DECENTRALIZED NONSMOOTH NONCONVEX OPTIMIZATION WITH CLIENT-DATA SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies decentralized nonsmooth nonconvex optimization with Lipschitz-continuous local functions. We propose an efficient stochastic first-order method with a novel client-data sampling strategy in which the mini-batch sizes of the stochastic gradients on all clients follow a multinomial distribution. We show that our method can find a (δ, ϵ)-Goldstein stationary point on each client with the best-known dependence on ϵ and δ in both computational and communication complexity. Moreover, our results tighten the upper complexity bounds with respect to the Lipschitz parameter, the variance upper bound, the number of clients, and the spectral gap. We further conduct numerical experiments to demonstrate the empirical advantages of the proposed method.

1 INTRODUCTION

The large scale nonsmooth nonconvex optimization covers many applications in fields such as machine learning (Nair & Hinton, 2010; Xiao et al., 2024), statistics (Fan & Li, 2001; Zhang, 2010a;b), and economics (Duffie, 2010; Stadtler, 2014). In this paper, we focus on the decentralized stochastic optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \tag{1}$$

over the network with n clients, where the local function at the ith client has the form of

$$f_i(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\xi}_i\sim\mathcal{D}_i}[F_i(\mathbf{x};\boldsymbol{\xi}_i)] \tag{2}$$

such that the $f_i : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous but possibly nonconvex nonsmooth and the random index $\boldsymbol{\xi}_i \in \Xi_i$ follows the distribution $\mathcal{D}_i$. It is well known that achieving approximate stationary points in terms of the classical Clarke subdifferential (Clarke, 1990) for the general Lipschitz continuous function is intractable (Jordan et al., 2022; Kornowski & Shamir, 2021; Tian & So, 2024; Zhang et al., 2020). Instead, we typically target to find (δ, ϵ)-Goldstein stationary points (Zhang et al., 2020). This criterion suggests studying the convex hull of Clarke subdifferential at points in the δ-radius neighborhood of the given point.

The stochastic optimization methods for finding (δ, ϵ)-Goldstein stationary points in non-distributed setting have been widely studied in recent years (Chen et al., 2023; Cutkosky et al., 2023; Davis et al., 2022; Kornowski & Shamir, 2024; Lin et al., 2022; Tian et al., 2022; Zhang et al., 2020). Specifically, Tian et al. (2022); Zhang et al. (2020) proposed the (perturbed) stochastic interpolated normalized gradient descent with the stochastic first-order oracle (SFO) complexity of $\mathcal{O}(\delta^{-1}\epsilon^{-4})$. In a seminal work, Cutkosky et al. (2023) established the conversion from nonsmooth nonconvex optimization to online learning, achieving the SFO complexity of $\mathcal{O}(\delta^{-1}\epsilon^{-3})$. They also extends the lower bound of Arjevani et al. (2023) to show their SFO complexity is optimal.

In decentralized setting, we have to consider the consensus error among the variables on different clients in the network. The popular technique of gradient tracking can successfully bound the consensus error in for smooth optimization problems (Nedic & Ozdaglar, 2009; Qu & Li, 2017; Shi et al., 2015), while it cannot be directly used in the nonsmooth setting since the Lipschitz continuity of the gradient (subgradient) may not hold. Kovalev et al. (2024); Lan et al. (2020) proposed efficient decentralized algorithms based on the primal-dual framework for the nonsmooth objective but only

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: We present the upper complexity bounds of our methods and related work for finding (δ, ϵ)-Goldstein stationary points in stochastic decentralized nonsmooth nonconvex optimization problem. Here, the sample complexity refers to the total local stochastic first-order oracle calls on all of the n clients.

Methods	†ME-DOL (Sahinoglu & Shahrampour, 2024)	DOC^{2}S (Theorem 1)
Sample Complexity	$\mathcal{O}\Big(\frac{L+\Delta}{\gamma^2\delta\epsilon^3}\big(G(L+G)n^2+(G^2+L^2d)\gamma n\sqrt{n}\big)\Big)$	$\mathcal{O}\Big(\frac{(bL^2+\sigma^2)\Delta}{\delta\epsilon^3}\Big)$
Computation Rounds	$\mathcal{O}\Big(\frac{L+\Delta}{\gamma^2\delta\epsilon^3}\big(G(L+G)n+(G^2+L^2d)\gamma\sqrt{n}\big)\Big)$	$\tilde{\mathcal{O}}\Big(\frac{(bL^2+\sigma^2)\Delta}{n\delta\epsilon^3}\Big)$
Communication Rounds	$\mathcal{O}\Big(\frac{L+\Delta}{\gamma^2\delta\epsilon^3}\big(G(L+G)n+(G^2+L^2d)\gamma\sqrt{n}\big)\Big)$	$\tilde{\mathcal{O}}\Big(\frac{(L^2+b^{-1}\sigma^2)\Delta}{\sqrt{\gamma}\delta\epsilon^3}\Big)$

† ME-DOL (Sahinoglu & Shahrampour, 2024, Assumption 4) additionally assumes the stochastic gradient has bounded second moment, i.e., there exists constant $G > 0$ such that $\mathbb{E}[\|\nabla F_i(\mathbf{x}, \boldsymbol{\xi}_i)\|^2] \leq G^2$ for all $\mathbf{x} \in \mathbb{R}^d$.

limit to the convex problem. A natural idea for decentralized nonsmooth optimization is using the randomized smoothing to establish the smooth surrogate for the original problem, which works for both the convex (Scaman et al., 2018) and the nonconvex settings (Lin et al., 2024; Sahinoglu & Shahrampour, 2024). Notably, Sahinoglu & Shahrampour (2024) proposed multi-epoch decentralized online learning (ME-DOL) method for decentralized stochastic stochastic nonsmooth nonconvex optimization, which incorporates the decentralized online mirror descent (Shahrampour & Jadbabaie, 2017) into the online-to-nonconvex conversion (Cutkosky et al., 2023; Kornowski & Shamir, 2024). The theoretical analysis shows that ME-DOL can find (δ, ϵ)-Goldstein stationary points under the assumption of bounded second moment for the stochastic gradient (Sahinoglu & Shahrampour, 2024).

It is worth noting that the objective in distributed optimization problem (1) naturally has the finite-sum structure in the view of local functions $\{f_i\}_{i=1}^n$. This motivates us to design the partial participated methods, which performs the client sampling during iterations and only executes the computation on the selected clients (Chen et al., 2020; Maranjyan et al., 2022; Mishchenko et al., 2022). Some recent works (Bai et al., 2024; Liu et al., 2024; Luo et al., 2022; Chen et al., 2025) studied partial participated methods by considering the balance among the first-order oracle complexity, the computation rounds, and the communication rounds in decentralized optimization. However, these results heavily depend on the smoothness assumptions. To the best of our knowledge, all existing methods for decentralized nonsmooth optimization (Chen et al., 2020; Kovalev et al., 2024; Lan et al., 2020; Lin et al., 2024; Sahinoglu & Shahrampour, 2024; Wang et al., 2023; Zhang et al., 2024) require all clients accessing their local oracle in per computation rounds, which limits the sampling efficiency.

