i are the eigenvalues of Q defined in Assumption 1. Therefore,

$$\begin{aligned}\|(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau\| &= \max_{i=2,\ldots,n}(1 - \gamma(1 - \lambda_i))^\tau\\ &= (1 - \gamma(1 - \lambda_2))^\tau.\end{aligned} \tag{52}$$

From (14), it follows that $\tau \ge 2\left(1 - \frac{\ln\sqrt{\delta}/4}{\gamma(1-\lambda_2)}\right) > -\frac{\ln\sqrt{\delta}/2}{\gamma(1-\lambda_2)}$. Substituting this inequality in (52), we get,

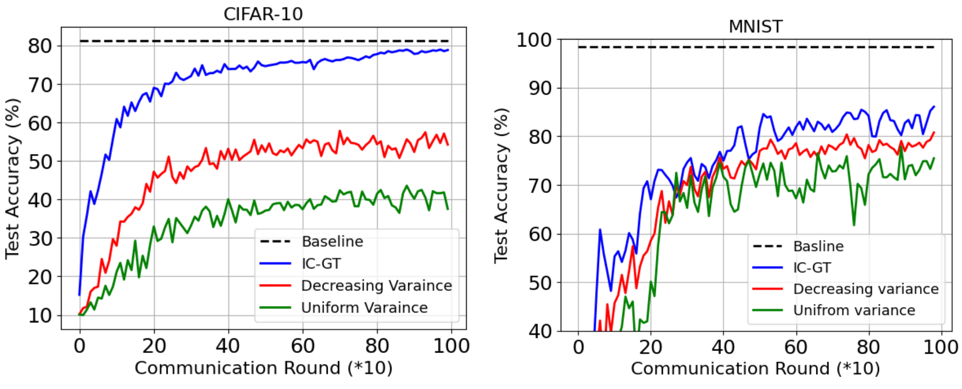
$$\big(1 - \gamma(1 - \lambda_2)\big)^\tau \le \exp\big(-\tau\gamma(1 - \lambda_2)\big) \le \frac{\sqrt{\delta}}{2}. \tag{53}$$

We next bound the second term in (51). For convenience, we define $\mathbf{Q}_1 \overset{\text{def}}{=} \tau(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^{\tau-1}$. The smallest eigenvalue of $\mathbf{Q}_1$ is zero and the remaining eigenvalues are of the form $\tau(1 - \gamma(1 - \lambda_i))^{\tau-1}$, for $i = 2, \ldots, n$. Therefore,

$$\|\mathbf{Q}_1\| \le \tau(1 - \gamma(1 - \lambda_2))^{\tau-1} \le \tau\exp(-(\tau - 1)\gamma(1 - \lambda_2)). \tag{54}$$

Taking logarithm on both sides of (54) yields,

$$\ln\|\mathbf{Q}_1\| \le \ln\tau - (\tau - 1)\gamma(1 - \lambda_2). \tag{55}$$



(a)

(b)

Fig. 2: **(a)** Test Accuracy vs. communication rounds for CIFAR-10 dataset with star topology ($n = 10$). **(b)** Test Accuracy vs. communication rounds for MNIST dataset with star topology ($n = 10$).

Now, consider $\frac{\ln\tau}{\tau}$ as a function of τ and observe that it is monotonically decreasing for any $\tau > \exp(1)$ since its first derivative $\frac{1-\ln\tau}{\tau^2} < 0$. From (14), we have $\tau \geq 16\ln 4 > \exp(1)$ since $\gamma < 1/4$ and $\lambda_2 \in (-1, 1)$. For convenience, we define $\epsilon_{\gamma,\lambda_2} = \frac{\gamma(1-\lambda_2)}{2} \in [0, 1/4)$. Therefore, from(14), it follows that,

$$\frac{\ln\tau}{\tau} \leq \epsilon_{\gamma,\lambda_2}\frac{\ln\frac{4}{\epsilon_{\gamma,\lambda_2}} + \ln\ln\frac{1}{\epsilon_{\gamma,\lambda_2}}}{4\ln 1/\epsilon_{\gamma,\lambda_2}} \leq \epsilon_{\gamma,\lambda_2} = \frac{\gamma(1-\lambda_2)}{2}. \tag{56}$$

Using (56) and $\tau \geq 2\left(1 - \frac{\ln\sqrt{\delta}/4}{\gamma(1-\lambda_2)}\right)$ in (55), we get,

$$\begin{aligned}\ln\|\mathbf{Q}_1\| &\leq \ln\tau - (\tau-1)\gamma(1-\lambda_2)\\ &\leq \frac{\tau\gamma(1-\lambda_2)}{2} - (\tau-1)\gamma(1-\lambda_2)\\ &= \gamma(1-\lambda_2)\left(1 - \frac{\tau}{2}\right)\\ &\leq \ln\sqrt{\delta}/4.\end{aligned}$$

Therefore,

$$\|\mathbf{Q}_1\| \leq \sqrt{\delta}/4. \tag{57}$$

Finally, we bound the third term in (51) as,

$$\tau\|(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^\tau\| \leq \|\mathbf{Q}_1\| \leq \sqrt{\delta}/4. \tag{58}$$

Combining, (51), (55),(57) and (58), we get,

$$\|\mathbf{J}_\gamma^\tau\|^2 \leq \left(\frac{\sqrt{\delta}}{2} + \frac{\sqrt{\delta}}{4} + \frac{\sqrt{\delta}}{4}\right)^2 = \delta.$$

.

