Single Point-Based Distributed Zeroth-Order Optimization with a Non-Convex Stochastic Objective Function

Elissa Mhanna [* 1] **Mohamad Assaad** [* 1]

Abstract

Zero-order (ZO) optimization is a powerful tool for dealing with realistic constraints. On the other hand, the gradient-tracking (GT) technique proved to be an efficient method for distributed optimization aiming to achieve consensus. However, it is a first-order (FO) method that requires knowledge of the gradient, which is not always possible in practice. In this work, we introduce a zero-order distributed optimization method based on a one-point estimate of the gradient tracking technique. We prove that this new technique converges with a single noisy function query at a time in the non-convex setting. We then establish a convergence rate of $O(\frac{1}{\sqrt[3]{K}})$ after a number of iterations K, which competes with that of $O(\frac{1}{\sqrt[4]{K}})$ of its centralized counterparts. Finally, a numerical example validates our theoretical results.

1. Introduction

We consider a network encompassing a set of agents $\mathcal{N} = \{1, 2, \ldots, n\}$, where the communication is limited to network neighbors. Each agent i maintains a local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$ and the aim is for the agents to collaboratively find the decision variable $x \in \mathbb{R}^d$ that minimizes the stochastic global objective function:

$$\min_{x \in \mathbb{R}^d} \mathcal{F}(x) = \frac{1}{n} \sum_{i=1}^{n} F_i(x) \tag{1}$$

with

$$F_i(x) = \mathbb{E}_{\xi \sim D_i} f_i(x, \xi), \tag{2}$$

where ξ denotes an i.i.d. ergodic stochastic process that follows a local distribution D_i. Hence, the goal is for agents

to achieve consensus while the local objective functions are kept private. Further, these functions are assumed to be non-convex, and to solve this problem, agent i can only query the function values of f_i at exactly *one point* at a time. These function queries are assumed to be noisy $\tilde{f}_i = f_i + \zeta_i$ with ζ_i some additive noise. We note that ζ_i is added to differentiate between the stochastic nature of f_i and the noise imposed on its querying. While the expectation of the local function might not change with this addition, its variance increases, presenting an added difficulty to the analysis

Distributed optimization appears across many applications, in distributed data processing and control in sensor networks, parallel computing and data storing, and big-data analytics. It has many interests in machine learning applications such as data rights, privacy, and scalability (Lian et al., 2017; Nedic, 2020). One important technique in this domain is the gradient tracking method, which is agnostic to the data distribution (Koloskova et al., 2022), meaning there's no need for data to be i.i.d across all the devices. In addition, this technique was shown to have interesting rates. For example, it converges linearly to the optimal solution with constant step size (Qu & Li, 2018; Nedić et al., 2017; Pu, 2020).

However, all these references assume the availability of first-order information, which is not always the case, especially when gradient computation can be impractical like when the system is constantly changing, costly when the system is too complex or too huge in dimension, or even impossible such as in black-box optimization. For example, when training a deep neural network, the loss output is an intricate function of the weights, which complicates the derivation of the gradient (Lian et al., 2016). In bandit settings, an adversary reveals a cost value at every period, and the player must compete with this adversary without any knowledge of the underlying cost function (Flaxman et al., 2004). In centralized and distributed settings, much work has been done on zero-order (ZO) optimization. ZO methods include substituting the gradient by one-point estimates with one functional value at a time (Li & Assaad, 2021; Flaxman et al., 2004; Roy et al., 2022; Mhanna & Assaad, 2022),

[1] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France. Correspondence to: Elissa Mhanna <elissa.mhanna@centralesupelec.fr>, Mohamad Assaad <mohamad.assaad@centralesupelec.fr>.

generally of the form

$$g = \frac{d}{\gamma} f(x + \gamma z, \xi) z, \qquad (3)$$

with $\gamma > 0$ a small value and z a random vector with symmetrical distribution. It includes two- and multi-points estimates (Agarwal et al., 2010; Roy et al., 2022; Balasubramanian & Ghadimi, 2018; Tang et al., 2021) having the respective forms

$$g = d \frac{f(x + \gamma z, \xi) - f(x - \gamma z, \xi)}{2\gamma} z \qquad (4)$$

$$\text{and } g = \sum_{j=1}^{d} \frac{f(x + \gamma e_j, \xi) - f(x - \gamma e_j, \xi)}{2\gamma} e_j \qquad (5)$$

where $\{e_j\}_{j=1,...,d}$ is the canonical basis and other techniques which are derivative-free (Golovin et al., 2019; Bubeck et al., 2021).

While the gradient tracking technique has been extended to the ZO case (Tang et al., 2021; Mhanna & Assaad, 2022), the prior utilizes a multi-point estimator of the gradient and a static objective function *without* any stochasticity/noise whatsoever in the system, and the latter deals with the problem in the *strongly convex* setting. Both of which do not satisfy the assumptions of our problem.

As various stochastic sources and noisy environments can be present in the model, such as distinct device data distributions, uncertainties in electronic devices, discretization errors, lossy data compression, and time-varying communication channels, static objective functions no longer serve realistic purposes. Moreover, multi-point estimates assume that it is possible to observe multiple realizations of the objective function under the same system conditions, as can be seen from (4) and (5), many function queries are done for the same realization of ξ. This assumption does not hold when the environment is continually changing. For example, in wireless networks, when the query is an estimate is subject to a wireless channel between the transmitter and the user, that usually changes every $1 - 2$ ms. Thus, by the time the next observation is queried, a stochastic process will have already impacted the new function value.

While the distributed convex setting is present in many applications, countless problems call for non-convex optimization, such as in distributed learning (Omidshafiei et al., 2017) and resource allocation (Tychogiorgos et al., 2013).

For all these reasons, we extend the gradient tracking technique to stochastic objective functions and one-point gradient estimators in the non-convex setting.

1.1. Related Work

Centralized ZO Optimization: Adversarial convex bandit problems have been long studied in the zero-order domain.

Flaxman et al. introduce a one-point gradient estimator to the online gradient descent technique to compete with oblivious and adaptive adversaries, which accomplishes a regret bound of $O(K^{\frac{3}{4}})$ (2004). Agarwal et al. then extend the problem to the multi-point gradient estimate (2010); In the noise-free setting, this estimate has bounded and sometimes vanishing variance. Their algorithm acquires regret bounds of $O(\sqrt{K})$ and of $O(\log(K))$ for strongly convex loss functions. Nesterov and Spokoiny study the complexity with a two-point estimator in different noise-free settings (2017). In a stochastic convex case and another static non-convex case, a convergence rate of $O(\frac{1}{\sqrt{K}})$ is attained.

Roy et al. present and examine noise-free stochastic algorithms for general non-convex stochastic objective functions that change with time (2022). They propose nonstationary versions of regret measures and define W_K as a bound on the amount of nonstationarity. W_K is allowed to increase with K. They present a gradient descent algorithm that achieves a regret bound of $O(\sqrt{KW_K})$ with a two-point gradient estimator and $O(K^{\frac{2}{3}} W_K^{\frac{1}{3}})$ with a one-point estimator.

Balasubramanian and Ghadimi also propose and investigate zeroth-order algorithms for stochastic objective functions (2018) in the noiseless setting. They employ structural sparsity assumptions in gradient descent to deal with high dimensions. With a two-point estimate, they acquire a rate of $O(\frac{1}{\sqrt[4]{K}})$ for non-convex optimization and $O(\frac{1}{\sqrt{K}})$ for convex one.

Distributed ZO Optimization: Li and Assaad present a distributed stochastic gradient descent technique in networks with a one-point gradient estimator based on a stochastic perturbation (2021). The communication in the network is restricted to a partial exchange of the observation of the agents' local objectives that each agent uses to estimate the global objective. This objective is subject to a stochastic process, and its solution is distributed where each agent is responsible for optimizing its own scalar variable. Their convergence rate evolves as $O(\frac{1}{\sqrt{K}})$ with the assumption of a strongly concave objective function.

Tang et al. present two consensus-aiming stochastic multi-agent optimization techniques in the static non-convex setting (2021). One technique is based on a 2-point gradient estimator without gradient tracking, and the other is based on a noise-free $2d$-point with gradient tracking. The first achieves a rate of $O(\frac{1}{\sqrt{K}} \log K)$.

Gradient Tracking with FO Information: All Qu and Li (2018), Lorenzo and Scutari (2016), Nedic et al. (2017), Shi et al. (2015), Li et al. (2022), and Jiang et al. (2022) present a variant of the technique with the assumption of accurate gradient information availability.

Table 1. A summary of convergence rates attained in related work.

Reference	Gradient Estimate	Decision	GT	Setting	Convergence Rate
Flaxman et al. (2004)	One-point	Centralized	-	Convex	$O(\frac{1}{\sqrt[4]{K}})$
Roy et al. (2022)	One-point	Centralized	-	Non-Convex	$O(\frac{\sqrt[3]{W_K}}{\sqrt[3]{K}})$
Balasubramanian and Ghadimi (2018)	Two-point	Centralized	-	Non-Convex	$O(\frac{1}{\sqrt[4]{K}})$
Li and Assaad (2021)	One-point	Distributed	$\times$	Strongly Convex	$O(\frac{1}{\sqrt{K}})$
Tang et al. (2021)	Two-point	Distributed	$\times$	Non-Convex	$O(\frac{1}{\sqrt{K}} \log K)$
Pu and Nedić (2018)	Unbiased/BV	Distributed	$\surd$	Strongly Convex	$O(\frac{1}{K})$
Tang et al. (2021)	$2d$-point	Distributed	$\surd$	Non-Convex	$O(\frac{1}{K})$
Mhanna and Assaad (2022)	One-point	Distributed	$\surd$	Strongly Convex	$O(\frac{1}{\sqrt{K}})$
This Paper	One-point	Distributed	$\surd$	Non-Convex	$O(\frac{1}{\sqrt[3]{K}})$

Other references assume access to local stochastic first-order oracles where the gradient estimate is unbiased, and with a bounded variance (Pu & Nedić, 2018; Xin et al., 2019; Pu, 2020; Lu et al., 2019; Koloskova et al., 2022). The first three references consider smooth and strongly convex local objective functions, and the convergence rate is linear when the step size is constant. When the step size is decreasing, it is that of $O(\frac{1}{K})$ (Pu & Nedić, 2018). Pu and Nedić propose a distributed stochastic gradient tracking method (DSGT) and a gossip-like stochastic gradient tracking method (GSGT) that allows the agent to wake up at each round with a certain probability (2018). Xin et al. extend the technique to strongly connected graphs (2019). Pu then extends it to directed graphs assuming noisy information exchange between agents (2020). Both Lu et al. (2019) and Koloskova et al. (2022) study the technique in the non-convex stochastic setting and achieve a convergence rate of $O(\frac{1}{\sqrt{K}})$ under constant step size.