In this paper, we propose decentralized online-to-nonconvex conversion with client-data sampling (DOC^{2}S), incorporating a novel client-data sampling scheme and Chebyshev acceleration (Ye et al., 2023; Arioli & Scott, 2014). The proposed client-data sampling scheme repeats the sampling procedure b times, where each of them first samples a client and then draws a stochastic gradient from the selected client. This two-stage sampling strategy enables DOC^{2}S to perform mini-batch update with unfixed batch sizes on different clients, resulting the sharper dependence on the Lipschitz parameter and variance upper bound. On the other hand, the Chebyshev acceleration allows DOC^{2}S significantly reduce the consensus error in each round and therefore improve the dependence on the number of clients n and the dimensionality d in the upper complexity bounds of ME-DOL (Sahinoglu & Shahrampour, 2024). We show that DOC^{2}S can achieve an (δ, ϵ)-Goldstein stationary point on each client with the total sample complexity of $\mathcal{O}((bL^2+\sigma^2)\delta^{-1}\epsilon^{-3}\Delta)$, the computation rounds of $\mathcal{O}((L^2+b^{-1}\sigma^2)\delta^{-1}\epsilon^{-3}\Delta)$, and the communication rounds of $\mathcal{O}((bL^2+\sigma^2)\gamma^{-1/2}\delta^{-1}\epsilon^{-3}\Delta)$, where b is the overall batch size in each iteration, L is the Lipschitz constant of local functions, σ^2 is the upper bound of the stochastic gradient variance, γ is the spectral gap of the mixing matrix, and Δ is the initial function value gap. These upper complexity bounds indicate the mini-batch size b in our algorithm provides a tradeoff between the computation complexity and the communication complexity. Specifically, even by taking $b = 1$, all of our upper bounds are tighter than those of ME-DOL (Sahinoglu & Shahrampour, 2024). We present the comparison of theoretical results of our methods and related work in Table 1.

2 PRELIMINARIES

In this section, we formalize our problem setting and introduce the background of nonsmooth analysis.

2.1 NOTATION AND ASSUMPTIONS

We use the bold lowercase letter with the subscript to present the local variable, e.g., the row vector $\mathbf{x}_i \in \mathbb{R}^{1\times d}$ presents the local variable on the ith client. The bold lowercase letter with a bar presents the average of local variables, e.g., $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$. We also use bold uppercase letters to denote matrices that consists of local variables, e.g., $\mathbf{X} = [\mathbf{x}_1; \ldots; \mathbf{x}_n] \in \mathbb{R}^{n\times d}$. For the ease of presentation, we allow the input of the function and its gradient to be organized as either a column vector or a row vector. We denote $\|\cdot\|$ and $\|\cdot\|_2$ as the Frobenius norm and the spectral norm of the matrix, respectively, also the Euclidean norm of the vector. We let $\mathbf{1}_n = [1, \ldots, 1]^\top \in \mathbb{R}^n$ and $\mathbf{I}$ be the identity matrix. The notation $\mathrm{conv}(\cdot)$ denotes the convex hull of given set. Additionally, we use notations $\mathbb{B}^d(\mathbf{x}, \delta)$ and $\mathbb{S}^{d-1}$ to present the Euclidean ball centered at $\mathbf{x} \in \mathbb{R}^d$ with radius $\delta > 0$ and the unit sphere centered at the origin, respectively.

We impose following assumptions for formulations (1) and (2).

Assumption 1. We suppose each local function $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz continuous, i.e., there exists $L > 0$ such that $|f_i(\mathbf{x}) - f_i(\mathbf{y})| \le L||\mathbf{x} - \mathbf{y}||$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $i \in [n]$.

Assumption 2. We suppose the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is lower bounded by f^*, i.e., it holds $f^* \triangleq \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$.

Note that Rademacher's theorem (Evans, 2018) states a Lipschitz continuous function is differentiable almost everywhere. Thus, the LSFO is well-defined almost everywhere under Assumption 1.

We make the following assumption for the local stochastic first-order oracle (LSFO).

Assumption 3. We suppose the algorithms can access the local function $f_i : \mathbb{R}^d \to \mathbb{R}$ via the LSFO consisting of the local gradient estimator $F_i : \mathbb{R}^d \times \Xi_i \to \mathbb{R}$ and the random index $\boldsymbol{\xi}_i \sim \mathcal{D}_i$ such that $\mathbb{E}_{\boldsymbol{\xi}_i\sim\mathcal{D}_i}[\nabla F_i(\mathbf{x}; \boldsymbol{\xi}_i)] = \nabla f_i(\mathbf{x})$ and $\mathbb{E}_{\boldsymbol{\xi}_i\sim\mathcal{D}_i}[\|\nabla F_i(\mathbf{x}; \boldsymbol{\xi}_i) - \nabla f_i(\mathbf{x})\|^2] \le \sigma^2$ for some $\sigma \ge 0$.

We aim for all n clients in the network to collaborate in solving stochastic decentralized optimization problem (1). We use the doubly stochastic matrix $\mathbf{P} = [p_{ij}] \in \mathbb{R}^{n\times n}$ to describe the topology of the network. Specifically, the communication step at the ith client is built upon the weighted average $\mathbf{x}_i^+ = \sum_{j=1}^n p_{ij}\mathbf{x}_j$, where $\mathbf{x}_j$ is the local variable on the jth client. We impose the following standard assumption in decentralized optimization for the matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ (Schmidt et al., 2017; Scaman et al., 2018).

Assumption 4. We suppose that the mixing matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ associated with the network satisfies: (a) The matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ is symmetric and holds $p_{ij} \ge 0$ for all $i, j \in [n]$; (b) It holds $p_{ij} \ne 0$ if and only if clients i and j are connected or $i = j$; (c) It holds $\mathbf{0} \preceq \mathbf{P} \preceq \mathbf{I}$ and $\mathbf{P}^\top\mathbf{1}_n = \mathbf{P}\mathbf{1}_n = \mathbf{1}_n$.

Under Assumption 4, the largest eigenvalue of the mixing matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ is one. Consequently, we define the spectral gap of $\mathbf{P} \in \mathbb{R}^{n\times n}$ as $\gamma = 1 - \lambda_2(\mathbf{P}) \in (0, 1]$, where $\lambda_2(\mathbf{P})$ is the second largest eigenvalue of $\mathbf{P}$.

2.2 GOLDSTEIN STATIONARY POINTS

We present the notion of Clarke subdifferential (Clarke et al., 2008) and its relaxation Goldstein subdifferential (Goldstein, 1977) for the Lipschitz continuous objective in the nonconvex nonsmooth problem as follows.

Definition 1 (Clarke et al. (2008))**.** The Clarke subdifferential of a Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $\mathbf{x} \in \mathbb{R}^d$ is defined by $\partial f(\mathbf{x}) := \mathrm{conv}\big\{\mathbf{g} : \mathbf{g} = \lim_{\mathbf{x}_k\to\mathbf{x}} \nabla f(\mathbf{x}_k)\big\}$.

Definition 2 (Goldstein (1977))**.** For given $\delta \ge 0$ and a Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, the Goldstein δ-subdifferential of at point $\mathbf{x} \in \mathbb{R}^d$ is defined by $\partial_\delta f(\mathbf{x}) := \mathrm{conv}\big(\cup_{\mathbf{y}\in\mathbb{B}^d(\mathbf{x},\delta)} \partial f(\mathbf{y})\big)$, where the $\partial f(\mathbf{y})$ is Clarke subdifferential.

We are interested in finding the (δ, ϵ)-Goldstein stationary point (Zhang et al., 2020), which is defined as follows.

Definition 3 (Zhang et al. (2020))**.** For given Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, $\delta \geq 0$, and $\mathbf{x} \in \mathbb{R}^d$, we denote $\|\nabla f(\mathbf{x})\|_\delta := \min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x})\}$. We call the point $\mathbf{x} \in \mathbb{R}^d$ as a (δ, ϵ)-Goldstein stationary point of f if $\|\nabla f(\mathbf{x})\|_\delta \leq \epsilon$ holds.

2.3 Randomized Smoothing

Randomized smoothing is a popular technique in stochastic optimization (Duchi et al., 2012; Lin et al., 2022; Nesterov & Spokoiny, 2017; Shamir, 2017; Yousefian et al., 2012). This paper focuses on the uniform smoothing as follows.