Appendix III: Proof of Lemma 3

Proof: We begin by iterating the relation (12) with $k = t + \tau$:

$$\begin{aligned}\Psi_{t+\tau} &= \mathbf{J}_\gamma\Psi_{t+\tau-1} + \alpha\mathbf{E}_{t+\tau-1}\\ &= \mathbf{J}_\gamma^\tau\Psi_t + \alpha\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}\end{aligned} \tag{59}$$

We next consider $\mathbf{E}_k$ whose definition is recalled here:

$$\begin{aligned}&\mathbf{E}_{k-1} =\\ &\underbrace{\frac{\gamma}{\alpha}\begin{bmatrix}\tilde{\mathbf{Q}}\boldsymbol{\epsilon}_{k,c}\\ \tilde{\mathbf{Q}}\boldsymbol{\epsilon_{k-1,c}}\\ \alpha\tilde{\mathbf{Q}}\boldsymbol{\epsilon_{k-1,c}}\end{bmatrix}}_{\mathbf{E}_{k-1}^c} + \underbrace{\begin{bmatrix}0\\0\\ \bar{\mathbf{I}}_n(\nabla\mathbf{F}(\mathbf{x}_k,\boldsymbol{\xi}_k) - \nabla\mathbf{F}(\mathbf{x}_{k-1},\boldsymbol{\xi}_{k-1}))\end{bmatrix}}_{\mathbf{E}_{k-1}^g}\end{aligned} \tag{60}$$

for all $k \in \mathbb{N}$. We note that

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}\right\|^2\right] &\leq 2\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}^c\right\|^2\right]\\ &+ 2\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}^g\right\|^2\right]\end{aligned} \tag{61}$$

We first bound the first term on the right hand side of (61). Using the expression for the matrix product $\mathbf{J}_\gamma^{\tau-i-1}$ for any $0 \leq i \leq \tau - 1$ (cf. (50)), we have,

$$\begin{aligned}&\mathbf{J}_\gamma^{\tau-i-1}\ \mathbf{E}_{t+i}^c = \frac{\gamma}{\alpha}\times\\ &\begin{bmatrix}(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^{\tau-i-1}\tilde{\mathbf{Q}}(\boldsymbol{\epsilon}_{t+i+1,c} - \alpha(\tau-i-1)\boldsymbol{\epsilon}_{t+i,c})\\ (\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^{\tau-i-1}\tilde{\mathbf{Q}} - \alpha(\tau-i-1)(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^{\tau-i-2}\tilde{\mathbf{Q}})\boldsymbol{\epsilon}_{t+i,c}\\ \alpha(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^{\tau-i-1}\tilde{\mathbf{Q}}\boldsymbol{\epsilon}_{t+i,c}\end{bmatrix}.\end{aligned} \tag{62}$$

Note that, using $\|Q\|$, $\|\mathrm{diag}(Q)\| \leq 1$, we have,

$$\begin{aligned}\|(\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')^{\tau-i-1}\|^2 \leq 1,\ \|\hat{\mathbf{Q}}\| = \|(Q' - \mathrm{diag}(Q'))\otimes I_d\| \leq 2,\\ \|\tilde{\mathbf{Q}}\| \leq \left\|\mathbf{I}_n - \frac{11^T}{n}\right\|\|\hat{\mathbf{Q}}\| \leq 2.\end{aligned} \tag{63}$$

Taking norms in (62) and using the bounds (63), we get,

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}^c\|^2\right] \leq \frac{4\gamma^2}{\alpha^2}(2(1+\alpha^2(\tau-i-1)^2)+\alpha^2)\times\\ &\qquad\max\{\mathbb{E}\left[\|\boldsymbol{\epsilon}_{t+i+1,c}\|^2\right], \mathbb{E}\left[\|\boldsymbol{\epsilon}_{t+i,c}\|^2\right]\}\\ &\leq \frac{4\gamma^2}{\alpha^2}(2+2\alpha^2\tau^2+\alpha^2)\max\{\mathbb{E}\left[\|\boldsymbol{\epsilon}_{t+i+1,c}\|^2\right], \mathbb{E}\left[\|\boldsymbol{\epsilon}_{t+i,c}\|^2\right]\}\\ &\leq \frac{8\gamma^2}{\alpha^2}(1+\alpha^2\left(\tau^2+1/2\right)n\sigma_c^2,\end{aligned} \tag{64}$$

where the last inequality is due to Assumption 2. Hence,

$$\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\ \mathbf{E}_{t+i}^c\right\|^2\right] = \sum_{i=0}^{\tau-1}\mathbb{E}\left[\left\|\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}^c\right\|^2\right] \leq \frac{8\gamma^2}{\alpha^2}\big(1+\alpha^2(\tau^2+1/2)\big)n\sigma_c^2\tau, \tag{65}$$

where the equality is due to Assumption 2 and the fact that the cross terms of the form $\langle\boldsymbol{\epsilon}_{i,c},\boldsymbol{\epsilon}_{j,c}\rangle$ are all zero. That is, if we denote $\mathcal{F}_k \overset{\text{def}}{=} \sigma(\mathbf{x}_0,\boldsymbol{\xi}_0,\boldsymbol{\epsilon}_{0,c},\cdots,\boldsymbol{\xi}_{k-1},\boldsymbol{\epsilon}_{k-1,c})$ to be the sigma algebra generated by the random variables up to iteration k, we have for any i,j with $i<j$, $\mathbb{E}[\langle\boldsymbol{\epsilon}_{i,c},\boldsymbol{\epsilon}_{j,c}\rangle] = \mathbb{E}[\mathbb{E}\left[\langle\boldsymbol{\epsilon}_{i,c},\boldsymbol{\epsilon}_{j,c}\rangle|\mathcal{F}_j\right]] = 0$.