While those are all good rates, comparing with them is not fair as there is naturally a gap in the convergence rate between first-order and zero-order methods. In zero-order techniques, a biased estimate of the gradient is obtained, contrary to first-order methods, where an unbiased estimate is considered. Besides, the use of a constant step size guarantees a convergence near the optimal/stationary point and not to the point itself.

Gradient Tracking with ZO Information: In their gradient tracking method, Tang et al. consider a non-convex static setting (2021). They employ a noise-free $2d$-point gradient estimator, which has a vanishing variance, and their algorithm acquires a rate of $O(\frac{1}{K})$ with constant step size. Mhanna and Assaad then extend this technique to a one-point gradient estimate achieving a rate of $O(\frac{1}{\sqrt{K}})$, but they assume strong convexity of the objective function (2022).

1.2. Challenges and Contribution

The first challenge of our work is that when the objective function is stochastic, the underlying stochastic law is not supposed to be known. Thus, the expectation cannot be computed. And when the observation is noisy, the algorithm's performance would be especially affected as both the variance and the norm squared of the gradient estimate would increase. Our next obstacle involves analyzing the convergence of existing algorithms in non-convex optimization problems without relying on the assumption that the functions involved are strongly convex. In this scenario, convergence analysis requires us to verify that the norm of the exact gradient approaches zero. However, due to the bias in our gradient estimate w.r.t. to the true gradient, even if the estimate approaches zero, it remains challenging to prove that the exact gradient also converges to zero. While all estimates are subject to bias, multi-point cases differ because their variance $\mathbb{E}[\|\nabla \mathcal{F} - g\|^2]$ is always bounded when there is no noise present in the optimization process. However, in our case, for single-point estimates, this variance is unbounded (Liu et al., 2020), meaning the estimate can deviate significantly from the true gradient. This demonstrates both technical and intuitive difficulties. Unlike convex optimization, where the characteristics of the function can guide the algorithm, non-convex optimization lacks such guidance. Additionally, we face the challenge of having access only to the expectation of our gradient estimate at a specific point while aiming to prove that the gradient at the agents' average of that point vanishes. This further complicates the problem. Nevertheless, we successfully overcome all of these challenges in our analyses of convergence and rate.

In this paper, we focus on single-point distributed zero-order optimization, which is motivated by the fact that it's not always possible to query multiple points, especially in stochastic environments, as mentioned earlier. Our contribution can then be summarized as follow:

- We extend the gradient tracking technique in the stochastic non-convex setting to the case where we don't have FO information and have to estimate the gradient with a noisy one-point query.

- We overcome the technical challenges of non-convexity, the stochastic impact of the environment, and biasedness alongside the unlimited variance of our gradient estimate.

- We prove the convergence of the proposed algorithm theoretically using stochastic approximation techniques.

- We prove a convergence rate of $O(\frac{1}{\sqrt[3]{K}})$ is attainable using this algorithm which is comparable to that of ZO non-convex methods as shown in Table 1.

1.3. Notation

Vectors are column-shaped unless otherwise specified and $\mathbf{1}$ denotes the vector of all entries equal to 1. For two vectors $a, b \in \mathbb{R}^n$, $\langle a, b \rangle$ is the inner product. For two matrices A, $B \in \mathbb{R}^{n \times d}$, $\langle A, B \rangle$ is the Frobenius inner product. $\|.\|$ is the 2-norm for vectors and the Frobenius norm for matrices.

1.4. Problem Assumptions

In this section, we introduce the assumptions necessary for convergence of our algorithm.

Assumption 1.1. (the local functions) We assume the Lipschitz continuity of all local objective functions $x \longmapsto f_i(x, \xi)$,

$$\|f_i(x, \xi) - f_i(x', \xi)\| \le L_\xi \|x - x'\|, \ \forall x, x' \in \mathbb{R}^d, \forall i \in \mathcal{N}, \tag{6}$$

where L_ξ denotes the Lipschitz constant. Further, we assume $\mathbb{E}_\xi f_i(x, \xi) < \infty, \forall i \in \mathcal{N}$, to ensure the boundedness of the objective $\mathcal{F}(x)$.

Assumption 1.2. (the objective function) Both $\nabla F_i(x)$ and $\nabla^2 F_i(x)$ exist and are continuous, and there exists a constant $\sigma_1 > 0$ such that

$$\|\nabla^2 F_i(x)\| \le \sigma_1, \ \forall x \in \mathbb{R}^d, \forall i \in \mathcal{N}.$$

Hence, we can say that the stochastic objective function $x \longmapsto \mathcal{F}(x)$ is L-smooth for some positive constant L,

$$\|\nabla \mathcal{F}(x) - \nabla \mathcal{F}(x')\| \le L \|x - x'\|, \ \forall x, x' \in \mathbb{R}^d, \tag{7}$$

or alternatively,

$$\mathcal{F}(x) \le \mathcal{F}(x') + \langle \nabla \mathcal{F}(x'), x - x' \rangle + \frac{L}{2} \|x - x'\|^2. \tag{8}$$

Assumption 1.3. (the additive noise) ζ_i is a zero-mean uncorrelated noise with bounded variance, meaning $E(\zeta_i) = 0$, $E(\zeta_i^2) = \sigma_2 < \infty, \forall i \in \mathcal{N}$, and $E(\zeta_i \zeta_j) = 0$ if $i \ne j$.

Assumption 1.4. (the graph) The network is described by an undirected and connected graph, meaning communication links work in both directions, and between any two agents, we can find a path of links.

We then define the agents' coupling matrix $W = [w_{ij}] \in \mathbb{R}^{n \times n}$, where agents i and j are connected iff $w_{ij} = w_{ji} > 0$ ($w_{ij} = w_{ji} = 0$ otherwise). We assume W is a non-negative matrix and doubly stochastic, i.e., $W\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T W = \mathbf{1}^T$. All diagonal elements w_{ii} are strictly positive.

2. Algorithm Description

This section describes our proposed distributed stochastic gradient tracking method in the non-convex setting with a one-point gradient estimator (1P-DSGT-NC). In the GT technique, every agent i must keep and tend to two variables, a local copy $x_i \in \mathbb{R}^d$ of the decision variable and another auxiliary variable $y_i \in \mathbb{R}^d$. GT entails one 'gradient descent' step and one auxiliary variable update step. The reason for the technique's name is that the auxiliary variable update usually allows it to track the gradient of the global objective function in the network. The gradient descent step updates the local optimization variables using the auxiliary variable, which itself tracks the global gradient. As shown in Algorithm 1, y_i is the sum of its previous iteration and the difference between the new and the old gradient estimates so that the update may include only the new information brought by the new estimate. w_{ij} is the weight of the link between agents i and j, so finally, the algorithm evolves by considering a weighted sum of neighboring updates and never sharing the local objective function nor its gradient (estimate). Moreover, we assume that every local function is influenced by a stochastic variable $\xi_i \in \mathbb{R}^m$, and its query is subject to an additive scalar noise ζ_i. At iteration k, the variables are denoted as $x_{i,k}$, $y_{i,k}$, $\xi_{i,k}$, and $\zeta_{i,k}$.

By querying the function once per algorithm iteration $k \in \mathbb{N}$, agent i assembles its one-point gradient estimator $g_{i,k}$:

$$\begin{aligned} g_{i,k} &= z_{i,k} \tilde{f}_i(x_{i,k} + \gamma_k z_{i,k}, \xi_{i,k}) \\ &= z_{i,k}(f_i(x_{i,k} + \gamma_k z_{i,k}, \xi_{i,k}) + \zeta_{i,k}), \end{aligned} \tag{9}$$

where $z_{i,k} \in \mathbb{R}^d$ is a random perturbation vector generated by agent i, and $\gamma_k > 0$ is a vanishing step size. We let $\eta_k > 0$ be another vanishing step size, and we summarize the algorithm updates in Algorithm 1. We note that at every iteration, every agent $i \in \mathcal{N}$ updates its variables independently and in parallel with all other agents in the network.

We next let $\mathbf{x}_k := [x_{1,k}, x_{2,k}, \dots, x_{n,k}]^T$ and $\mathbf{y}_k := [y_{1,k}, y_{2,k}, \dots, y_{n,k}]^T \in \mathbb{R}^{n \times d}$, and their means $\bar{x}_k := \frac{1}{n} \mathbf{1}^T \mathbf{x}_k$ and $\bar{y}_k := \frac{1}{n} \mathbf{1}^T \mathbf{y}_k \in \mathbb{R}^{1 \times d}$, respectively. The perturbation matrix is defined as

Algorithm 1 The 1P-DSGT-NC Algorithm

Input: Initial arbitrary value $x_{i,0} \in \mathbb{R}^d$, the agents' coupling matrix W, the initial step-sizes η_0 and γ_0, and setting $y_{i,0} = g_{i,0}$

1: **for** $k = 0$ **to** $K - 1$, in parallel on all agents i, **do**
2:     Send $x_{i,k}$ and $y_{i,k}$ to neighbors
3:     Update the decision variable:
      $x_{i,k+1} = \sum_{j=1}^n w_{ij}(x_{j,k} - \eta_k y_{j,k})$
4:     Sample $z_{i,k+1}$ and update η_{k+1} and γ_{k+1}
5:     Query the local function f_i at the point:
      $x_{i,k+1} + \gamma_{k+1} z_{i,k+1}$
6:     Assemble $g_{i,k+1}$ according to (9)
7:     Update the auxiliary variable:
      $y_{i,k+1} = \sum_{j=1}^n w_{ij} y_{j,k} + g_{i,k+1} - g_{i,k}$
8: **end for**

$\mathbf{z}_k := [z_{1,k}, z_{2,k}, \ldots, z_{n,k}]^T$, all $\in \mathbb{R}^{n \times d}$. Denoting the concatenated stochastic variables as $\xi_k := [\xi_{1,k}, \xi_{2,k}, \ldots, \xi_{n,k}]^T \in \mathbb{R}^{n \times m}$, then $\mathbf{g}_k := g(\mathbf{x}_k, \xi_k) := [g_1(x_{1,k}, \xi_{1,k}), g_2(x_{2,k}, \xi_{2,k}), \ldots, g_n(x_{n,k}, \xi_{n,k})]^T \in \mathbb{R}^{n \times d}$ denotes the concatenated gradient estimate of all agents and its mean is $\bar{g} := \frac{1}{n} \mathbf{1}^T \mathbf{g} \in \mathbb{R}^{1 \times d}$.