Definition 4 (Yousefian et al. (2012))**.** Given a Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its smooth surrogate as $f_\delta(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{u}\sim \mathrm{Unif}(\mathbb{B}^d(\mathbf{0},1))}[f(\mathbf{x} + \delta\mathbf{u})]$, where $\mathrm{Unif}(\mathbb{B}^d(\mathbf{0}, 1))$ is the uniform distribution on the unit Euclidean ball centered at the origin.

The smooth surrogate f_δ has the following properties.

Proposition 1 (Lin et al. (2022, Proposition 2.2), Kornowski & Shamir (2024, Lemma 4))**.** *Suppose the function $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz, then its smooth surrogate f_δ holds: (a) $|f_\delta(\cdot) - f(\cdot)| \leq \delta L$; (b) $f_\delta(\cdot)$ is L-Lipschitz; (c) $f_\delta(\cdot)$ is differentiable with $c_0\sqrt{d}L\delta^{-1}$-Lipschitz gradients for some numeric constant $c_0 > 0$; (d) $\partial_\mu f_\delta(x) \subseteq \partial_{\mu+\delta} f(x)$ for all $\mu \geq 0$.*

3 The Algorithm Design

We propose decentralized online-to-nonconvex conversion with client-data sampling (DOC^{2}S) in Algorithm 1, which incorporates the client-data sampling strategy and the multi-consensus step with Chebyshev acceleration (Algorithm 2) (Ye et al., 2023; Arioli & Scott, 2014) into the framework of online-to-nonconvex conversion (Cutkosky et al., 2023; Sahinoglu & Shahrampour, 2024).

Our algorithm design is based on iterating with mini-batch stochastic gradient, where the total sampling budget per iteration is b across all n clients. Specifically, each iteration allocates the local mini-batch sizes by multinomial distribution, i.e.,

$$[b_{t,1}, \ldots, b_{t,n}]^\top \sim \mathrm{Multinomial}\left(b, \frac{\mathbf{1}}{n}\right), \tag{3}$$

where $b_{t,i}$ is the batch size of the ith client at the tth iteration. We then establish the mini-batch stochastic gradient

$$\mathbf{g}_{t,i} = \sum_{j=1}^{b_{t,i}} \nabla F_i(\mathbf{w}_{t,i} + \mu\mathbf{z}_{t,i}; \boldsymbol{\xi}_{t,i,j}) \tag{4}$$

for each client $i \in [n]$, where $\mathbf{z}_{t,i} \sim \mathrm{Unif}(\mathbb{B}^d(\mathbf{0}, 1))$, $\boldsymbol{\xi}_{t,i,j} \sim \mathcal{D}_i$, and $\mu \geq 0$. We then combine the above local gradient estimator and Chebyshev acceleration (Algorithm 2) with online-to-nonconvex conversion to establish Algorithm 1, where we define

$$\mathbf{Clip}(\mathbf{x}, D) = \begin{bmatrix} \mathbf{clip}(\mathbf{x}_1, D) \\ \vdots \\ \mathbf{clip}(\mathbf{x}_n, D) \end{bmatrix} \in \mathbb{R}^{n\times d} \quad \text{with} \quad \mathbf{clip}(\mathbf{x}, D) = \min\left\{1, \frac{D}{\|\mathbf{x}\|}\right\}\mathbf{x}.$$

Remark 1. Note that the batch sizes allocation (3) gives $\sum_{i=1}^b b_{t,i} = b$ and $\mathbb{E}[b_{t,i}] = b/n$, allowing the client with $b_{t,i} = 0$ to skip local stochastic gradient evaluation, i.e., we let $\mathbf{g}_{t,i} = \mathbf{0}$ if $b_{t,i} = 0$. Since the batch sizes on different clients may be not identical, the tth iteration takes the computation rounds of $\max_{i\in[n]} b_{t,i}$, which is not always proportion to the sample complexity $b = \sum_{i=1}^n b_{t,i}$.

Remark 2. The factor n/b in line 11 of Algorithm 1 scales the mini-batch stochastic gradients by total sampling budget, which ensures the average of the rows of matrix $n\mathbf{G}_t/b$ to be

$$\hat{\mathbf{g}}_t = \frac{1}{b}\sum_{i=1}^n \sum_{j=1}^{b_{t,i}} \nabla F_i(\mathbf{w}_{t,i} + \mu\mathbf{z}_{t,i}; \boldsymbol{\xi}_{t,i,j}).$$

Since FastGossip (Algorithm 2) encourages all $\mathbf{w}_{t,i}$ be close to their average $\bar{\mathbf{w}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}$, the vector $\hat{\mathbf{g}}_t$ can be regarded as a stochastic gradient evaluation of the global smooth surrogate function $\frac{1}{n}\sum_{i=1}^n (f_i)_\mu(\cdot)$ at the point $\bar{\mathbf{w}}_t$ with batch size b.

Algorithm 1 Decentralized Online-to-Nonconvex Conversion with Client-Data Sampling (DOC^{2}S)

Input: $\mu \geq 0,\ K, T, R, b \in \mathbb{N},\ \eta, \mu, D > 0,\ \mathbf{P} \in \mathbb{R}^{n \times n}$
Initialization: $\mathbf{X}_0 = \mathbf{V}_0 = \mathbf{0}$
for $t = 1$ **to** KT **do**
  $\mathbf{X}_t = \mathrm{FastGossip}(\mathbf{X}_{t-1} + \mathbf{V}_t, \mathbf{P}, R)$
  $s_t \sim \mathrm{Unif}([0, 1])$
  $\mathbf{W}_t = \mathbf{X}_{t-1} + s_t \mathbf{V}_t$
  $[b_{t,1}, \ldots, b_{t,n}]^\top \sim \mathrm{Multinomial}\Big(b, \frac{\mathbf{1}}{n}\Big)$
  $\mathbf{z}_{t,i} \overset{\text{i.i.d}}{\sim} \mathrm{Unif}(\mathbb{B}^d(\mathbf{0}, 1))$ for all $i \in [n]$
  $\boldsymbol{\xi}_{t,i,j} \overset{\text{i.i.d}}{\sim} \mathcal{D}_i$ for all $i \in [n]$ and $j \in [b_{t,i}]$
  $\mathbf{G}_t = \left[\sum_{j=1}^{b_{t,1}} \nabla F_1(\mathbf{w}_{t,1} + \mu \mathbf{z}_{t,1}; \boldsymbol{\xi}_{t,1,j}); \ \ldots\ ; \sum_{j=1}^{b_{t,n}} \nabla F_n(\mathbf{w}_{t,n} + \mu \mathbf{z}_{t,n}; \boldsymbol{\xi}_{t,n,j})\right]$
  $\mathbf{V}_{t+1} = \mathbf{Clip}\left(\mathrm{FastGossip}\left(\mathbf{V}_t - \eta \cdot \frac{n\mathbf{G}_t}{b}, \mathbf{P}, R\right), D\right)$
end for
Output: $\mathbf{w}_i^{\mathrm{out}} \sim \mathrm{Unif}\big(\{\hat{\mathbf{w}}_i^{(1)}, \ldots, \hat{\mathbf{w}}_i^{(K)}\}\big)$ for all $i \in [n]$, where $\hat{\mathbf{w}}_i^{(k)} = \frac{1}{T}\sum_{t=(k-1)T+1}^{kT} \mathbf{w}_{t,i}$

Algorithm 2 $\mathrm{FastGossip}(\mathbf{Z}^{(0)}, \mathbf{P}, R)$

Initialize: $\mathbf{Z}^{(-1)} = \mathbf{Z}^{(0)}$, $\phi = \dfrac{1 - \sqrt{1 - (\lambda_2(\mathbf{P}))^2}}{1 + \sqrt{1 - (\lambda_2(\mathbf{P}))^2}}$
for $r = 0$ **to** $R - 1$
  $\mathbf{Z}^{(r+1)} = (1 + \phi)\mathbf{P}\mathbf{Z}^{(r)} - \phi \mathbf{Z}^{(r-1)}$
end for
Output: $\mathbf{Z}^{(R)}$

4 The Complexity Analysis

In this section, we provide a convergence analysis of DOC^{2}S (Algorithm 1). We first consider the case where each local loss function f_i has Lipschitz-continuous gradients and then adopt the analysis to the general nonsmooth nonconvex setting.