Next we consider $\mathbf{E}_k^g$ to bound the second summation in (61). Let $\mathbf{g}_k \overset{\text{def}}{=} \nabla\mathbf{F}(\mathbf{x}_k,\boldsymbol{\xi}_k) - \nabla\mathbf{f}(\mathbf{x}_k)$ and $\mathbf{d}_k = \nabla\mathbf{f}(\mathbf{x}_k) - \nabla\mathbf{f}(\mathbf{x}^*)$. We note from Assumption 4, $\mathbf{g}_k$ is a zero mean vector given $\mathbf{x}_k$ with variance $n\sigma_g^2$. Using $\bar{\mathbf{Q}} \overset{\text{def}}{=} (\bar{\mathbf{I}}_n - \gamma\mathbf{Q}')$ and the expression for the matrix product $\mathbf{J}_\gamma^{\tau-i-1}$ (cf. (50)), we have,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}^g\right\|^2\right] \\
&= \mathbb{E}\Big[\Big\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\nabla\mathbf{F}(\mathbf{x}_{t+i+1},\boldsymbol{\xi}_{t+i+1}) \\
&\qquad\qquad - \nabla\mathbf{F}(\mathbf{x}_{t+i},\boldsymbol{\xi}_{t+i}))\Big\|^2\Big] \\
&+ \mathbb{E}\Big[\Big\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-2}\bar{\mathbf{I}}_n(\nabla\mathbf{F}(\mathbf{x}_{t+i+1},\boldsymbol{\xi}_{t+i+1}) \\
&\qquad\qquad - \nabla\mathbf{F}(\mathbf{x}_{t+i},\boldsymbol{\xi}_{t+i}))\Big\|^2\Big] \\
&+ \mathbb{E}\Big[\Big\|\sum_{i=0}^{\tau-1}\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\nabla\mathbf{F}(\mathbf{x}_{t+i+1},\boldsymbol{\xi}_{t+i+1}) \\
&\qquad\qquad - \nabla\mathbf{F}(\mathbf{x}_{t+i},\boldsymbol{\xi}_{t+i}))\Big\|^2\Big] \\
&\leq 2\Bigg\{\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{g}_{t+i+1}-\mathbf{g}_{t+i})\right\|^2\right] \\
&\qquad + \mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{d}_{t+i+1}-\mathbf{d}_{t+i})\right\|^2\right] \\
&+ \mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-2}\bar{\mathbf{I}}_n(\mathbf{g}_{t+i+1}-\mathbf{g}_{t+i})\right\|^2\right] \\
&\qquad + \mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-2}\bar{\mathbf{I}}_n(\mathbf{d}_{t+i+1}-\mathbf{d}_{t+i})\right\|^2\right] \\
&+ \mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{g}_{t+i+1}-\mathbf{g}_{t+i})\right\|^2\right] \\
&\qquad + \mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{d}_{t+i+1}-\mathbf{d}_{t+i})\right\|^2\right]\Bigg\},
\end{aligned} \tag{66}$$

where the inequality is obtained by adding and subtracting the terms $\nabla\mathbf{f}(\mathbf{x}_{t+i+1})$, $\nabla\mathbf{f}(\mathbf{x}_{t+i})$, and $\nabla\mathbf{f}(\mathbf{x}^*)$ in each of the three terms in the first equality. We bound the first term on the right hand side of (66) and follow a similar approach to bound the rest of the terms. However, before proceeding, we state the following fact whose proof is provided at the end of this appendix:

$$\left\|(i+1)\bar{\mathbf{Q}}^{i+1}\bar{\mathbf{I}}_n - i\bar{\mathbf{Q}}^i\bar{\mathbf{I}}_n\right\|^2 \leq 4, \quad \forall i\in\mathbb{N}. \tag{67}$$

The first term on the right hand side of (66) is bounded as,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{g}_{t+i+1}-\mathbf{g}_{t+i})\right\|^2\right] \\
&= \mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{\tau-1}\big((\tau-i)\bar{\mathbf{Q}}^{\tau-i}\bar{\mathbf{I}}_n - (\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-1}\big)\mathbf{g}_{t+i} \\
&\qquad - (\tau-1)\bar{\mathbf{Q}}^{\tau-1}\bar{\mathbf{I}}_n\mathbf{g}_t\Bigg\|^2\Bigg] \\
&\leq \sum_{i=1}^{\tau-1}\left\|(\tau-i)\bar{\mathbf{Q}}^{\tau-1}\bar{\mathbf{I}}_n - (\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n\right\|^2\mathbb{E}[\|\mathbf{g}_{t+i}\|^2] \\
&\qquad + \left\|(\tau-1)\bar{\mathbf{Q}}^{\tau-1}\bar{\mathbf{I}}_n\right\|^2\mathbb{E}[\|\mathbf{g}_t\|^2] \\
&\leq 4\sum_{i=1}^{\tau-1}\mathbb{E}[\|\mathbf{g}_{t+i}\|^2] + n\sigma_g^2 \leq 4\tau n\sigma_g^2
\end{aligned} \tag{68}$$

where the first inequality is due to Assumption 4 and the fact that the cross terms of the form $\mathbb{E}[\langle\mathbf{g}_p,\mathbf{g}_{p'}\rangle] = 0$, for any $p<p'$, and the second the inequality is due to Assumption 4, (67), and the fact that $\tau\|\mathbf{I}_n - \gamma\mathbf{Q}'\|^{\tau-1}\|\bar{\mathbf{I}}_n\| \leq 1$(cf. (57)). Following a similar approach, we can bound the rest of the terms involving $\mathbf{g}_k$ as:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-2}\bar{\mathbf{I}}_n(\mathbf{g}_{t+i+1}-\mathbf{g}_{t+i})\right\|^2\right] \leq 4\tau n\sigma_g^2 \\
&\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{g}_{t+i+1}-\mathbf{g}_{t+i})\right\|^2\right] \leq 4(\tau+1)n\sigma_g^2.
\end{aligned} \tag{69}$$