The algorithm's updates can thus be written in matrix notation:

$$\mathbf{x}_{k+1} = W(\mathbf{x}_k - \eta_k \mathbf{y}_k)$$
$$\text{Query at: } \mathbf{x}_{k+1} + \gamma_{k+1} \mathbf{z}_{k+1} \qquad (10)$$
$$\mathbf{y}_{k+1} = W\mathbf{y}_k + \mathbf{g}_{k+1} - \mathbf{g}_k,$$

with

$$\bar{y}_k = \bar{g}_k \text{ and } \bar{x}_{k+1} = \bar{x}_k - \eta_k \bar{g}_k, \qquad (11)$$

where the first equality in (11) is by $\mathbf{y}_0 = \mathbf{g}_0$ and the recursion of the algorithm's updates, and the second equality is due to the doubly stochastic property of W.

Remark 2.1. We only use matrix notation to simplify the analysis. In fact, all agents perform the update $x_{i,k+1} = \sum_{j=1}^n w_{ij}(x_{j,k} - \eta_k y_{j,k})$ in parallel, so the per-iteration cost is $O(nd)$ (as w_{ij} is a scalar and both $x_{j,k}$ and $y_{j,k}$ are of dimension d). Same for step 7 in Algorithm 1. The sparsity of matrix W helps the complexity by considering at most $m < n$ neighbors for any agent i, then we obtain $O(md)$.

3. Convergence Result

In this section, we present the convergence results of Algorithm 1. We begin by introducing necessary assumptions on the algorithm parameters.

Assumption 3.1. (the step-sizes) η_k and γ_k are two decreasing step sizes that converge to 0 as $k \to \infty$ such that

$$\sum_{k=1}^\infty \eta_k \gamma_k = \infty, \quad \sum_{k=1}^\infty \eta_k \gamma_k^3 < \infty, \text{ and } \sum_{k=1}^\infty \eta_k^2 < \infty.$$

An example of the step sizes satisfying Assumption 3.1 is the following form:

$$\eta_k = \eta_0 (1 + k)^{-\upsilon_1} \text{ and } \gamma_k = \gamma_0 (1 + k)^{-\upsilon_2}. \qquad (12)$$

Then, it's sufficient to find υ_1 and υ_2 such that $0 < \upsilon_1 + \upsilon_2 \leq 1$, $\upsilon_1 + 3\upsilon_2 > 1$, and $\upsilon_1 > 0.5$.

Assumption 3.2. (the random perturbation) The perturbation vector $z_{i,k} = (z_{i,k}^1, z_{i,k}^2, \ldots, z_{i,k}^d)^T$ has the same dimension as the local gradient. It is chosen independently by every agent $i \in \mathcal{N}$ from others and previous samples. Besides, it has i.i.d elements with $\mathbb{E}(z_{i,k}^{d_j})^2 = \sigma_3 > 0, \forall d_j, \forall i$ and there exists a constant $\sigma_4 > 0$ such that $\|z_{i,k}\| \leq \sigma_4, \forall i$.

To generate a perturbation vector satisfying Assumption 3.2, we can choose every dimension of $z_{i,k}$ from the symmetric Bernoulli distribution on $\{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}$. Then, we would have $\sigma_3 = \frac{1}{d}$ and $\sigma_4 = 1$.

Let $\nabla F_i(x_i) \in \mathbb{R}^{1 \times d}$ and $\nabla^2 F_i(x_i) \in \mathbb{R}^{d \times d}$ denote the gradient of F_i at the local variable and its Hessian matrix, respectively. Then, the gradient at $\mathbf{x} \in \mathbb{R}^{n \times d}$ is defined as

$$\nabla F(\mathbf{x}) := [\nabla F_1(x_1), \nabla F_2(x_2), \ldots, \nabla F_n(x_n)]^T \in \mathbb{R}^{n \times d}. \qquad (13)$$

Let $h(\mathbf{x}) := \frac{1}{n} \mathbf{1}^T \nabla F(\mathbf{x}) \in \mathbb{R}^{1 \times d}$ and denote by $\mathcal{H}_k = \{\mathbf{x}_0, \mathbf{y}_0, \xi_0, \ldots, \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, \xi_{k-1}, \mathbf{x}_k\}$ the history sequence.

Proposition 3.3. *(Li & Assaad, 2021; Mhanna & Assaad, 2022) Let Assumptions 1.3 and 3.2 hold. Then, $g_{i,k}$ is a biased estimator of the agent's gradient $\nabla F_i(x_{i,k}), \forall i \in \mathcal{N}$, and*

$$\mathbb{E}_{z,\xi,\zeta}[g_{i,k}|\mathcal{H}_k] = \sigma_3 \gamma_k [\nabla F_i(x_{i,k}) + b_{i,k}], \qquad (14)$$

where $b_{i,k}$ denotes the bias with respect to the true gradient. Refer to Appendix A for details.

Lemma 3.4. *When Assumptions 1.1, 1.3, and 3.2 are satisfied and $\|\mathbf{x}_k\| < \infty$ almost surely, there exists a bounded constant $M > 0$, such that $E[\|\mathbf{g}_k\|^2 |\mathcal{H}_k] < M$ almost surely.*

Proof: Refer to Appendix A.1.

Lemma 3.5. *By Assumption 1.2, we have*

$$\|\nabla \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\| \leq \frac{L}{\sqrt{n}} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|. \qquad (15)$$

Lemma 3.6. *(Qu & Li, 2018) Let Assumption 1.4 hold, then*

$$\|W\omega - \mathbf{1}\bar{\omega}\| \leq \rho_w \|\omega - \mathbf{1}\bar{\omega}\|, \quad \forall \omega \in \mathbb{R}^{n \times d}, \qquad (16)$$

where $\rho_w < 1$ is the spectral norm of the matrix $W - \frac{1}{n} \mathbf{1}\mathbf{1}^T$ and $\bar{\omega} = \frac{1}{n} \mathbf{1}^T \omega$.

The main idea is that we use the smoothness property of the objective function in (7). This property means that when the variable is varied, the gradient's variation does not shoot to infinity and stays confined within a scale of the variable's variation. We then study the variation at the average of the variable from one iteration to the next employing the algorithm's properties in (11). The variation is quantized by a step size multiplied by the average estimate. We use the facts that the expectation of estimate gives us biased access to the exact gradient (Proposition 3.3) and that the norm squared of the estimate is bounded (Lemma 3.4). The smoothness inequality allows finding an upper bound on the exact gradient's norm where the key players are the step sizes: (updating γ_k) How far away from the variable do we need to go to query the function and find an estimate of the gradient at that variable, and (updating α_k) how far the algorithm should advance when updating a descent step? Answering these question provide the right assumptions on the step sizes (Assumption 3.1). Finally, taking the telescoping sum of (8) provides the following technical result and justifies the summation assumptions on the step-sizes.

Theorem 3.7. *Suppose there exists $\mathcal{F}^* := \inf_{x \in \mathbb{R}^d} \mathcal{F} > -\infty$ and denote $\delta_k = \mathcal{F}(\bar{x}_k) - \mathcal{F}^*$. Suppose also that $\|\mathbf{x}_k\| < \infty$ almost surely. Then, when Assumptions 1.1-1.4 and 3.1-3.2 hold, the consensus error $\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$ converges to 0 as $k \to \infty$ and*

$$\sum_{k=0}^{+\infty} \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2 < \infty, \tag{17}$$

which implies

$$\lim_{k \to \infty} \|\nabla \mathcal{F}(\bar{x}_k)\| = 0 \tag{18}$$

almost surely.

Proof: Refer to Appendix B.

4. Convergence Rate

We begin by defining the following constant terms:

$$A_1 = \frac{2}{\sigma_3 \eta_0 \gamma_0} \delta_0 + \frac{4L^2}{n} \frac{1}{1 - \rho_w^2} \|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2$$

$$A_2 = \frac{\sigma_4^6 \sigma_1^2}{2\sigma_3^2} \gamma_0^2 (v_1 + 3v_2)$$

$$A_3 = 2\frac{L\bar{M}\eta_0}{\sigma_3 \gamma_0} v_1 \tag{19}$$

$$A_4 = \frac{12L^2 G^2 \eta_0^2}{n} \frac{\rho_w^2 (1 + \rho_w^2)}{(1 - \rho_w^2)^2} v_1$$

where $\bar{M}, G > 0$ are defined such that the norm squared of the mean estimate $\|\bar{g}_k\|^2 \leq \bar{M}$ and the gradient tracking

error $\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \leq G^2$ (refer to (34) in Appendix B.1 and to (51) in Appendix C.1, respectively).

As the convergence relies on the step sizes, we provide a form to the step sizes to locate an upper bound on the convergence rate. We present the resulting rate in the following theorem.

Theorem 4.1. *Let the step sizes satisfy (12) with the further assumption $0 < v_1 + v_2 < 1$ and suppose $\|\mathbf{x}_k\| < \infty$ almost surely. When Assumptions 1.1-1.4 and 3.1-3.2 are fulfilled,*

$$\frac{\sum_{k=0}^{K} \eta_k \gamma_k \mathbb{E}\left[\|\nabla \mathcal{F}(\bar{x}_k)\|^2\right]}{\sum_{k=0}^{K} \eta_k \gamma_k} \leq \frac{(1 - v_1 - v_2)}{(K + 2)^{1 - v_1 - v_2} - 1}$$

$$\times \left(A_1 + \frac{A_2}{v_1 + 3v_2 - 1} + \frac{A_3}{2v_1 - 1} + \frac{A_4}{3v_1 - 1}\right). \tag{20}$$

Proof: Refer to Appendix C.

Optimizing for the time-varying part $O(\frac{1}{K^{1-v_1-v_2}})$, we find that the best choice for the exponents is $v_1 = 0.5$ and $v_2 = \frac{1}{6}$ for a rate of $O(\frac{1}{\sqrt[3]{K}})$. To avoid the constant term growing too large, one can take $v_1 = 0.5 + \frac{\epsilon}{2}$ and $v_2 = \frac{1}{6} + \frac{\epsilon}{2}$, for an arbitrarily small $\epsilon > 0$, which achieve a rate $O\left(\frac{1}{K^{\frac{1}{3} - \epsilon}}\right)$. We remark that when we use the perturbation vector example in Section 3, our bound scales as $O(d^2)$ in terms of the problem's dimension.