4.1 Analysis for the Smooth Case

We first introduce the following proposition for the subroutine of multi-consensus steps with Chebyshev acceleration (Algorithm 2) (Ye et al., 2023).

Proposition 2 (Ye et al. (2023, Proposition 1))**.** *For Algorithm 2, we denote* $\bar{\mathbf{z}} = \frac{1}{n}\mathbf{1}^\top \mathbf{Z}^{(0)}$*, then it holds that* $\frac{1}{n}\mathbf{1}^\top \mathbf{Z}^{(R)} = \bar{\mathbf{z}}$ *and*

$$\|\mathbf{Z}^{(R)} - \mathbf{1}\bar{\mathbf{z}}^\top\|_F^2 \leq 14\left(1 - \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{\gamma}\right)^{2R} \|\mathbf{Z}^{(0)} - \mathbf{1}^\top \bar{\mathbf{z}}\|_F^2.$$

Based on Proposition 2, we can bound the consensus errors of $\mathbf{V}_t$ and $\mathbf{X}_t$ by following lemmas.

Lemma 1. *Under Assumptions 3 and 4, Algorithm 1 with*

$$R \geq \left\lceil \frac{1}{(1 - 1/\sqrt{2})\sqrt{\gamma}} \log \frac{\sqrt{28nD^2 + n^2 D(G^2 + \sigma^2)}}{\epsilon'} \right\rceil \tag{5}$$

satisfies $\mathbb{E}[\|\mathbf{V}_t - \bar{\mathbf{v}}_t\|] \leq \epsilon'$ *for all* $i \in [n]$ *and* $\epsilon' > 0$.

Lemma 2. *Under the setting of Lemma 1, Algorithm 1 holds* $\|\mathbf{X}_t - \bar{\mathbf{x}}_t\| \leq \epsilon'$, *for all* $i \in [n]$ *and* $\epsilon' < D$.

We then consider the decrease of the objective function value at the mean vector in the smooth case, where each f_i is H-smooth. According to the update rule of $\mathbf{X}_t$ in Algorithm 1 (line 4) and the doubly stochastic assumption of matrix $\mathbf{P}$ (Assumption 4), we have:

$$\bar{\mathbf{x}}_t = \bar{\mathbf{x}}_{t-1} + \bar{\mathbf{v}}_t.$$

Let $\tilde{\boldsymbol{\nabla}}_t = \int_0^1 \nabla f(\bar{\mathbf{x}}_{t-1} + s\bar{\mathbf{v}}_t)\mathrm{d}s$, then the continuity of the function f means

$$f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t-1}) = \int_0^1 \langle \nabla f(\bar{\mathbf{x}}_{t-1} + s\bar{\mathbf{v}}_t), \bar{\mathbf{v}}_t\rangle \mathrm{d}s = \langle \tilde{\boldsymbol{\nabla}}_t, \bar{\mathbf{v}}_t\rangle.$$

Summing the above equation over $t = (k-1)T, (k-1)T+1, \ldots, kT$ and taking expectation yields

$$\mathbb{E}[f(\bar{\mathbf{x}}_{kT}) - f(\bar{\mathbf{x}}_{(k-1)T})] = \sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \tilde{\boldsymbol{\nabla}}_t, \bar{\mathbf{v}}_t\rangle] \tag{6}$$

$$= \underbrace{\sum_{t=(k-1)T+1}^{kT} \mathbb{E}\left[\left\langle \tilde{\boldsymbol{\nabla}}_t - \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t\right\rangle\right]}_{S_1} + \underbrace{\sum_{t=(k-1)T+1}^{kT} \mathbb{E}\left[\left\langle \frac{n}{b}\bar{\mathbf{g}}_t, \mathbf{u}_k\right\rangle\right]}_{S_2} + \underbrace{\sum_{t=(k-1)T+1}^{kT} \mathbb{E}\left[\left\langle \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t - \mathbf{u}_k\right\rangle\right]}_{S_3}. \tag{7}$$

hold for all $\mathbf{u}_k \in \mathbb{R}^d$. We can bound S_1, S_2 and S_3 by the following three lemmas.

Lemma 3. *Under Assumptions 1, 2, 3 and 4, we further suppose each* f_i *is* H*-smooth, then Algorithm 1 with* $\mu = 0$ *holds*

$$\sum_{t=(k-1)T+1}^{kT} \mathbb{E}\left[\left\langle \tilde{\boldsymbol{\nabla}}_t - \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t\right\rangle\right] \leq 2DH\epsilon' T.$$

Lemma 4. *Under Assumptions 1, 2, 3 and 4, we let* $\mathbf{u}_k = -D\frac{\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})}{\left\|\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right\|}$. *Then Algorithm 1 with* $\mu = 0$ *holds*

$$\mathbb{E}\left[\left\langle \sum_{t=(k-1)T+1}^{kT} \frac{n}{b}\bar{\mathbf{g}}_t, \mathbf{u}_k\right\rangle\right] \leq -D\mathbb{E}\left[\left\|\frac{1}{n}\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right\|\right] + \frac{2\sqrt{2}D(\sigma+2L)\sqrt{T}}{\sqrt{b}}.$$

Lemma 5. *Under the setting of Lemma 4, Algorithm 1 with* $\mu = 0$ *holds*

$$\sum_{t=(k-1)T+1}^{kT} \mathbb{E}\left[\left\langle \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t - \mathbf{u}_k\right\rangle\right] \leq D\sqrt{T(L^2+\sigma^2/b)} + \epsilon' T\sqrt{T(L^2+\sigma^2/b)}$$

We can also bound the difference between the gradients of the global and local functions as follows.

Lemma 6 (Sahinoglu & Shahrampour (2024, Lemma 2))**.** *Let the functions* $\{f_i\}_{i=1}^n$ *be* H*-smooth and* $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$. *For given sequence* $\{\mathbf{w}_{t,i}\}_{t,i=1}^{T,n}$, *we suppose there exists some* $r > 0$ *such that* $\|\mathbf{w}_{t,i} - \bar{\mathbf{w}}_t\| \leq r$ *for all* $i \in [n]$, $k \in [K]$ *and* $t \in [T]$, *then it holds*

$$\left\|\frac{1}{nT}\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f(\mathbf{w}_{t,i})\right\| \leq \left\|\frac{1}{nT}\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right\| + 2rH,$$

where $\bar{\mathbf{w}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_{t,i}$.

Combing all above results, we establish the theoretical guarantee for our method in the smooth case.

Lemma 7. *Under the settings of Lemma 3, running Algorithms 1 with* $\mu = 0$, $T = \mathcal{O}((L^2 + \sigma^2/b)\epsilon^{-2})$, $K = \mathcal{O}(\delta^{-1}\epsilon^{-1}\Delta)$, $D = \mathcal{O}(\delta/T)$, $R = \tilde{\mathcal{O}}(\gamma^{-1/2})$, *and* $\eta = D/\sqrt{T(L^2+\sigma^2/b)}$ *can output* $\{\mathbf{w}_i^{\mathrm{out}}\}_{i=1}^n$ *such that* $\mathbb{E}\left[\|\nabla f(\mathbf{w}_i^{\mathrm{out}})\|_\delta\right] \leq \epsilon$ *holds for all* $i \in [n]$.