Similarly, considering the second term in (66), we have,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}(\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{d}_{t+i+1}-\mathbf{d}_{t+i})\right\|^2\right] \\
&\quad \leq \tau\Big((\tau-1)\|\bar{\mathbf{Q}}^{\tau-1}\bar{\mathbf{I}}_n\|^2\mathbb{E}[\|\mathbf{d}_t\|^2] \\
&+ \sum_{i=1}^{\tau-1}\left\|(\tau-i)\bar{\mathbf{Q}}^{\tau-1}\bar{\mathbf{I}}_n - (\tau-i-1)\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n\right\|^2\mathbb{E}[\|\mathbf{d}_{t+i}\|^2]\Big) \\
&\quad \leq 4\tau\sum_{i=0}^{\tau-1}\mathbb{E}[\|\mathbf{d}_{t+i}\|^2],
\end{aligned} \tag{70}$$

where the first inequality is due to the fact that $\left\|\sum_{i=0}^{\tau-1}a_i\right\|^2 \leq \tau\sum_{i=0}^{\tau-1}\|a_i\|^2$ for any $a\in\mathbb{R}^d$. The same bound also holds for the fourth term in (66) while for the last term, we have,

$$\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{d}_{t+i+1}-\mathbf{d}_{t+i})\right\|^2\right] \leq 4\tau\sum_{i=0}^{\tau}\mathbb{E}[\|\mathbf{d}_{t+i}\|^2]. \tag{71}$$

We next bound the summation $\sum_{i=0}^{\tau}\mathbb{E}[\|\mathbf{d}_{t+i}\|^2$. For all $i<\tau$:

$$\begin{aligned}\|\mathbf{d}_{t+i}\|^2 &= \|\nabla\mathbf{f}(\mathbf{x}_{t+i})-\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})+\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2\\ &\leq 2L^2\|\Psi_{t+i}\|^2+2\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2,\end{aligned} \tag{72}$$

where the second inequality is due to Assumption 3. Now, for $i=\tau$, we have,

$$\begin{aligned}\|\mathbf{d}_{t+\tau}\|^2 &\leq 2L^2\|\Psi_{t+\tau}\|^2+2\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2\\ &\leq 2L^2\|\Psi_{t+\tau}\|^2+4\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau})-\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})\|^2\\ &\qquad+4\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2\\ &\leq 2L^2\|\Psi_{t+\tau}\|^2+4L^2\|\bar{\mathbf{x}}_{t+\tau}-\bar{\mathbf{x}}_{t+\tau-1}\|^2\\ &\qquad+4\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2.\end{aligned} \tag{73}$$

The expression for $\bar{\mathbf{x}}_{t+\tau}$ can be written as (cf. (28)),

$$\begin{aligned}\bar{\mathbf{x}}_{t+\tau} = \bar{\mathbf{x}}_{t+\tau-1} &- \frac{\alpha}{n}\big(1_n1_n^T\otimes I_d\big)\nabla\mathbf{f}(\mathbf{x}_{t+\tau-1})+\alpha\boldsymbol{\epsilon}_{t+\tau-1,g}\\ &+\gamma\bar{\boldsymbol{\epsilon}}_{t+\tau-1,c}-\alpha\gamma\sum_{j=0}^{t+\tau-2}\bar{\boldsymbol{\epsilon}}_{j,c},\end{aligned}$$

where $\boldsymbol{\epsilon}_{k,g}\overset{\text{def}}{=}\frac{1}{n}\big(1_n1_n^T\otimes I_d\big)\Big(\nabla\mathbf{f}(\mathbf{x}_k)-\nabla\mathbf{F}(\mathbf{x}_k,\boldsymbol{\xi}_k)\Big)$ and $\bar{\boldsymbol{\epsilon}}_{k,c}\overset{\text{def}}{=}\frac{1}{n}\big(1_n1_n^T\otimes I_d\big)\hat{\mathbf{Q}}\boldsymbol{\epsilon}_{k,c}$ for any $k\in\mathbb{N}$. Taking square norms and expectations, we get,

$$\begin{aligned}&\mathbb{E}[\|\bar{\mathbf{x}}_{t+\tau}-\bar{\mathbf{x}}_{t+\tau-1}\|^2|\mathcal{F}_{t+\tau-1}]\\ &\leq 2\mathbb{E}\Big[\Big\|\frac{\alpha}{n}\big(1_n1_n^T\otimes I_d\big)\left(\nabla\mathbf{f}(\mathbf{x}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\right)+\alpha\boldsymbol{\epsilon}_{t+\tau-1,g}\\ &+\gamma\bar{\boldsymbol{\epsilon}}_{t+\tau-1,c}\Big\|^2\Big|\mathcal{F}_{t+\tau-1}\Big]+2\alpha^2\gamma^2\mathbb{E}\left[\left\|\sum_{j=0}^{t+\tau-2}\bar{\boldsymbol{\epsilon}}_{j,c}\right\|^2\Big|\mathcal{F}_{t+\tau-1}\right],\end{aligned} \tag{74}$$