Another interesting aspect of our bound is that the employment of a biased one-point gradient estimate without bounded variance did not affect the dependence on the network size n as compared with the first order-based gradient tracking method, for example, in (Pu & Nedić, 2018). The only dependence on n remains through the spectral norm ρ_w, as $G = O(\sqrt{n})$ and $\bar{M} = O(1)$ and the quantity $\frac{1}{n}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|x_{i,0} - x_{i,0}\|^2$ generally does not depend on n as it's the average of the initial agents' decision variables. Thus, the graph topology determines the bounds' dependence on n.

5. Numerical Example

This section presents a numerical example of our developed algorithm 1P-DSGT-NC to verify its efficacy. We aim to classify m images of two digits taken from the MNIST data set (LeCun & Cortes, 2005) using logistic regression. These images labeled with $y = \{-1, +1\}$ are represented by 784-dimensional vectors X, which are then compressed to dimension $d = 10$ using a lossy autoencoder. They are split equally between n agents, where every agent i can only access their $m_i = \frac{m}{n}$ images. The collective objective is to locate $\theta \in \mathbb{R}^d$ that minimizes the following binary

non-convex loss function:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m_i} \mathbb{E}_u \frac{1}{1 + \exp(-u_{ij} y_{ij} . X_{ij}^T \theta)} + c\|\theta\|^2, \tag{21}$$

where $u \sim \mathcal{N}(1, \sigma_u)$ is a random variable modeling stochastic variations on the querying of local functions. For example, unstable communications while fetching the data points from a local server. c is the regularization constant.

5.1. Simulation Setup

A connected Erdős-Rényi random graph with probability 0.3 is used to model our agent network. We assemble the weight matrix W using the Laplacian method (Qu & Li, 2018). Our algorithm is compared with a distributed stochastic gradient tracking method (DSGT-NC) based on an unbiased estimator with a bounded variance where the objective function is deterministic. This estimator is formed by calculating the exact gradient and adding noise. The final results are simulations averaged over 30 instances.

We first classify the images of the digits 6 and 7. They are $m = 12183$ images in total and divided equally over $n = 31$ agents. Our parameters are generated as follows. The querying noise is $\zeta_{i,k} \sim \mathcal{N}(0,1)$, $\forall i \in \mathcal{N}$, the stochastic variable's standard deviation is $\sigma_u = 0.01$, the regularization constant is $c = 0.1$, the step sizes are $\eta_k = 1.5(k+1)^{-0.51}$ and $\gamma_k = 3.5(k+1)^{-0.17}$, and every dimension of the perturbation vector z_k is chosen from $\{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}$ with equal probability. For the DSGT-NC algorithm, the step size is $\eta_k = 2.5(k+1)^{-0.51}$, and no other noise than that on the exact gradient is considered. Both algorithms are initialized with the same random weights vectors $\theta_{i,0} \sim \mathcal{U}([-1,1]^d)$, $\forall i \in \mathcal{N}$, per simulation instance.

We then classify images of digits 1 and 2, splitting the total $m = 12700$ images over $n = 50$ agents first and $n = 100$ agents second and varying ρ_w for $n = 100$ agents to see its effect. When we kept the same model of Erdős-Rényi, we obtained a smaller ρ_w for a bigger n, i.e for 100 agents, which makes sense for the same probability of edges. To obtain the same ρ_w as for 50 agents, we decreased the probability to 0.2. All other parameters are left the same.

5.2. Simulation Results

Figure 1 describes the development of the expected loss function for the digits 6 and 7 as the iterations advance. As FO methods naturally have better rates, it makes sense that DSGT-NC converges faster than our algorithm. However, the difference is still very small. On the other hand, the much faster convergence of the expected consensus error attained by our algorithm in Figure 2 is surprising. To get a sense of the rate of convergence of this error, we add to the plot in Figure 3 the two graphs $\frac{S_1}{k+1}$ and $\frac{S_2}{k+1}$, where
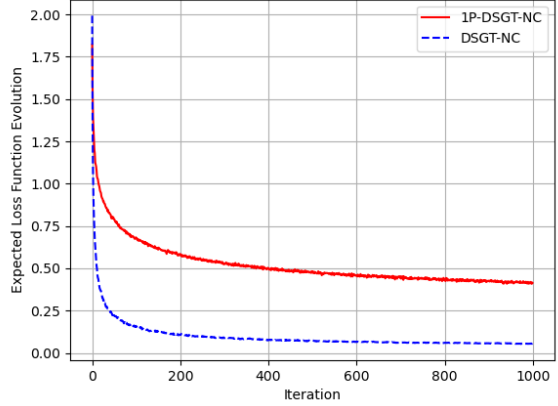

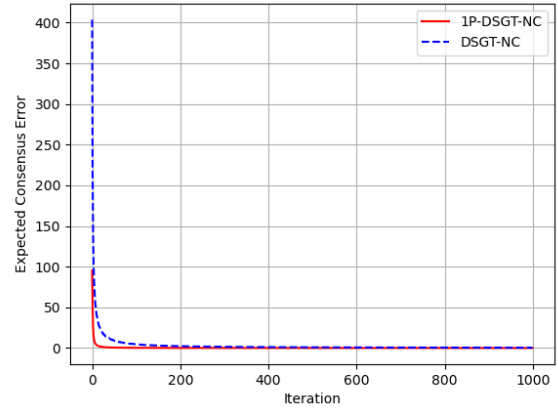*Figure 1.* The evolution of the expected loss function for the digits 6 and 7.

Figure 2. The evolution of the consensus error for the digits 6 and 7.

$S_1 = \mathbb{E}[\|\mathbf{x}_1 - \mathbf{1}\bar{x}_1\|^2]$ the starting expected consensus error resulting from 1P-DSGT-NC and $S_2 = \mathbb{E}[\|\mathbf{x}_1 - \mathbf{1}\bar{x}_1\|^2]$ that resulting from DSGT-NC. We notice that 1P-DSGT-NC has a slightly better rate than $O(\frac{1}{K})$ and DSGT-NC a slightly worse rate.

Figure 4 illustrates the expected gradient tracking error. While both this paper and that of (Pu & Nedić, 2018) prove that this error is bounded by constant and is linear in n, it could be that their constant is affected by other factors, and thus it's much greater.

In Figure 5, we test the accuracy of both algorithms against an independent test set of size 1986 images using the mean decision variable $\bar{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} \theta_{i,k}$ as weight. The final
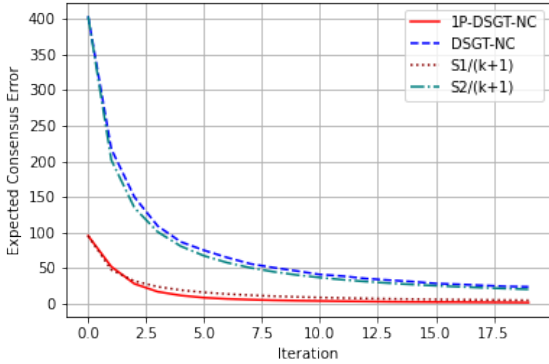
Figure 3. The evolution of the consensus error for the digits 6 and 7 as compared with the rate $O(\frac{1}{K+1})$.
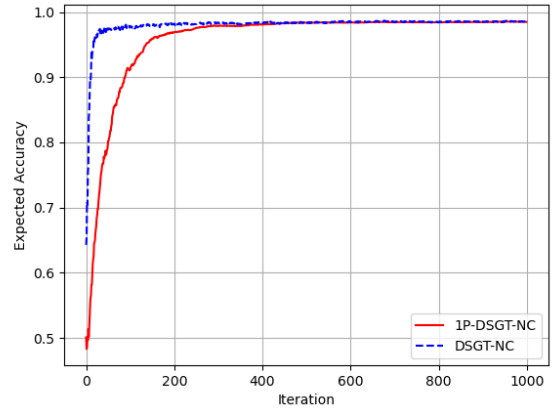

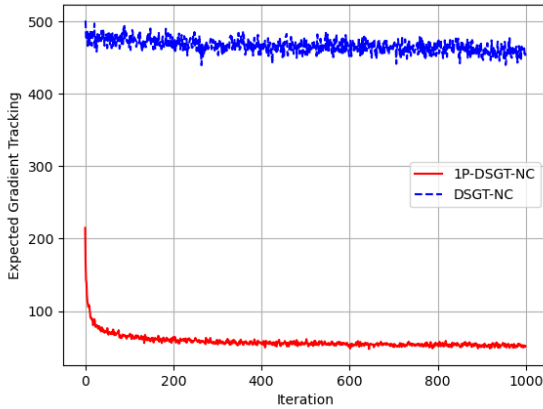*Figure 5.* The evolution of the accuracy for the digits 6 and 7.

Figure 4. The evolution of the gradient tracking error for the digits 6 and 7.

Figure 6. The evolution of the expected loss function for the digits 1 and 2.

accuracy achieved by our algorithm is 98.494461% against 98.539778% of the baseline. We conclude that, finally, the convergence rate did not impact the classification quality as both algorithms achieved excellent results.

Similarly, Figures 6-10 show the same results as previous ones. In Figure 9, the test set is of size 2167 images and the final accuracy with 1P-DSGT-NC is 97.303492% for $(n, \rho_w) = (50, 0.67)$, 97.285033% for $(100, 0.67)$, and 97.340409% for $(100, 0.58)$. With DSGT-NC it is 98.681741% for $(50, 0.67)$, 98.777111% for $(100, 0.67)$, and 98.780188% for $(100, 0.58)$. We notice a slight improvement in the accuracy's convergence rate with the addition of agents for both algorithms in Figures 9 and 10, despite keeping ρ_w the same for one of the examples. The final performance generally stays the same for all cases,
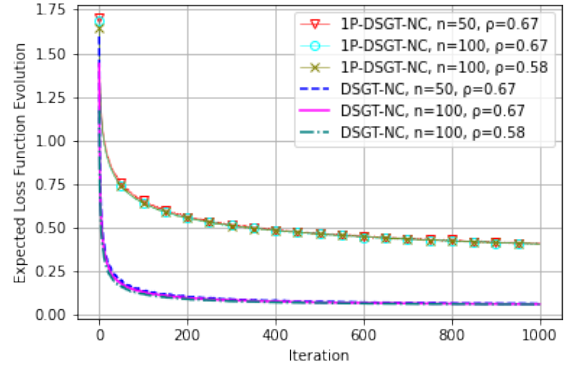
with the edge given to the smaller ρ_w. This is an interesting result as it proves that this algorithm is indeed scalable.