4.2 Analysis for the Nonsmooth Case

For nonsmooth case, we analyze our algorithm by considering its smooth surrogate function. Specifically, Proposition 1 states the function f_δ has H-Lipschitz continuous gradients with $H = cL\sqrt{d}\delta^{-1}$. Additionally, we can connect the Goldstein subdifferential of a Lipschitz continuous function to that of its smooth surrogate as follows.

Lemma 8 (Sahinoglu & Shahrampour (2024, Proposition 2))**.** *From Proposition 1 and Definition 3, we have* $\|\nabla f(\mathbf{x})\|_\delta \leq \|\nabla f_{a\delta}(\mathbf{x})\|_{(1-a)\delta}$ *for all* $a \in (0,1)$.

Connecting the above lemma with Lemma 7, we can establish our main results for the nonsmooth case as follows.

Theorem 1. *Under Assumptions 1, 2, 3, and 4, Algorithm 1 with* $\mu = \delta/2$, $T = \mathcal{O}((L^2+\sigma^2/b)\epsilon^{-2})$, $K = \mathcal{O}(\delta^{-1}\epsilon^{-1}\Delta)$, $D = \mathcal{O}(\delta/T)$, $R = \tilde{\mathcal{O}}(\gamma^{-1/2})$, *and* $\eta = D/\sqrt{T(L^2+\sigma^2/b)}$ *can output* $\{\mathbf{w}_i^{\text{out}}\}_{i=1}^n$ *such that* $\mathbb{E}\left[\|\nabla f(\mathbf{w}_i^{\text{out}})\|_\delta\right] \leq \epsilon$ *for all* $i \in [n]$.

Corollary 2. *Following the setting of Theorem 1, each client can achieve an* (δ,ϵ)*-Goldstein stationary point within the overall sample complexity of* $\mathcal{O}((bL^2+\sigma^2)\Delta\delta^{-1}\epsilon^{-3})$ *, the computation rounds of* $\tilde{\mathcal{O}}((bL^2+\sigma^2)\Delta n^{-1}\delta^{-1}\epsilon^{-3})$ *, and the communication rounds of* $\tilde{\mathcal{O}}((L^2+b^{-1}\sigma^2)\Delta\gamma^{-1/2}\delta^{-1}\epsilon^{-3})$.

Remark 3. The mini-batch size b provides a trade-off between the complexity of computation and communication. Specifically, increasing b reduces the variance-dependent term in the communication complexity from σ^2 to σ^2/b, In particular, choosing $b = \Theta(\sigma^2/L^2)$ balances the terms bL^2 and σ^2 in the sample complexity and the computation rounds. Notably, even with $b = 1$, our method still achieves better complexity bounds than existing methods (Sahinoglu & Shahrampour, 2024) in terms of all three measures.

Remark 4. Besides the sharper complexity bounds than ME-DOL (Sahinoglu & Shahrampour, 2024), the proposed DOC^{2}S also guarantees every client can achieve a (δ,ϵ)-Goldstein stationary point in expectation. Recall that the theoretical analysis of ME-DOL only indicates that there exists some point $\bar{\mathbf{w}}^k = \frac{1}{nT}\sum_{t=1}^T\sum_{i=1}^n \mathbf{w}_i^{k,t}$ which is an (δ,ϵ)-Goldstein stationary point, where $\mathbf{w}_i^{k,t}$ is generated on the ith client. However, achieving such mean vector $\bar{\mathbf{w}}^k$ is non-trivial in practice for the decentralized setting. In contrast, the point $\mathbf{w}_i^{\text{out}}$ in our algorithm and theory only depends on its local variables.

5 Experiments

This section empirically compare our DOC^{2}S with baseline methods ME-DOL (Shahrampour & Jadbabaie, 2017) by conduction experiments on the following two models:

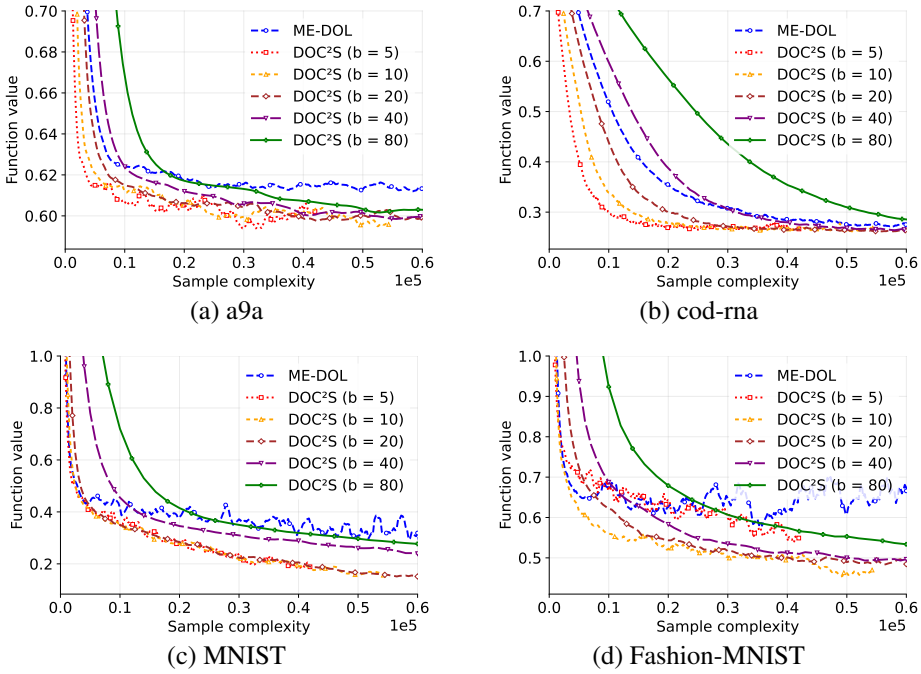
Nonconvex SVM with Capped-ℓ_1 Penalty: We first consider the model of nonconvex penalized SVM with capped-ℓ_1 regularizer (Zhang, 2010b) on dataset "a9a" and "cod-rna", which targets to train the binary classifier $\mathbf{x} \in \mathbb{R}^d$ on dataset $\{(\mathbf{a}_i, b_i)\}_{i=1}^m$, where $\mathbf{a}_i \in \mathbb{R}^d$ and $b_i \in \{-1,1\}$ are the feature vector and label for the i-th sample. We formulate this problem as the following nonsmooth nonconvex problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{m}\sum_{i=1}^m g_i(\mathbf{x}),$$

where $g_i(\mathbf{x}) = l(b_i\mathbf{a}_i^\top\mathbf{x}) + \nu(\mathbf{x})$, $l(z) = \max\{1-z, 0\}$, $\nu(\mathbf{x}) = \lambda\sum_{j=1}^d \min\{|x(j)|, \alpha\}$, and $\lambda, \alpha > 0$. Here, the notation $x(j)$ means the jth coordinate of $\mathbf{x}$. We evenly partition $\{g_i\}_{i=1}^m$ into m clients. and set $\lambda = 10^{-5}/m$ and $\alpha = 2$ in our experiments.