where we used the fact that $\frac{1}{n}\left(1_n1_n^T\otimes I_d\right)\nabla\mathbf{f}(\mathbf{x}^*)=0$. From Assumptions 2 and 4, we have for all $k\in\mathbb{N}$,

$$\mathbb{E}[\alpha\boldsymbol{\epsilon}_{k,g}+\gamma\bar{\boldsymbol{\epsilon}}_{k,c}|\mathcal{F}_k]=0,\ \mathbb{E}[\|\alpha\boldsymbol{\epsilon}_{k,g}+\gamma\bar{\boldsymbol{\epsilon}}_{k,c}\|^2]\leq\alpha^2\sigma_g^2+\gamma^2\sigma_c^2, \tag{75}$$

and

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\bar{\boldsymbol{\epsilon}}_{j,c}\right\|^2\right] &= \mathbb{E}\left[\sum_{j=0}^{k-1}\|\bar{\boldsymbol{\epsilon}}_{j,c}\|^2\right]+\sum_{\substack{1\leq p,p'\\ \leq k-1}}\mathbb{E}\left[\langle\bar{\boldsymbol{\epsilon}}_{p,c},\bar{\boldsymbol{\epsilon}}_{p',c}\rangle\right]\\ &\leq\sum_{j=0}^{k-1}\sigma_c^2=k\sigma_c^2,\end{aligned} \tag{76}$$

where the last inequality is due to the fact that $\mathbb{E}[\langle\bar{\boldsymbol{\epsilon}}_{p,c},\bar{\boldsymbol{\epsilon}}_{p',c}\rangle]=\mathbb{E}[\mathbb{E}\left[\langle\bar{\boldsymbol{\epsilon}}_{p,c},\bar{\boldsymbol{\epsilon}}_{p',c}\rangle|\mathcal{F}_{p'}\right]]=0$ for any $p<p'$. Combining (74), (75) and (76), we have,

$$\begin{aligned}&\mathbb{E}[\|\bar{\mathbf{x}}_{t+\tau}-\bar{\mathbf{x}}_{t+\tau-1}\|^2]\\ &\leq 2\mathbb{E}\left[\left\|\frac{\alpha}{n}\big(1_n1_n^T\otimes I_d\big)\left(\nabla\mathbf{f}(\mathbf{x}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\right)\right\|^2\right]\\ &\qquad+2\alpha^2\sigma_g^2+2\left(1+\alpha^2(t+\tau)\right)\gamma^2\sigma_c^2\\ &\leq 2\mathbb{E}\Big[\Big\|\frac{\alpha}{n}\big(1_n1_n^T\otimes I_d\big)\Big(\nabla\mathbf{f}(\mathbf{x}_{t+\tau-1})-\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})\\ &\qquad+\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\Big)\Big\|^2\Big]\\ &\qquad+2\alpha^2\sigma_g^2+2\left(1+\alpha^2(t+\tau)\right)\gamma^2\sigma_c^2\\ &\leq 4\alpha^2L^2\mathbb{E}\left[\|\Psi_{t+\tau-1}\|^2\right]+4\alpha^2\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2\right]\\ &\qquad+2\alpha^2\sigma_g^2+2\left(1+\alpha^2(t+\tau)\right)\gamma^2\sigma_c^2.\end{aligned} \tag{77}$$

Taking expectations in (73) and using (77), we get,

$$\begin{aligned}&\mathbb{E}[\|\mathbf{d}_{t+\tau}\|^2]\\ &\leq 2L^2\mathbb{E}[\|\Psi_{t+\tau}\|^2]+4\mathbb{E}[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2]\\ &+16L^4\alpha^2\mathbb{E}[\|\Psi_{t+\tau-1}\|]^2+16\alpha^2L^2\mathbb{E}[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2]\\ &\qquad+8L^2\left(\alpha^2\sigma_g^2+\gamma^2(1+\alpha^2(t+\tau))\sigma_c^2\right)\\ &\leq 2L^2\mathbb{E}[\|\Psi_{t+\tau}\|^2]+5\mathbb{E}[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+\tau-1})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2]\\ &+L^2\mathbb{E}[\|\Psi_{t+\tau-1}\|^2]+8L^2\left(\alpha^2\sigma_g^2+\gamma^2(1+\alpha^2(t+\tau))\sigma_c^2\right),\end{aligned} \tag{78}$$

where the last inequality is due to $\alpha^2<1/16L^2$. Using (72) and (78) in (71), we have,

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\bar{\mathbf{Q}}^{\tau-i-1}\bar{\mathbf{I}}_n(\mathbf{d}_{t+i+1}-\mathbf{d}_{t+i})\right\|^2\right]\\ &\leq 12\tau L^2\sum_{i=0}^{\tau-1}\mathbb{E}[\|\Psi_{t+i}\|^2]+28\tau\sum_{i=0}^{\tau-1}\mathbb{E}[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2]\\ &+8\tau L^2\mathbb{E}[\|\Psi_{t+\tau}\|^2]+32\tau L^2\left(\alpha^2\sigma_g^2+\gamma^2(1+\alpha^2(t+\tau))\sigma_c^2\right)\end{aligned} \tag{79}$$