In Figure 8, the linear dependence on n is clear for both algorithms, with a much better bound for 1P-DSGT-NC.

6. Conclusion

In this paper, we have analyzed and demonstrated the convergence of the gradient tracking technique with non-convex objective functions that are subject to a stochastic process. Under the assumption that the gradient is not available and that only noisy single queries of the objective function are available at a time, we assembled a biased estimate of the gradient and proved a convergence rate of $O(\frac{1}{\sqrt[3]{K}})$ is possible. We then confirmed the efficacy of our algorithm with a binary logistic classification simulation model.
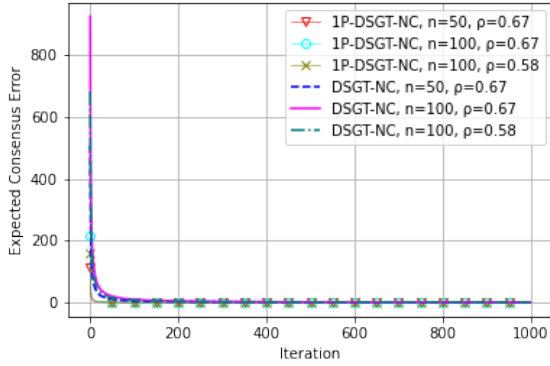
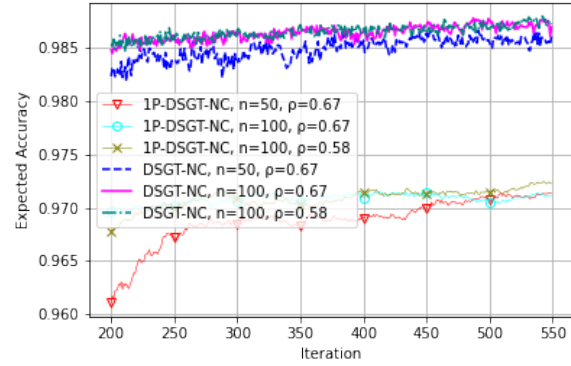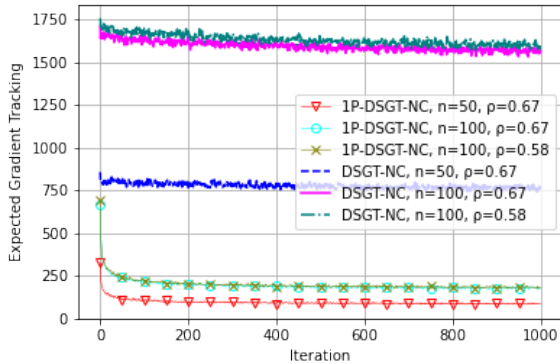*Figure 7.* The evolution of the consensus error for the digits 1 and 2.

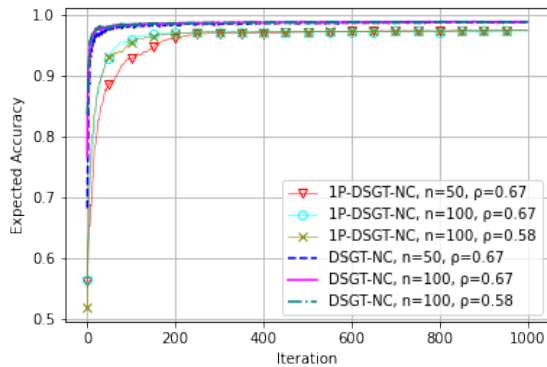*Figure 8.* The evolution of the gradient tracking error for the digits 1 and 2.

Figure 9. The evolution of the accuracy for the digits 1 and 2.

Figure 10. Zooming on the evolution of the accuracy for the digits 1 and 2.

Acknowledgments

References

Agarwal, A., Dekel, O., and Xiao, L. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, 2010.

Balasubramanian, K. and Ghadimi, S. Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality and saddle-points, 2018. URL `https://arxiv.org/abs/1809.06474`.

Bubeck, S., Eldan, R., and Lee, Y. T. Kernel-based methods for bandit convex optimization. *J. ACM*, 68(4), jun 2021. ISSN 0004-5411. doi: 10.1145/3453721. URL `https://doi.org/10.1145/3453721`.

Flaxman, A., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. *CoRR*, cs.LG/0408007, 2004. URL `http://arxiv.org/abs/cs.LG/0408007`.

Golovin, D., Karro, J., Kochanski, G., Lee, C., Song, X., and Zhang, Q. R. Gradientless descent: High-dimensional zeroth-order optimization. *CoRR*, abs/1911.06317, 2019. URL `http://arxiv.org/abs/1911.06317`.

Jiang, X., Zeng, X., Sun, J., and Chen, J. Distributed stochastic gradient tracking algorithm with variance reduction for non-convex optimization. *IEEE Transactions on Neu-*

ral Networks and Learning Systems, pp. 1–12, 2022. doi: 10.1109/TNNLS.2022.3170944.

Koloskova, A., Lin, T., and Stich, S. U. An improved analysis of gradient tracking for decentralized machine learning. 2022. doi: 10.48550/ARXIV.2202.03836. URL https://arxiv.org/abs/2202.03836.

LeCun, Y. and Cortes, C. The mnist database of handwritten digits. 2005.

Li, H., Zheng, L., Wang, Z., Yan, Y., Feng, L., and Guo, J. S-diging: A stochastic gradient tracking algorithm for distributed optimization. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(1):53–65, 2022. doi: 10.1109/TETCI.2020.3017242.

Li, W. and Assaad, M. Distributed stochastic optimization in networks with low informational exchange. *IEEE Transactions on Information Theory*, 67(5):2989–3008, 2021. doi: 10.1109/TIT.2021.3064925.

Lian, X., Zhang, H., Hsieh, C.-J., Huang, Y., and Liu, J. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order, 2016. URL https://arxiv.org/abs/1606.00498.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f75526659f31040afeb61cb7133e4e6d-Paper.pdf.

Liu, S., Chen, P.-Y., Kailkhura, B., Zhang, G., Hero III, A. O., and Varshney, P. K. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020. doi: 10.1109/MSP.2020.3003837.

Lorenzo, P. D. and Scutari, G. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016. doi: 10.1109/TSIPN.2016.2524588.

Lu, S., Zhang, X., Sun, H., and Hong, M. Gnsd: a gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pp. 315–321, 2019. doi: 10.1109/DSW.2019.8755807.

Mhanna, E. and Assaad, M. Zero-order one-point estimate with distributed stochastic gradient-tracking technique, 2022. URL https://arxiv.org/abs/2210.05618.

Nedić, A., Olshevsky, A., and Shi, W. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017. doi: 10.1137/16M1084316. URL https://doi.org/10.1137/16M1084316.

Nedic, A. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020. doi: 10.1109/MSP.2020.2975210.

Nesterov, Y. and Spokoiny, V. G. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Omidshafiei, S., Pazis, J., Amato, C., How, J. P., and Vian, J. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. 2017. doi: 10.48550/ARXIV.1703.06182. URL https://arxiv.org/abs/1703.06182.

Pu, S. A robust gradient tracking method for distributed optimization over directed networks. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 2335–2341, 2020. doi: 10.1109/CDC42340.2020.9303917.

Pu, S. and Nedić, A. Distributed stochastic gradient tracking methods, 2018. URL https://arxiv.org/abs/1805.11454.

Qu, G. and Li, N. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018. doi: 10.1109/TCNS.2017.2698261.

Roy, A., Balasubramanian, K., Ghadimi, S., and Mohapatra, P. Stochastic zeroth-order optimization under nonstationarity and nonconvexity. *Journal of Machine Learning Research*, 23(64):1–47, 2022. URL http://jmlr.org/papers/v23/19-750.html.

Shi, W., Ling, Q., Wu, G., and Yin, W. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015. doi: 10.1137/14096668X. URL https://doi.org/10.1137/14096668X.

Tang, Y., Zhang, J., and Li, N. Distributed zero-order algorithms for nonconvex multiagent optimization. *IEEE Transactions on Control of Network Systems*, 8(1):269–281, 2021. doi: 10.1109/TCNS.2020.3024321.

Tychogiorgos, G., Gkelias, A., and Leung, K. K. A non-convex distributed optimization framework and its application to wireless ad-hoc networks. *IEEE Transactions on Wireless Communications*, 12(9):4286–4296, 2013. doi: 10.1109/TW.2013.072313.120739.

Xin, R., Sahu, A. K., Khan, U. A., and Kar, S. Distributed stochastic optimization with gradient tracking over strongly-connected networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 8353–8358, 2019. doi: 10.1109/CDC40024.2019.9029217.