Multilayer Perceptron with ReLU activation: We then consider the applications of image classification on datasets "MNIST" and "fashion-MNIST" (28×28 pixels for each image, 10 classes). Specifically, we consider the two-layer Multilayer Perceptron (MLP) with ReLU activation and a 256-dimensional hidden layer. Specifically, the local function at the i-th client can be written as

$$f_i(\mathbf{x}) \triangleq \frac{1}{m_i}\sum_{j=1}^{m_i} \ell\big(g(\mathbf{x};\mathbf{a}_i^j), b_i^j\big) + \lambda\|\mathbf{x}\|_2^2,$$



(a) a9a (b) cod-rna

(c) MNIST (d) Fashion-MNIST

Figure 1: Comparison on the sample complexity of ME-DOL and DOC^{2}S with different batch sizes on datasets "a9a", "cod-rna", "MNIST", and "Fashion-MNIST".

where we organize the parameters of the model as $\mathbf{x} = (\mathbf{W}_1, \mathbf{c}_1, \mathbf{W}_2, \mathbf{c}_2)$ with $\mathbf{W}_1 \in \mathbb{R}^{256\times 784}$, $\mathbf{c}_1 \in \mathbb{R}^{256}$, $\mathbf{W}_2 \in \mathbb{R}^{10\times 256}$, $\mathbf{c}_2 \in \mathbb{R}^{10}$ and denote $g(\mathbf{x}; \mathbf{a}_i^j) = \mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1 \mathbf{a}_i^j + \mathbf{c}_1) + \mathbf{c}_2$ with $\mathrm{ReLU}(x) = \max(0, x)$. Additionally, we let

$$\ell(\hat{\mathbf{y}}, y) = -\sum_{k=0}^{9} \mathbf{1}_{[y=k]} \log\left(\frac{\exp(\hat{y}_{[k]})}{\sum_{j=0}^{9}\exp(\hat{y}_{[j]})}\right),$$

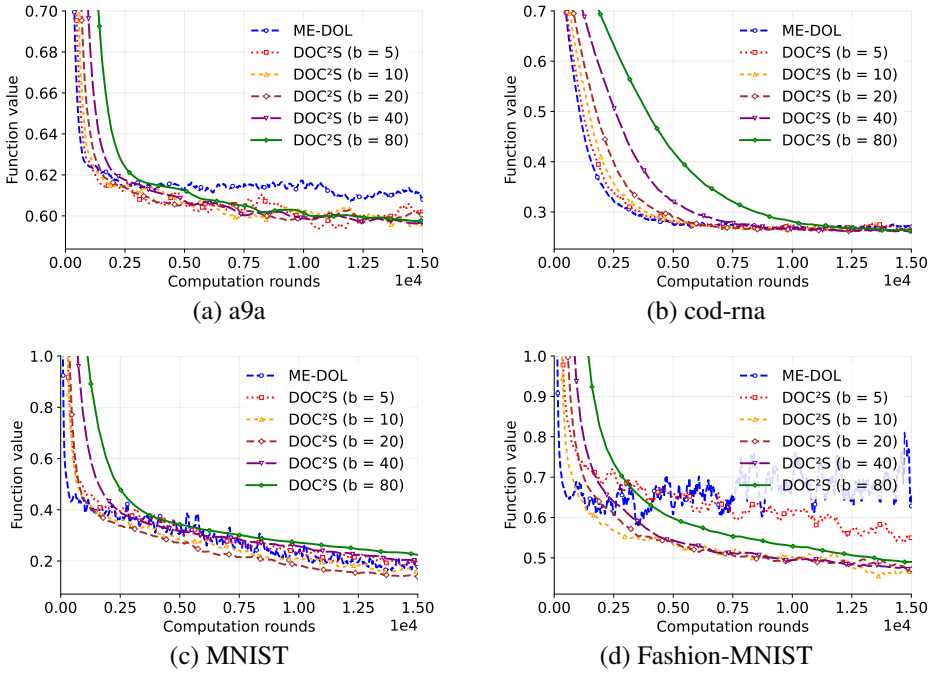
where $\hat{y}_j$ is the j-th coordinate of $\hat{\mathbf{y}}$. We also denote $\mathbf{a}_i^j \in \mathbb{R}^{784}$ and $b_i^j \in \{0, 1, \ldots, 9\}$ as the feature (flattened 28×28 images) of the jth sample on the ith client and its corresponding label.

We perform our experiments on a random-walk ring graph with $n = 10$ nodes. For both ME-DOL and DOC^{2}S, we tune the stepsize η and diameter D from $\{0.001, 0.005, 0.01\}$ and $\{0.05, 0.01, 0.005\}$, respectively. For our DOC^{2}S, we compare the performance of the algorithm across different mini-batch sizes $b \in \{5, 10, 20, 40, 80\}$ and set $R = 2$ for the subroutine of Algorithm 2.

We present experimental results in terms of sample complexity, computation rounds, and communication rounds in Figures 1, 2, and 3, respectively. We can observe that by taking appropriate mini-batch size, the proposed DOC^{2}S performs better than baselines with respect to all measures. For DOC^{2}S, Figure 3 shows that the larger mini-batch size b generally lead to less communication complexity, while the marginal benefit gradually diminishes as b further increases. This phenomenon supports our theoretical results in Corollary 2 and discussion in Remark 3.

6 Conclusion

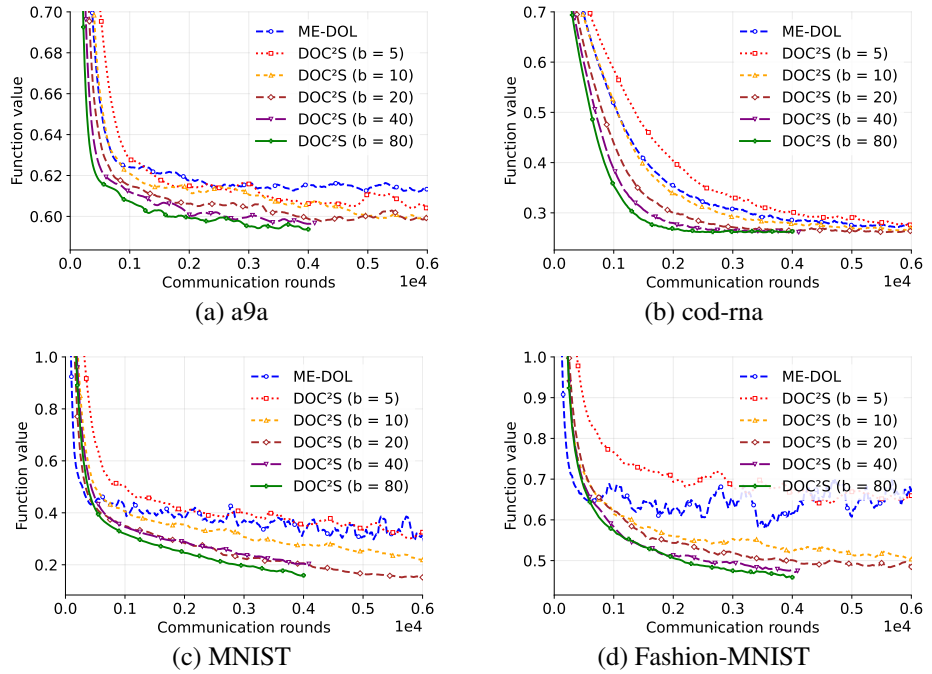
The paper presents a novel stochastic optimization methods with client-data sampling for decentralized nonsmooth nonconvex problem. We further provide theoretical analysis to show our method significantly improves the computation and the communication efficiencies. The advantage of proposed method is also validated by empirical studies. In future work, it is possible to extend our ideas to design the efficient zeroth-order methods (Lin et al., 2022; Chen et al., 2023; Lin et al., 2024; Kornowski & Shamir, 2024) and we can also apply the client-data sampling strategy to address the optmization problem in time varying networks (Kovalev et al., 2024).