The rest of the terms involving d_k in (66) can be bounded in the same manner. Using (68), (69) and (79) in (66), we get,

$$\begin{aligned}\sum_{i=0}^{\tau-1}\mathbb{E}\left[\|\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}^g\|^2\right] &\leq 24n(\tau+1)\sigma_g^2+72\tau L^2\sum_{i=0}^{\tau-1}\mathbb{E}[\|\Psi_{t+i}\|^2]\\ &+168\tau\sum_{i=0}^{\tau-1}\mathbb{E}[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2]+48\tau L^2\mathbb{E}[\|\Psi_{t+\tau}\|^2]\\ &\qquad+192\tau L^2\left(\alpha^2\sigma_g^2+\gamma^2(1+\alpha^2(t+\tau))\sigma_c^2\right).\end{aligned} \tag{80}$$

Using (65) and (80) to bound the right hand side in (61), we have,

$$\begin{aligned}
&\mathbb{E}\left[\Big\|\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}\Big\|^2\right]\\
&\leq \frac{16\gamma^2}{\alpha^2}\big(1+\alpha^2(\tau^2+1/2)\big)n\sigma_c^2\tau+144\tau L^2\sum_{i=0}^{\tau-1}\mathbb{E}[\|\Psi_{t+i}\|^2]\\
&\quad+96\tau L^2\mathbb{E}[\|\Psi_{t+\tau}\|^2]+336\tau\sum_{i=0}^{\tau-1}\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2\right]\\
&\quad+48n(\tau+1)\sigma_g^2+384\tau L^2\left(\alpha^2\sigma_g^2+\gamma^2(1+\alpha^2(t+\tau))\sigma_c^2\right)\\
&\leq 144\tau L^2\sum_{i=0}^{\tau-1}\mathbb{E}[\|\Psi_{t+i}\|^2]+\frac{1}{4\alpha^2}\mathbb{E}[\|\Psi_{t+\tau}\|^2]\\
&\quad+336\tau\sum_{i=0}^{\tau-1}\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2\right]+49n(\tau+1)\sigma_g^2\\
&\quad+\frac{16\gamma^2}{\alpha^2}\big(2+\alpha^2(\tau^2+1/2)+\alpha^2(t+\tau)\big)n\sigma_c^2\tau \qquad (81)
\end{aligned}$$

where we used $\alpha^2 < 1/384\tau L^2$. Next, taking square norms and expectations in (59), we get,

$$\begin{aligned}
\mathbb{E}\left[\|\Psi_{t+\tau}\|^2\right]&=\mathbb{E}\left[\left\|\mathbf{J}_\gamma^\tau\Psi_t+\alpha\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}\right\|^2\right]\\
&\leq 2\mathbb{E}\left[\left\|\mathbf{J}_\gamma^\tau\Psi_t\right\|^2\right]+2\alpha^2\mathbb{E}\left[\left\|\sum_{i=0}^{\tau-1}\mathbf{J}_\gamma^{\tau-i-1}\mathbf{E}_{t+i}\right\|^2\right].
\end{aligned} \qquad (82)$$

From Lemma 2, it follows that there there exists a τ such that $\|\mathbf{J}_\gamma^\tau\|^2 \leq \frac{1}{4}\rho'$ for a given $\rho' \in (0, 1/4]$. Therefore,

$$4\|\mathbf{J}_\gamma^\tau\Psi_t\|^2 \leq 4\|\mathbf{J}_\gamma^\tau\|^2\|\Psi_t\|^2 \leq \rho'\|\Psi_t\|^2. \qquad (83)$$

To conclude, we substitute (81) and (83) in (82) to get the required inequality,

$$\begin{aligned}
&\mathbb{E}\left[\|\Psi_{t+\tau}\|^2\right]\leq 2\mathbb{E}\left[\left\|\mathbf{J}_\gamma^\tau\Psi_t\right\|^2\right]+288\alpha^2\tau L^2\sum_{i=0}^{\tau-1}\mathbb{E}\left[\|\Psi_{t+i}\|^2\right]\\
&+672\alpha^2\tau\sum_{i=0}^{\tau-1}\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2\right]+\frac{1}{2}\mathbb{E}[\|\Psi_{t+\tau}\|^2]+\\
&32\gamma^2\big(2+\alpha^2(\tau^2+1/2)+\alpha^2(t+\tau)\big)n\sigma_c^2\tau+98n(\tau+1)\alpha^2\sigma_g^2\\
&\mathbb{E}\left[\|\Psi_{t+\tau}\|^2\right]\leq\rho'\mathbb{E}\|\Psi_t\|^2+576\alpha^2\tau L^2\sum_{i=0}^{\tau-1}\mathbb{E}[\|\Psi_{t+i}\|^2]+\\
&1344\alpha^2\tau\sum_{i=0}^{\tau-1}\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{x}}_{t+i})-\nabla\mathbf{f}(\mathbf{x}^*)\|^2\right]+196n(\tau+1)\alpha^2\sigma_g^2\\
&\qquad+64\gamma^2\big(2+\alpha^2(\tau^2+1/2)+\alpha^2(t+\tau)\big)n\sigma_c^2\tau,
\end{aligned}$$

which proves the bound (15). The bound (16) for $\ell < \tau$ is proved exactly along the same lines with the only modification being that the first term is scaled by $\|\mathbf{J}_\gamma^\ell\|^2$, $\ell < \tau$ instead of $\|\mathbf{J}_\gamma^\tau\|^2$. The former can be bounded by using the expression for $\mathbf{J}_\gamma^\ell$ (cf. (50)) as follows:

$$\begin{aligned}
\|\mathbf{J}_\gamma^\ell\Psi_0\|^2&\leq\left\|(\bar{\mathbf{I}}_n-\gamma\mathbf{Q}')^\ell-\ell(\bar{\mathbf{I}}_n-\gamma\mathbf{Q}')^\ell\right\|^2\|\Delta\mathbf{v}_0\|^2\\
&\quad+\left\|(\bar{\mathbf{I}}_n-\gamma\mathbf{Q}')^\ell-\ell(\bar{\mathbf{I}}_n-\gamma\mathbf{Q}')^{\ell-1}\right\|^2\|\Delta\mathbf{x}_0\|^2\\
&\quad+\alpha^2\left\|\left(\bar{\mathbf{I}}_n-\gamma\mathbf{Q}'\right)^\ell\right\|^2\|\Delta\mathbf{y}_0\|^2\\
&\leq 2(1+\ell^2)(\|\Delta\mathbf{v}_0\|^2+\|\Delta\mathbf{x}_0\|^2+\alpha^2\|\Delta\mathbf{y}_0\|^2)\\
&\leq 2(1+\tau^2)\|\Psi_0\|^2
\end{aligned}$$

where the second inequality is due to $\left\|\left(\bar{\mathbf{I}}_n-\gamma\mathbf{Q}'\right)^\ell\right\|^2 \leq 1$ and the last inequality is due to $\ell < \tau$.

To conclude, we provide the proof of (67).

Claim: $\|(i+1)\bar{\mathbf{Q}}^{i+1}\bar{\mathbf{I}}_n - i\bar{\mathbf{Q}}^i\bar{\mathbf{I}}_n\|^2 \leq 4$.

Proof: We have

$$\begin{aligned}
&\|(i+1)\bar{\mathbf{Q}}^{i+1}\bar{\mathbf{I}}_n-i\bar{\mathbf{Q}}^i\bar{\mathbf{I}}_n\|^2\\
&\leq\|(i+1)\bar{\mathbf{Q}}^{i+1}-i\bar{\mathbf{Q}}^i\|^2\|\bar{\mathbf{I}}_n\|^2\\
&\leq\max_{j\in[n]}|(i+1)(1-\gamma(1-\lambda_j)^{i+1}-i(1-\gamma(1-\lambda_j)^i|^2\\
&=\left|(1+i)(1-\gamma(1-\bar{\lambda}))^{i+1}-i(1-\gamma(1-\bar{\lambda}))^i\right|^2\\
&\qquad\qquad\qquad\qquad\qquad\text{(for some }\bar{\lambda})\\
&=\left|(1-\gamma(1-\bar{\lambda}))^{i+1}+i(1-\gamma(1-\bar{\lambda}))^i(1-\gamma(1-\bar{\lambda})-1)\right|^2\\
&=|(1-\gamma(1-\bar{\lambda}))^{i+1}-i\gamma(1-\bar{\lambda})(1-\gamma(1-\bar{\lambda}))^i|^2\\
&\leq 2|(1-\gamma(1-\bar{\lambda}))|^{2i+2}+2\gamma^2(1-\bar{\lambda})^2\Big(\underbrace{i(1-\gamma(1-\bar{\lambda}))^i}_{\leq\frac{1}{\gamma(1-\bar{\lambda})}}\Big)^2\\
&\leq 4
\end{aligned}$$

where the second inequality is due to $\|\bar{\mathbf{I}}_n\| \leq 1$ and the last inequality is due to $\gamma(1-\bar{\lambda}) \in [0, 1]$.

References

[1] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE transactions on automatic control*, vol. 31, no. 9, pp. 803–812, 1986.

[2] J. N. Tsitsiklis, "Problems in decentralized decision making and computation." Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, Tech. Rep., 1984.

[3] J. Wu, W. Huang, J. Huang, and T. Zhang, "Error compensated quantized sgd and its applications to large-scale distributed optimization," in *International conference on machine learning*. PMLR, 2018, pp. 5325–5333.

[4] C. Iakovidou and E. Wei, "S-near-dgd: A flexible distributed stochastic gradient method for inexact communication," *IEEE Transactions on Automatic Control*, 2022.

[5] A. Koloskova, T. Lin, and S. U. Stich, "An improved analysis of gradient tracking for decentralized machine learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 422–11 435, 2021.

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

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

[8] K. Yuan, W. Xu, and Q. Ling, "Can primal methods outperform primal-dual methods in decentralized dynamic optimization?" *IEEE Transactions on Signal Processing*, vol. 68, pp. 4466–4480, 2020.

[9] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2017.

[10] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized sgd with changing topology and local updates," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5381–5393.

[11] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Mathematical Programming*, vol. 187, no. 1, pp. 409–457, 2021.

[12] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, 2019.

[13] C. Iakovidou and E. Wei, "S-near-dgd: A flexible distributed stochastic gradient method for inexact communication," *IEEE Transactions on Automatic Control*, 2022.

[14] T. T. Doan, S. T. Maguluri, and J. Romberg, "Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach," *IEEE Transactions on Automatic Control*, vol. 66, no. 10, pp. 4469–4484, 2020.

[15] R. L. Cavalcante and S. Stanczak, "A distributed subgradient method for dynamic convex optimization problems under noisy information exchange," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 243–256, 2013.

[16] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE journal of selected topics in signal processing*, vol. 5, no. 4, pp. 772–790, 2011.