A. Estimated gradient

In this part, we elaborate on the properties of our gradient estimate. Denote by $\breve{g}_{i,k} = \mathbb{E}_{z,\xi,\zeta}[g_{i,k}|\mathcal{H}_k]$ the expected gradient estimate given the history sequence. Then,

$$
\begin{aligned}
\breve{g}_{i,k} &= \mathbb{E}_{z,\xi,\zeta}[z_{i,k}(f_i(x_{i,k} + \gamma_k z_{i,k}, \xi_{i,k}) + \zeta_{i,k})|\mathcal{H}_k] \\
&= \mathbb{E}_{z,\xi}[z_{i,k}f_i(x_{i,k} + \gamma_k z_{i,k}, \xi_{i,k})|\mathcal{H}_k] \\
&= \mathbb{E}_z[z_{i,k}F_i(x_{i,k} + \gamma_k z_{i,k})|\mathcal{H}_k].
\end{aligned}
\tag{22}
$$

By Taylor's theorem and the mean-valued theorem, there exists $\tilde{x}_{i,k} \in [x_{i,k}, x_{i,k} + \gamma_k z_{i,k}]$ such that

$$
F_i(x_{i,k} + \gamma_k z_{i,k}) = F_i(x_{i,k}) + \gamma_k \langle z_{i,k}, \nabla F_i(x_{i,k})\rangle + \frac{\gamma_k^2}{2}\langle z_{i,k}, \nabla^2 F_i(\tilde{x}_{i,k})z_{i,k}\rangle.
\tag{23}
$$

Substituting in the expected gradient estimate,

$$
\begin{aligned}
\breve{g}_{i,k} &= F_i(x_{i,k})\mathbb{E}_z[z_{i,k}] + \gamma_k \mathbb{E}_z[z_{i,k}z_{i,k}^T]\nabla F_i(x_{i,k}) + \frac{\gamma_k^2}{2}\mathbb{E}_z[z_{i,k}z_{i,k}^T\nabla^2 F_i(\tilde{x}_{i,k})z_{i,k}|\mathcal{H}_k] \\
&= \sigma_3\gamma_k[\nabla F_i(x_{i,k}) + b_{i,k}],
\end{aligned}
\tag{24}
$$

where

$$
\begin{aligned}
b_{i,k} &= \frac{\breve{g}_{i,k}}{\sigma_3\gamma_k} - \nabla F_i(x_{i,k}) \\
&= \frac{\gamma_k}{2\sigma_3}\mathbb{E}_z[z_{i,k}z_{i,k}^T\nabla^2 F_i(\tilde{x}_{i,k})z_{i,k}|\mathcal{H}_k].
\end{aligned}
\tag{25}
$$

By Assumptions 1.2 and 3.2,

$$
\begin{aligned}
\|b_{i,k}\| &\le \frac{\gamma_k}{2\sigma_3}\mathbb{E}_z[\|z_{i,k}\|_2\|z_{i,k}^T\|_2\|\nabla^2 F_i(\tilde{x}_{i,k})\|_2\|z_{i,k}\|_2|\mathcal{H}_k] \\
&\le \gamma_k\frac{\sigma_4^3\sigma_1}{2\sigma_3}.
\end{aligned}
\tag{26}
$$

Thus, the expected value of the mean gradient estimate is

$$
\begin{aligned}
\tilde{g}_k &= \mathbb{E}[\bar{g}_k|\mathcal{H}_k] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[g_{i,k}|\mathcal{H}_k] \\
&= \frac{1}{n}\sum_{i=1}^n \sigma_3\gamma_k[\nabla F_i(x_{i,k}) + b_{i,k}] \\
&= \sigma_3\gamma_k[h(\mathbf{x}_k) + \bar{b}_k],
\end{aligned}
\tag{27}
$$

where the mean bias is

$$
\begin{aligned}
\|\bar{b}_k\| &= \|\frac{1}{n}\sum_{i=1}^n b_{i,k}\| \\
&\le \frac{1}{n}\sum_{i=1}^n \|b_{i,k}\| \\
&\le \frac{1}{n}\sum_{i=1}^n \gamma_k\frac{\sigma_4^3\sigma_1}{2\sigma_3} \\
&= \gamma_k\frac{\sigma_4^3\sigma_1}{2\sigma_3}.
\end{aligned}
\tag{28}
$$

A.1. Gradient Estimate Norm Squared Bound

We start by bounding the expected norm squared of the individual gradient estimates. $\forall i \in \mathcal{N}$,

$$
\begin{aligned}
\mathbb{E}[\|g_{i,k}\|^2|\mathcal{H}_k] &= \mathbb{E}[\|z_{i,k}(f_i(x_{i,k} + \gamma_k z_{i,k}, \xi_{i,k}) + \zeta_{i,k})\|^2|\mathcal{H}_k] \\
&= \mathbb{E}[\|z_{i,k}\|^2\|f_i(x_{i,k} + \gamma_k z_{i,k}, \xi_{i,k}) + \zeta_{i,k}\|^2|\mathcal{H}_k] \\
&\overset{(a)}{\leq} \sigma_4^2 \mathbb{E}[(f_i(x_{i,k} + \gamma_k z_{i,k}, \xi_{i,k}) + \zeta_{i,k})^2|\mathcal{H}_k] \\
&\overset{(b)}{=} \sigma_4^2 \mathbb{E}[f_i^2(x_{i,k} + \gamma_k z_{i,k}, \xi_{i,k})|\mathcal{H}_k] + \sigma_4^2 \sigma_2 \\
&\overset{(c)}{\leq} \sigma_4^2 \mathbb{E}[(\|f_i(0, \xi_{i,k})\| + L_{\xi_{i,k}}\|x_{i,k} + \gamma_k z_{i,k}\|)^2|\mathcal{H}_k] + \sigma_4^2 \sigma_2 \\
&\overset{(d)}{\leq} 2\sigma_4^2 \mathbb{E}[\mu_{\xi_{i,k}}^2 + L_{\xi_{i,k}}^2(\|x_{i,k}\| + \gamma_k \sigma_4)^2|\mathcal{H}_k] + \sigma_4^2 \sigma_2 \\
&\overset{(e)}{=} 2\sigma_4^2(\mu + L'(\|x_{i,k}\| + \gamma_k \sigma_4)^2) + \sigma_4^2 \sigma_2 \\
&:= M_i \\
&< \infty,
\end{aligned}
\tag{29}
$$

where (a) is by Assumption 3.2, (b) Assumption 1.3, (c) Assumption 1.1. In (d), we denote $\|f_i(0, \xi_{i,k})\| = \mu_{\xi_{i,k}}$ and the inequality is due to $\frac{x+y}{2} \leq \sqrt{\frac{x^2+y^2}{2}}$, $\forall x, y \in \mathbb{R}$. In (e), $\mu = \mathbb{E}[\mu_{\xi_{i,k}}^2]$ and $L' = \mathbb{E}[L_{\xi_{i,k}}^2]$.

Then, we can write

$$
\mathbb{E}[\|\mathbf{g}_k\|^2|\mathcal{H}_k] = \sum_{i=1}^n \mathbb{E}[\|g_{i,k}\|^2|\mathcal{H}_k] < n \times \sup_{i \in \mathcal{N}} M_i := M < \infty.
\tag{30}
$$

B. Proof of Convergence

Lemma B.1. *If all Assumptions 1.1, 1.3, 1.4, 3.1, and 3.2 are fulfilled and the inequality $\|\mathbf{x}_k\| < \infty$ holds almost surely, then $\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 = 0$. In addition, both sums*

$$
\sum_{k=0}^\infty \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 < \infty \text{ and } \sum_{k=0}^\infty \gamma_k \eta_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 < \infty
\tag{31}
$$

converge almost surely.

Proof: See Appendix C.1.

B.1. Proof of Theorem 3.7

As a reminder of (11), $\bar{x}_{k+1} = \bar{x}_k - \eta_k \bar{g}_k$. By the L-smoothness of the objective function in Assumption (1.2), we have

$$
\mathcal{F}(\bar{x}_{k+1}) \leq \mathcal{F}(\bar{x}_k) - \eta_k \langle \nabla \mathcal{F}(\bar{x}_k), \bar{g}_k \rangle + \frac{\eta_k^2 L}{2} \|\bar{g}_k\|^2.
\tag{32}
$$

Taking the conditional expectation given $\mathcal{H}_k$ on both sides,

$$\mathbb{E}[\mathcal{F}(\bar{x}_{k+1})|\mathcal{H}_k] \stackrel{(a)}{\leq} \mathcal{F}(\bar{x}_k) - \sigma_3 \eta_k \gamma_k \langle \nabla \mathcal{F}(\bar{x}_k), h(\mathbf{x}_k) + \bar{b}_k \rangle + \frac{\eta_k^2 L}{2} \bar{M}$$

$$= \mathcal{F}(\bar{x}_k) - \sigma_3 \eta_k \gamma_k \langle \nabla \mathcal{F}(\bar{x}_k), h(\mathbf{x}_k) + \bar{b}_k + \mathcal{F}(\bar{x}_k) - \mathcal{F}(\bar{x}_k) \rangle + \frac{L\bar{M}}{2} \eta_k^2$$

$$= \mathcal{F}(\bar{x}_k) - \sigma_3 \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2 - \sigma_3 \eta_k \gamma_k \langle \nabla \mathcal{F}(\bar{x}_k), \bar{b}_k \rangle + \sigma_3 \eta_k \gamma_k \langle \nabla \mathcal{F}(\bar{x}_k), \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k) \rangle + \frac{L\bar{M}}{2} \eta_k^2$$

$$\stackrel{(b)}{\leq} \mathcal{F}(\bar{x}_k) - \sigma_3 \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2 + \sigma_3 \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\| \|\bar{b}_k\| + \sigma_3 \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\| \|\mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\|$$

$$+ \frac{L\bar{M}}{2} \eta_k^2$$

$$\stackrel{(c)}{\leq} \mathcal{F}(\bar{x}_k) - \sigma_3 \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2 + \frac{\sigma_3 \eta_k \gamma_k}{4} \|\nabla \mathcal{F}(\bar{x}_k)\|^2 + \sigma_3 \eta_k \gamma_k \|\bar{b}_k\|^2 + \frac{\sigma_3 \eta_k \gamma_k}{4} \|\nabla \mathcal{F}(\bar{x}_k)\|^2$$

$$+ \sigma_3 \eta_k \gamma_k \|\mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\|^2 + \frac{L\bar{M}}{2} \eta_k^2$$

$$= \mathcal{F}(\bar{x}_k) - \frac{\sigma_3 \eta_k \gamma_k}{2} \|\nabla \mathcal{F}(\bar{x}_k)\|^2 + \sigma_3 \eta_k \gamma_k \|\bar{b}_k\|^2 + \sigma_3 \eta_k \gamma_k \|\mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\|^2 + \frac{L\bar{M}}{2} \eta_k^2$$

$$\stackrel{(d)}{\leq} \mathcal{F}(\bar{x}_k) - \frac{\sigma_3 \eta_k \gamma_k}{2} \|\nabla \mathcal{F}(\bar{x}_k)\|^2 + \frac{\sigma_4^6 \sigma_1^2}{4\sigma_3} \eta_k \gamma_k^3 + \frac{\sigma_3 L^2}{n} \eta_k \gamma_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{L\bar{M}}{2} \eta_k^2.$$

(33)

where (a) is by (27) and (34), (b) is by Cauchy-Schwarz inequality, (c) is by $-2\epsilon \times \frac{1}{\epsilon} \langle a, b \rangle = -2\langle \epsilon a, \frac{1}{\epsilon} b \rangle \leq \epsilon^2 \|a\|^2 + \frac{1}{\epsilon^2} \|b\|^2$, and (d) is by (28) and Lemma 3.5.