(a) a9a (b) cod-rna

(c) MNIST (d) Fashion-MNIST

Figure 2: Comparison on the computation rounds of ME-DOL and DOC^{2}S with different batch sizes on datasets "a9a", "cod-rna", "MNIST", and "Fashion-MNIST".



(a) a9a (b) cod-rna

(c) MNIST (d) Fashion-MNIST

Figure 3: Comparison on the communication1 rounds of ME-DOL and DOC^{2}S with different batch sizes on datasets "a9a", "cod-rna", "MNIST", and "Fashion-MNIST".

AI USE STATEMENT

In this work, We used generative AI tools solely to proofread the manuscript and polish language. All AI-assisted edits were carefully reviewed and revised by the authors. We take responsibility for the final content of this work.

Reproducibility Statement

We provide complete proofs for all theoretical claims in the appendix. To facilitate reproducibility, we include the source code in the supplementary material.

References

Mario Arioli and Jennifer Scott. Chebyshev acceleration of iterative refinement. *Numerical Algorithms*, 66(3):591–608, 2014.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023.

Yunyan Bai, Yuxing Liu, and Luo Luo. On the complexity of finite-sum smooth optimization under the Polyak–Łojasiewicz condition. In *International Conference on Machine Learning*, pp. 2392–2417, 2024.

Hongxu Chen, Ke Wei, Haishan Ye, and Luo Luo. A near-optimal algorithm for decentralized convex-concave finite-sum minimax optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 121043–121093, 2025.

Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In *International Conference on Machine Learning*, pp. 5219–5233, 2023.

Wenlin Chen, Samuel Horvath, and Peter Richtárik. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020.

Francis H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, and Peter R. Wolenski. *Nonsmooth Analysis and Control Theory*, volume 178. Springer Science & Business Media, 2008.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pp. 6643–6670, 2023.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In *Advances in Neural Information Processing Systems*, pp. 6692–6703, 2022.

John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Darrell Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, 2010.

LawrenceCraig Evans. *Measure Theory and Fine Properties of Functions*. Routledge, 2018.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

A.A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13: 14–22, 1977.

Michael I. Jordan, Tianyi Lin, and Manolis Zampetakis. On the complexity of deterministic nonsmooth and nonconvex optimization. *arXiv preprint arXiv:2209.12463*, 2022.

Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 324–334, 2021.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122): 1–14, 2024.

Dmitry Kovalev, Ekaterina Borodich, Alexander Gasnikov, and Dmitrii Feoktistov. Lower bounds and optimal algorithms for non-smooth convex decentralized optimization over time-varying networks. In *Advances in Neural Information Processing Systems*, pp. 96566–96606, 2024.

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180(1):237–284, 2020.

Tianyi Lin, Zeyu Zheng, and Michael I. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 26160–26175, 2022.

Zhenwei Lin, Jingfan Xia, Qi Deng, and Luo Luo. Decentralized gradient-free methods for stochastic non-smooth non-convex optimization. In *AAAI Conference on Artificial Intelligence*, pp. 17477–17486, 2024.

Yuxing Liu, Lesi Chen, and Luo Luo. Decentralized convex finite-sum optimization with better dependence on condition numbers. In *International Conference on Machine Learning*, pp. 30807–30841, 2024.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, volume 2025, pp. 92529–92554, 2025.

Luo Luo, Yunyan Bai, Lesi Chen, Yuxing Liu, and Haishan Ye. On the complexity of decentralized smooth nonconvex finite-sum optimization. *arXiv preprint arXiv:2210.13931*, 2022.

Artavazd Maranjyan, Mher Safaryan, and Peter Richtárik. Gradskip: Communication-accelerated local gradient methods with better computational complexity. *arXiv preprint arXiv:2210.16402*, 2022.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. ProxSkip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pp. 15750–15769, 2022.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning*, pp. 807–814, 2010.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Emre Sahinoglu and Shahin Shahrampour. An online optimization perspective on first-order and zero-order decentralized nonsmooth nonconvex stochastic optimization. In *International Conference on Machine Learning*, pp. 43043–43059, 2024.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Laurent Massoulié, and Yin Tat Lee. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems*, pp. 2745–2754, 2018.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

Shahin Shahrampour and Ali Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3):714–725, 2017.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Hartmut Stadtler. Supply chain management: An overview. *Supply Chain Management and Advanced Planning: Concepts, Models, Software, and Case Studies*, pp. 3–28, 2014.

Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, 208:51–74, 2024.

Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In *International Conference on Machine Learning*, pp. 21360–21379, 2022.

Jinxin Wang, Jiang Hu, Shixiang Chen, Zengde Deng, and Anthony Man-Cho So. Decentralized weakly convex optimization over the Stiefel manifold. *arXiv preprint arXiv:2303.17779*, 2023.

Nachuan Xiao, Xiaoyin Hu, Xin Liu, and Kim-Chuan Toh. Adam-family methods for nonsmooth optimization with convergence guarantees. *Journal of Machine Learning Research*, 25:1–53, 2024.

Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *Journal of Machine Learning Research*, 24(306):1–50, 2023.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010a.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182, 2020.

Siyuan Zhang, Nachuan Xiao, and Xin Liu. Decentralized stochastic subgradient methods for nonsmooth nonconvex optimization. *arXiv preprint arXiv:2403.11565*, 2024.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(3):1081–1107, 2010b.

A SOME USEFUL LEMMAS

We first provide some useful lemmas for our later analysis.

Lemma 9. *Under Assumptions 1 and 3, it holds*

$$\mathbb{E}[\|\nabla F_i(\mathbf{x};\boldsymbol{\xi}_i)\|^2] \le 2\mathbb{E}\left[\|\nabla f_i(\mathbf{x})\|^2 + \|\nabla f_i(\mathbf{x}) - \nabla F_i(\mathbf{x};\boldsymbol{\xi}_i)\|^2\right] \le 2(L^2+\sigma^2).$$

Lemma 10 ((Ye et al., 2023, Lemma 18))**.** *For all $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $D > 0$, it holds*

$$\left\|\mathbf{Clip}\left(\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{X}, D\right) - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{Clip}(\mathbf{X}, D)\right\|_F \le \left\|\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{X}\right\|_F. \tag{8}$$

Lemma 11 ((Liu & Zhou, 2025, Lemma 4.3))**.** *Given a vector-valued martingale difference sequence $\mathbf{v}_1, \dots, \mathbf{v}_T$, we have*

$$\mathbb{E}\left[\left\|\sum_{t=1}^T \mathbf{v}_t\right\|\right] \le 2\sqrt{2}\mathbb{E}\left[\left(\sum_{t=1}^T \|\mathbf{v}_t\|^p\right)^{\frac{1}{p}}\right], \forall p \in [1,2].$$

Lemma 12 ((Shahrampour & Jadbabaie, 2017, Lemma 6))**.** *For given $\mathbf{g}, \mathbf{V} \in \mathbb{R}^d$ and $D > 0$, the problem*

$$\min_{\|\mathbf{x}\|\le D}\left\{\langle \mathbf{x}, \mathbf{g}\rangle + \frac{1}{2\eta}\|\mathbf{x}-\mathbf{V}\|^2\right\} \tag{9}$$

has the unique solution

$$\mathbf{x}^* = \min\left\{1, \frac{D}{\|\boldsymbol{\Delta} - \eta\mathbf{g}\|}\right\}(\mathbf{V} - \eta\mathbf{g}).$$

Additionally, we have

$$\langle \mathbf{g}, \mathbf{x}^* - \mathbf{u}\rangle \le \frac{1}{2\eta}\|\mathbf{V}-\mathbf{u}\|^2 - \frac{1}{2\eta}\|\mathbf{u}-\mathbf{x}^*\|^2 - \frac{1}{2\eta}\|\mathbf{V}-\mathbf{x}^*\|^2,$$

for all $\mathbf{u} \in \mathbb{R}^d$.