[17] C.-S. Lee, N. Michelusi, and G. Scutari, "Finite rate quantized distributed optimization with geometric convergence," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 1876–1880.

[18] T. T. Doan, S. T. Maguluri, and J. Romberg, "Fast convergence rates of distributed subgradient methods with adaptive quantization," *IEEE Transactions on Automatic Control*, vol. 66, no. 5, pp. 2191–2205, 2020.

[19] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc wsns with noisy links—part i: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2007.

[20] K. Srivastava, A. Nedić, and D. M. Stipanović, "Distributed constrained optimization over noisy networks," in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 1945–1950.

[21] X. Liu, Y. Li, R. Wang, J. Tang, and M. Yan, "Linear convergent decentralized optimization with compression," *arXiv preprint arXiv:2007.00232*, 2020.

[22] Y. Liao, Z. Li, K. Huang, and S. Pu, "A compressed gradient tracking method for decentralized optimization with linear convergence," *IEEE Transactions on Automatic Control*, vol. 67, no. 10, pp. 5622–5629, 2022.

[23] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4494–4506, 2019.

[24] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "Robustness of accelerated first-order algorithms for strongly convex optimization problems," *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2480–2495, 2020.

[25] D. Zelazo and M. Mesbahi, "Edge agreement: Graph-theoretic performance bounds and passivity analysis," *IEEE Transactions on Automatic Control*, vol. 56, no. 3, pp. 544–555, 2010.

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

[27] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D^2 : Decentralized training over decentralized data," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4848–4856.

[28] A. Nedić, A. Olshevsky, W. Shi, and C. A. Uribe, "Geometrically convergent distributed optimization with uncoordinated step-sizes," in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 3950–3955.

[29] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 8353–8358.

[30] B. Li, S. Cen, Y. Chen, and Y. Chi, "Communication-efficient distributed optimization in networks with gradient tracking and variance reduction," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1662–1672.

[31] A. S. Berahas, R. Bollapragada, and S. Gupta, "Balancing communication and computation in gradient tracking algorithms for decentralized optimization," *Journal of Optimization Theory and Applications*, pp. 1–34, 2024.

[32] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized sgd with changing topology and local updates," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5381–5393.

[33] H. Reisizadeh, B. Touri, and S. Mohajer, "Distributed optimization over time-varying graphs with imperfect sharing of information," *IEEE Transactions on Automatic Control*, vol. 68, no. 7, pp. 4420–4427, 2023.

[34] ——, "Dimix: Diminishing mixing for sloppy agents," *SIAM Journal on Optimization*, vol. 33, no. 2, pp. 978–1005, 2023.

[35] D. R. Pauluzzi and N. C. Beaulieu, "A comparison of snr estimation techniques for the awgn channel," *IEEE Transactions on communications*, vol. 48, no. 10, pp. 1681–1691, 2000.

[36] S. Stein, "Fading channel issues in system engineering," *IEEE Journal on selected areas in communications*, vol. 5, no. 2, pp. 68–89, 1987.

[37] T. S. Rappaport, *Wireless communications: Principles and practice*. Pearson Education India, 2010.

[38] D. Yuan, S. Xu, H. Zhao, and L. Rong, "Distributed dual averaging method for multi-agent optimization with quantized communication," *Systems & Control Letters*, vol. 61, no. 11, pp. 1053–1061, 2012.

[39] B. Widrow and I. Kollár, *Quantization noise: roundoff error in digital computation, signal processing, control, and communications*. Cambridge University Press, 2008.

[40] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 805–812, 1993.

[41] L. Chen, W. Liu, Y. Chen, and W. Wang, "Communication-efficient design for quantized decentralized federated learning," *IEEE Transactions on Signal Processing*, vol. 72, pp. 1175–1188, 2024.

[42] L. M. Nguyen, P. H. Nguyen, P. Richtárik, K. Scheinberg, M. Takác, and M. van Dijk, "New convergence aspects of stochastic gradient algorithms," *JMLR*, vol. 20, pp. 1–49, 2019.

[43] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[44] B. Widrow, J. McCool, M. Larimore, and C. Johnson, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proceedings of the IEEE*, vol. 64, no. 8, pp. 1151–1162, 1976.

[45] S. M. Shah and R. Bollapragada, "A stochastic gradient tracking algorithm for decentralized optimization with inexact communication," *arXiv preprint arXiv:2307.14942*, 2020.

[46] Y. Nesterov, "Introductory lectures on convex programming volume i: Basic course," *Lecture notes*, vol. 3, no. 4, p. 5, 1998.

Raghu Bollapragada is an assistant professor in the Operations Research and Industrial Engineering program at the University of Texas at Austin (UT). Before joining UT, he was a postdoctoral researcher in the Mathematics and Computer Science Division at Argonne National Laboratory. He received both PhD and MS degrees in Industrial Engineering and Management Sciences from Northwestern University. He has received the ME Walker Scholar award at UT, the IEMS Nemhauser Dissertation Award, the IEMS Arthur P. Hurter Award, and the McCormick terminal year fellowship at Northwestern University.

Suhail M. Shah is currently a research associate in the Operations Research and Industrial Engineering Program at the University of Texas at Austin. He obtained his PhD from the Electrical Engineering (EE) Dept. at Indian Institute of Technology, Bombay in 2019, working under the supervision of Professor Vivek Borkar. Since then, he has held various visiting and research scientist positions at Boston University, Microsoft Research Lab at Bangalore and Huawei Labs at HKUST among others.