By Lemma 3.4,

$$\|\bar{g}_k\|^2 = \|\frac{1}{n} \sum_{i=1}^n g_{i,k}\|^2 \leq \frac{n}{n^2} \sum_{i=1}^n \|g_{i,k}\|^2 = \frac{1}{n} \sum_{i=1}^n \|g_{i,k}\|^2 \leq \bar{M} < \infty, \text{ almost surely.}$$

(34)

By taking the telescoping sum of (33),

$$\mathbb{E}[\mathcal{F}(\bar{x}_{K+1})|\mathcal{H}_K] \leq \mathcal{F}(\bar{x}_0) - \frac{\sigma_3}{2} \sum_{k=0}^K \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2 + \frac{\sigma_4^6 \sigma_1^2}{4\sigma_3} \sum_{k=0}^K \eta_k \gamma_k^3 + \frac{\sigma_3 L^2}{n} \sum_{k=0}^K \eta_k \gamma_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{L\bar{M}}{2} \sum_{k=0}^K \eta_k^2.$$

(35)

We know that $\delta_k = \mathcal{F}(\bar{x}_k) - \mathcal{F}^* > 0$ by definition. Then,

$$0 \leq \mathbb{E}[\delta_{K+1}|\mathcal{H}_K] \leq \delta_0 - \frac{\sigma_3}{2} \sum_{k=0}^K \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2 + \frac{\sigma_4^6 \sigma_1^2}{4\sigma_3} \sum_{k=0}^K \eta_k \gamma_k^3 + \frac{\sigma_3 L^2}{n} \sum_{k=0}^K \eta_k \gamma_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{L\bar{M}}{2} \sum_{k=0}^K \eta_k^2.$$

(36)

Finally,

$$\sum_{k=0}^K \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2 \leq \frac{2}{\sigma_3} \delta_0 + \frac{\sigma_4^6 \sigma_1^2}{2\sigma_3^2} \sum_{k=0}^K \eta_k \gamma_k^3 + \frac{2L^2}{n} \sum_{k=0}^K \eta_k \gamma_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{L\bar{M}}{\sigma_3} \sum_{k=0}^K \eta_k^2.$$

(37)

Since as $K \to \infty$, by Assumption 3.1, $\sum_k \eta_k \gamma_k^3$ and $\sum_k \eta_k^2$ converge, and by Lemma B.1 $\sum_k \eta_k \gamma_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$ converges, then $\sum_k \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2$ must converge as well.

As $\sum_k \eta_k \gamma_k$ diverges, this implies that $\lim_{k \to \infty} \inf \|\nabla \mathcal{F}(\bar{x}_k)\| = 0$.

Next, let $\epsilon > 0$ be an arbitrary value, and consider the event

$$A_\epsilon := \left\{ \lim_{k \to \infty} \sup \|\nabla \mathcal{F}(\bar{x}_k)\| \geq \epsilon \right\}.$$

When event A_ϵ is true, an arbitrary subsequence $\left(\|\nabla\mathcal{F}(\bar{x}_{k_l})\|\right)_{l\in\mathbb{N}}$ of $\|\nabla\mathcal{F}(\bar{x}_k)\|$ can always be found, such that $\|\nabla\mathcal{F}(\bar{x}_{k_l})\| \geq \epsilon - \nu, \forall l$, for $\epsilon - \nu > 0$ and $\nu > 0$.

Then,

$$
\begin{aligned}
\|\nabla\mathcal{F}(\bar{x}_{k_l+1})\| &\geq \|\nabla\mathcal{F}(\bar{x}_{k_l})\| - \|\nabla\mathcal{F}(\bar{x}_{k_l+1}) - \nabla\mathcal{F}(\bar{x}_{k_l})\| \\
&\geq \epsilon - \nu - L\|\bar{x}_{k_l+1} - \bar{x}_{k_l}\| \\
&= \epsilon - \nu - L\eta_{k_l}\|\bar{g}_{k_l}\| \\
&\geq \epsilon - \nu - L\bar{M}\eta_{k_l},
\end{aligned}
\tag{38}
$$

Since $k_l \to \infty$ as $l \to \infty$, we can find a subsequence of $(k_{l_p})_{p\in\mathbb{N}}$ such that $k_{l_{p+1}} - k_{l_p} > 1$. Thus,

$$
\begin{aligned}
\sum_{k=0}^{\infty} \eta_{k+1}\gamma_{k+1}\|\nabla\mathcal{F}(\bar{x}_{k+1})\|^2 &\geq (\epsilon-\nu)^2 \sum_{k=0}^{\infty} \eta_{k+1}\gamma_{k+1} - 2(\epsilon-\nu)L\bar{M} \sum_{k=0}^{\infty} \eta_{k+1}\gamma_{k+1}\eta_k + L^2\bar{M}^2 \sum_{k=0}^{\infty} \eta_{k+1}\gamma_{k+1}\eta_k^2 \\
&\geq (\epsilon-\nu)^2 \sum_{k=0}^{\infty} \eta_{k+1}\gamma_{k+1} - 2(\epsilon-\nu)L\bar{M} \sum_{k=0}^{\infty} \eta_k^2\gamma_k + L^2\bar{M}^2 \sum_{k=0}^{\infty} \eta_{k+1}\gamma_{k+1}\eta_k^2
\end{aligned}
\tag{39}
$$

$$
= +\infty,
$$

as the first series diverges and the second and the third converge by Assumption 3.1. This implies that on A_ϵ the series $\sum_k \eta_k\gamma_k\|\nabla\mathcal{F}(\bar{x}_k)\|^2$ diverges. This is a contradiction as this series converges almost surely by (37). Therefore, $\mathbb{P}(A_\epsilon) = 0$ and as a result,

$$
\mathbb{P}\left(\lim_{k\to\infty} \|\nabla\mathcal{F}(\bar{x}_k)\| > 0\right) = \mathbb{P}\left(\bigcup_{\epsilon>0} A_\epsilon\right) = 0.
\tag{40}
$$

We conclude that $\nabla\mathcal{F}(\bar{x}_k)$ converges almost surely.

C. Rate of Convergence

Let the step sizes satisfy (12). We know that, $\forall K > 0$,

$$
\begin{aligned}
\sum_{k=0}^{K} \eta_k\gamma_k^3 = \eta_0\gamma_0^3 + \sum_{k=1}^{K} \eta_k\gamma_k^3 &\leq \eta_0\gamma_0^3\left(1 + \int_0^K (x+1)^{-\upsilon_1-3\upsilon_2}dx\right) \\
&= \eta_0\gamma_0^3\left(1 + \frac{1}{\upsilon_1+3\upsilon_2-1} - \frac{(K+1)^{-\upsilon_1-3\upsilon_2+1}}{\upsilon_1+3\upsilon_2-1}\right) \\
&\leq \eta_0\gamma_0^3\left(1 + \frac{1}{\upsilon_1+3\upsilon_2-1}\right) \\
&= \eta_0\gamma_0^3\left(\frac{\upsilon_1+3\upsilon_2}{\upsilon_1+3\upsilon_2-1}\right).
\end{aligned}
\tag{41}
$$

Similarly,

$$
\sum_{k=0}^{K} \eta_k^2 \leq \eta_0^2\left(\frac{2\upsilon_1}{2\upsilon_1-1}\right) \quad \text{and} \quad \sum_{k=0}^{K} \eta_k^3 \leq \eta_0^3\left(\frac{3\upsilon_1}{3\upsilon_1-1}\right).
\tag{42}
$$

Next, when $0 < \upsilon_1 + \upsilon_2 < 1$,

$$
\begin{aligned}
\sum_{k=0}^{K} \eta_k\gamma_k &\geq \eta_0\gamma_0 \int_0^{K+1} (x+1)^{-\upsilon_1-\upsilon_2}dx \\
&= \frac{\eta_0\gamma_0}{(1-\upsilon_1-\upsilon_2)}\left((K+2)^{1-\upsilon_1-\upsilon_2} - 1\right).
\end{aligned}
\tag{43}
$$

Then, substituting back in (37),

$$\frac{\sum_{k=0}^{K} \eta_k \gamma_k \|\nabla \mathcal{F}(\bar{x}_k)\|^2}{\sum_{k=0}^{K} \eta_k \gamma_k} \leq \frac{(1 - \upsilon_1 - \upsilon_2)}{(K+2)^{1-\upsilon_1-\upsilon_2} - 1} \left(\frac{2}{\sigma_3 \eta_0 \gamma_0} \delta_0 + \frac{\sigma_4^6 \sigma_1^2}{2\sigma_3^2} \gamma_0^2 \frac{\upsilon_1 + 3\upsilon_2}{\upsilon_1 + 3\upsilon_2 - 1} + \frac{L\bar{M}\eta_0}{\sigma_3 \gamma_0} \frac{2\upsilon_1}{2\upsilon_1 - 1} \right.$$
$$\left. + \frac{4L^2}{n} \frac{1}{(1 - \rho_w^2)} \|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \frac{4L^2 G^2 \eta_0^2}{n} \frac{\rho_w^2 (1 + \rho_w^2)}{(1 - \rho_w^2)^2} \frac{3\upsilon_1}{3\upsilon_1 - 1} \right). \tag{44}$$

Refer to (55) and (56) for the bounding of the term $\sum_k \gamma_k \eta_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$.

C.1. Proof of Lemma B.1

We substitute the variables by their updates in (10).

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 &= \|W\mathbf{x}_k - \eta_k W \mathbf{y}_k - \mathbf{1}\bar{x}_k + \eta_k \mathbf{1}\bar{y}_k\|^2 \\
&= \|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 - 2\eta_k \langle W\mathbf{x}_k - \mathbf{1}\bar{x}_k, W\mathbf{y}_k - \mathbf{1}\bar{y}_k \rangle + \eta_k^2 \|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
&\overset{(a)}{\leq} \|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \eta_k \Big[\frac{1 - \rho_w^2}{2\rho_w^2 \eta_k} \|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{2\rho_w^2 \eta_k}{1 - \rho_w^2} \|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\Big] \\
&\quad + \eta_k^2 \|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
&\overset{(b)}{\leq} \rho_w^2 \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \rho_w^2 \eta_k \Big[\frac{1 - \rho_w^2}{2\rho_w^2 \eta_k} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{2\rho_w^2 \eta_k}{1 - \rho_w^2} \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\Big] \\
&\quad + \rho_w^2 \eta_k^2 \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
&= \frac{1 + \rho_w^2}{2} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \eta_k^2 \frac{(1 + \rho_w^2)\rho_w^2}{1 - \rho_w^2} \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2,
\end{aligned} \tag{45}$$

where (a) is by $-2\epsilon \times \frac{1}{\epsilon} \langle a, b \rangle = -2\langle \epsilon a, \frac{1}{\epsilon} b \rangle \leq \epsilon^2 \|a\|^2 + \frac{1}{\epsilon^2} \|b\|^2$ and (b) is by Lemma 3.6. By taking the telescoping sum, we obtain

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \leq \left(\frac{1 + \rho_w^2}{2}\right)^{k+1} \|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \frac{2\rho_w^2}{1 - \rho_w^2} \sum_{j=0}^{k} \left(\frac{1 + \rho_w^2}{2}\right)^{j+1} \eta_{k-j}^2 \|\mathbf{y}_{k-j} - \mathbf{1}\bar{y}_{k-j}\|^2. \tag{46}$$

By the algorithm's updates of $\mathbf{y}_{k+1}$, we can write it in terms of current and all previous gradient estimates as

$$\begin{aligned}
\mathbf{y}_{k+1} &= W\mathbf{y}_k + \mathbf{g}_{k+1} - \mathbf{g}_k \\
&= W(W\mathbf{y}_{k-1} + \mathbf{g}_k - \mathbf{g}_{k-1}) + \mathbf{g}_{k+1} - \mathbf{g}_k \\
&= W^2 \mathbf{y}_{k-1} - W\mathbf{g}_{k-1} + (W - I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= W^2(W\mathbf{y}_{k-2} + \mathbf{g}_{k-1} - \mathbf{g}_{k-2}) - W\mathbf{g}_{k-1} + (W - I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= W^3 \mathbf{y}_{k-2} - W^2 \mathbf{g}_{k-2} + W(W - I)\mathbf{g}_{k-1} + (W - I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= \dots \\
&= W^{k+1} \mathbf{y}_0 - W^k \mathbf{g}_0 + \sum_{j=0}^{k-1} W^j (W - I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1} \\
&= W^k(W - I)\mathbf{g}_0 + \sum_{j=0}^{k-1} W^j (W - I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1} \\
&= \sum_{j=0}^{k} W^j (W - I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1},
\end{aligned} \tag{47}$$

then substituting in $\mathbf{y}_k - \mathbf{1}\bar{y}_k$, we obtain

$$
\begin{aligned}
\mathbf{y}_k - \mathbf{1}\bar{y}_k &= \sum_{j=0}^{k-1} W^j (W - I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \sum_{j=0}^{k-1} \frac{1}{n}\mathbf{1}\mathbf{1}^T W^j (W - I)\mathbf{g}_{k-1-j} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \mathbf{g}_k \\
&= \sum_{j=0}^{k-1} W^j (W - I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \sum_{j=0}^{k-1} \frac{1}{n}\mathbf{1}\mathbf{1}^T (W - I)\mathbf{g}_{k-1-j} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \mathbf{g}_k \\
&= \sum_{j=0}^{k-1} (W^j - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(W - I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \frac{1}{n}\mathbf{1}\mathbf{1}^T \mathbf{g}_k \\
&= \sum_{j=0}^{k-1} (W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^j (W - I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \mathbf{1}\bar{g}_k,
\end{aligned}
\tag{48}
$$

where the last equality is due to the matrix W being doubly stochastic (Assumption 1.4) and can be proved by recursion:
$(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^{j+1} = (W^j - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = W^{j+1} - \frac{1}{n}W^j\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T W + \frac{1}{n}\mathbf{1}\mathbf{1}^T = W^{j+1} - \frac{2}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n}\mathbf{1}\mathbf{1}^T = W^{j+1} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

By Lemma 3.6,

$$
\begin{aligned}
\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\| &\leq \sum_{j=0}^{k-1} \|(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^j (W - I)\mathbf{g}_{k-1-j}\| + \|\mathbf{g}_k - \mathbf{1}\bar{g}_k\| \\
&\leq \sum_{j=0}^{k-1} \rho_w^j \|(W - I)\mathbf{g}_{k-1-j}\| + \|\mathbf{g}_k - \mathbf{1}\bar{g}_k\|.
\end{aligned}
\tag{49}
$$

From Lemma 3.4, we have $\|\mathbf{g}_k\|^2 < \infty$ almost surely.

$$
\begin{aligned}
\|\mathbf{g}_k - \mathbf{1}\bar{g}_k\|^2 &= \sum_{i=1}^{n} \|g_{i,k} - \frac{1}{n}\sum_{j=1}^{n} g_{j,k}\|^2 \\
&= \sum_{i=1}^{n} \left(\|g_{i,k}\|^2 - 2\langle g_{i,k}, \frac{1}{n}\sum_{j=1}^{n} g_{j,k}\rangle + \|\bar{g}_k\|^2 \right) \\
&= \|\mathbf{g}_k\|^2 - 2n\|\bar{g}_k\|^2 + n\|\bar{g}_k\|^2 \\
&= \|\mathbf{g}_k\|^2 - n\|\bar{g}_k\|^2 \\
&\leq \|\mathbf{g}_k\|^2 \\
&\leq M'^2 < \infty
\end{aligned}
\tag{50}
$$

Thus, we get

$$
\begin{aligned}
\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\| &\leq \frac{M'}{1 - \rho_w}\|(W - I)\| + M' \\
&= G < \infty,
\end{aligned}
\tag{51}
$$

where the first term is the sum of a geometric series as $\rho_w < 1$.

1. **Proof of** $\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 = 0$

From (45), we can write

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \leq \frac{1+\rho_w^2}{2}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \eta_k^2 \frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2$$

$$\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 \leq \frac{1+\rho_w^2}{2}\|\mathbf{x}_{k-1} - \mathbf{1}\bar{x}_{k-1}\|^2 + \eta_{k-1}^2 \frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\|\mathbf{y}_{k-1} - \mathbf{1}\bar{y}_{k-1}\|^2 \tag{52}$$

$$\cdots$$

$$\|\mathbf{x}_1 - \mathbf{1}\bar{x}_1\|^2 \leq \frac{1+\rho_w^2}{2}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \eta_0^2 \frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\|\mathbf{y}_0 - \mathbf{1}\bar{y}_0\|^2.$$

If we add all inequalities in (52), we get

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \leq -\frac{1-\rho_w^2}{2}\sum_{i=1}^{k}\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2 + \frac{1+\rho_w^2}{2}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\sum_{i=0}^{k}\alpha_i^2\|\mathbf{y}_i - \mathbf{1}\bar{y}_i\|^2$$

$$\overset{(a)}{\leq} -\frac{1-\rho_w^2}{2}\sum_{i=1}^{k}\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2 + \frac{1+\rho_w^2}{2}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + G^2\frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\sum_{i=0}^{k}\alpha_i^2 \tag{53}$$

where (a) is due to (51). Let $k \to \infty$, then the second and third terms are bounded by Assumption 3.1. We then consider the only two possibilities: $\sum_i \|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2$ either diverges or converges.

Assume the hypothesis H) "$\sum_i \|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2$ diverges." to be true. This implies

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 < -\infty, \tag{54}$$

as $-\frac{1-\rho_w^2}{2} < 0$. However, $\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2$ is positive. Thus, hypothesis H cannot be true and $\sum_i \|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2$ converges. We conclude that $\lim_{k\to\infty}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 = 0$ almost surely.

2. **Proof of $\sum_{k=0}^{\infty}\gamma_k\eta_k\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 < \infty$**

Substituting (46) into the sum $\sum_{k=0}^{\infty}\gamma_k\eta_k\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$, we get

$$\sum_{k=0}^{\infty}\gamma_k\eta_k\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$$

$$=\gamma_0\eta_0\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \sum_{k=1}^{\infty}\gamma_k\eta_k\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$$

$$\leq\gamma_0\eta_0\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \sum_{k=1}^{\infty}\gamma_k\eta_k\left[\left(\frac{1+\rho_w^2}{2}\right)^k\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + G^2\frac{2\rho_w^2}{1-\rho_w^2}\sum_{j=0}^{k-1}\left(\frac{1+\rho_w^2}{2}\right)^{j+1}\eta_{k-1-j}^2\right] \tag{55}$$

$$\leq\gamma_0\eta_0\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \gamma_0\eta_0\frac{1+\rho_w^2}{1-\rho_w^2}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + G^2\frac{2\rho_w^2}{1-\rho_w^2}\sum_{k=1}^{\infty}\gamma_k\eta_k\sum_{j=0}^{k-1}\left(\frac{1+\rho_w^2}{2}\right)^{j+1}\eta_{k-1-j}^2$$

$$=\frac{2\gamma_0\eta_0}{1-\rho_w^2}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + G^2\frac{2\rho_w^2}{1-\rho_w^2}\sum_{k=1}^{\infty}\gamma_k\eta_k\sum_{j=0}^{k-1}\left(\frac{1+\rho_w^2}{2}\right)^{j+1}\eta_{k-1-j}^2,$$

where the last inequality is due to the fact that γ_k and η_k are both decreasing step-sizes and we have a sum of a

geometric series of ratio $\frac{1+\rho^2}{2} < 1$. We then study the last term,

$$
\begin{aligned}
\sum_{k=1}^{\infty} \gamma_k \eta_k \sum_{j=0}^{k-1} \left(\frac{1+\rho_w^2}{2}\right)^{j+1} \eta_{k-1-j}^2 &\leq \sum_{k=1}^{\infty} \gamma_k \sum_{j=0}^{k-1} \left(\frac{1+\rho_w^2}{2}\right)^{j+1} \eta_{k-1-j}^3 \\
&= \sum_{k=1}^{\infty} \gamma_k \sum_{j=1}^{k} \left(\frac{1+\rho_w^2}{2}\right)^{k-j+1} \eta_{j-1}^3 \\
&= \sum_{j=1}^{\infty} \eta_{j-1}^3 \sum_{k=j}^{\infty} \gamma_k \left(\frac{1+\rho_w^2}{2}\right)^{k-j+1} \\
&\leq \gamma_0 \sum_{j=1}^{\infty} \eta_{j-1}^3 \sum_{k=j}^{\infty} \left(\frac{1+\rho_w^2}{2}\right)^{k-j+1} \\
&= \gamma_0 \frac{1+\rho_w^2}{1-\rho_w^2} \sum_{j=1}^{\infty} \eta_{j-1}^3 \\
&< \infty,
\end{aligned}
\tag{56}
$$

as $\sum \eta_k^2$ converges by Assumption 3.1.

Finally, $\sum_{k=0}^{\infty} \gamma_k \eta_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 < \infty$.