B UPPER BOUNDS FOR CONSENSUS ERRORS

We present the proofs for Lemmas 1 and 2, which provide upper bounds for consensus errors.

B.1 PROOF OF LEMMA 1

Proof. We let

$$\hat{\mathbf{V}}_{t+\frac{1}{2}} = \mathbf{V}_t - \frac{\eta n \mathbf{G}_t}{b} \qquad \text{and} \qquad \mathbf{V}_{t+\frac{1}{2}} = \text{FastGossip}\left(\hat{\mathbf{V}}_{t+\frac{1}{2}}, \mathbf{P}, R\right),$$

then the iteration of Algorithm 1 and Proposition 2 implies

$$\left\|\mathbf{V}_{t+\frac{1}{2}} - \mathbf{1}\bar{\mathbf{v}}_{t+\frac{1}{2}}^\top\right\|_F^2 \le 14c^{2R}\left\|\hat{\mathbf{V}}_{t+\frac{1}{2}} - \mathbf{1}\bar{\mathbf{v}}_{t+\frac{1}{2}}^\top\right\|_F^2, \tag{10}$$

where $c = 1 - (1-1/\sqrt{2})\sqrt{\gamma}$. Thus, we have

$$\mathbb{E}\left[\left\|\hat{\mathbf{V}}_{t+\frac{1}{2}} - \mathbf{1}\bar{\mathbf{v}}_{t+\frac{1}{2}}^\top\right\|_F^2\right] \le \mathbb{E}\left[\left\|\hat{\mathbf{V}}_{t+\frac{1}{2}}\right\|_F^2\right] = \mathbb{E}\left[\left\|\mathbf{V}_t - \frac{\eta n\mathbf{G}_t}{b}\right\|_F^2\right] \le 2n(D^2 + \eta n(G^2+\sigma^2)),$$

where the last step comes from $\|\mathbf{v}_{t,i}\| \le D$ and the fact

$$\frac{1}{b}\mathbb{E}[\|\mathbf{G}_t\|^2] \le \frac{1}{b}\sum_{i=1}^n\sum_{j=1}^{b_{t,i}}\mathbb{E}\left[\|\nabla F_i(\mathbf{w}_{t,i} + \mu\mathbf{z}_{t,i};\boldsymbol{\xi}_{t,i,j})\|^2\right]$$

$$\leq \frac{1}{b}\sum_{i=1}^{n}\sum_{j=1}^{b_{t,i}}\mathbb{E}\left[2\left\|\nabla f_i(\mathbf{w}_{t,i}+\mu\mathbf{z}_{t,i})\right\|^2+2\left\|\nabla F_i(\mathbf{w}_{t,i}+\mu\mathbf{z}_{t,i};\boldsymbol{\xi}_{t,i,j})-\nabla f_i(\mathbf{w}_{t,i}+\mu\mathbf{z}_{t,i})\right\|^2\right]$$
$$\leq 2(L^2+\sigma^2)$$

By the property of the $\mathbf{Clip}$ operator, we have

$$\begin{aligned}\left\|\mathbf{V}_{t+1}-\mathbf{1}\bar{\mathbf{v}}_{t+1}^\top\right\|_F^2 &= \left\|\mathbf{Clip}(\mathbf{V}_{t+\frac{1}{2}},D)-\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{Clip}(\mathbf{V}_{t+\frac{1}{2}},D)\right\|_F^2\\ &\leq \left\|\mathbf{Clip}(\mathbf{V}_{t+\frac{1}{2}},D)-\mathbf{Clip}(\mathbf{1}\bar{\mathbf{v}}_{t+\frac{1}{2}}^\top,D)\right\|_F^2+\left\|\mathbf{Clip}(\mathbf{1}\bar{\mathbf{v}}_{t+\frac{1}{2}}^\top,D)-\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{Clip}(\mathbf{V}_{t+\frac{1}{2}},D)\right\|_F^2\\ &\leq 2\left\|\mathbf{V}_{t+\frac{1}{2}}-\mathbf{1}\bar{\mathbf{v}}_{t+\frac{1}{2}}^\top\right\|_F^2,\end{aligned}$$

where the last step is because of Lemma 10 and the non-expansiveness of the $\mathbf{Clip}$ operator.

Combining above results, we have

$$\mathbb{E}\left[\left\|\mathbf{V}_{t+1}-\mathbf{1}\bar{\mathbf{v}}_{t+1}^\top\right\|_F\right]\leq\sqrt{\mathbb{E}\left[\left\|\mathbf{V}_{t+1}-\mathbf{1}\bar{\mathbf{v}}_{t+1}^\top\right\|_F^2\right]}\leq c^R\sqrt{28(nD^2+\eta n^2(L^2+\sigma^2))}.$$

Recall that $c=1-(1-1/\sqrt{2})\sqrt{\gamma}$. Therefore, the setting of R and Proposition 2 implies

$$\mathbb{E}\left[\left\|\mathbf{V}_{t+1}-\mathbf{1}\bar{\mathbf{v}}_{t+1}^\top\right\|_F\right]\leq\epsilon' \tag{11}$$

for all $i\in[n]$. □

B.2 Proof of Lemma 2

Proof. Based on the update rule of $\mathbf{X}_t$ in Algorithm 1 (line 10), we get

$$\begin{aligned}&\left\|\mathbf{X}_t-\bar{\mathbf{x}}_t\right\|_F\\ \leq&\sqrt{14}c^R\left\|\mathbf{X}_{t-1}+\mathbf{V}_t-(\bar{\mathbf{x}}_{t-1}+\bar{\mathbf{v}}_t)\right\|_F\\ \leq&\sqrt{14}c^R\left\|\mathbf{X}_{t-1}-\bar{\mathbf{x}}_{t-1}\right\|_F+\sqrt{14}c^R\left\|\mathbf{V}_t-\bar{\mathbf{v}}_t\right\|_F\\ \leq&\sqrt{14}c^R\left\|\mathbf{X}_{t-1}-\bar{\mathbf{x}}_{t-1}\right\|_F+\sqrt{14}c^R\left\|\mathbf{V}_t\right\|_F\\ \leq&\sqrt{14}c^R\left\|\mathbf{X}_{t-1}-\bar{\mathbf{x}}_{t-1}\right\|_F+\sqrt{14}c^R\cdot\sqrt{n}D,\end{aligned} \tag{12}$$

where the first inequality is based on Proposition 2, the second inequality is based on the triangle inequality of Frobenius norm and the last step is based on the fact that $\|\mathbf{v}_{t,i}\|\leq D$ for $i\in[n]$.

We then use induction to prove

$$\left\|\mathbf{X}_t-\bar{\mathbf{x}}_t\right\|_F\leq\epsilon'$$

for all $t\geq 1$ and $\epsilon'<D$ as follows.

Induction Base: For $t=0$, we have

$$\mathbf{X}_0-\bar{\mathbf{x}}_0=0\leq\epsilon'.$$

Induction Step: We suppose

$$\left\|\mathbf{X}_{t-1}-\bar{\mathbf{x}}_{t-1}\right\|_F\leq\epsilon'$$

holds. Substituting the induction hypothesis into equation (12) implies

$$\left\|\mathbf{X}_t-\bar{\mathbf{x}}_t\right\|_F\leq\sqrt{14}c^R\cdot\epsilon'+\sqrt{14}c^R\cdot\sqrt{n}D\leq\epsilon',$$

where we take

$$R\geq\left\lceil\frac{1}{(1-1/\sqrt{2})\sqrt{\gamma}}\log\frac{2\sqrt{14n}D}{\epsilon'}\right\rceil.$$

□

C PROOFS FOR THE SMOOTH CASE

We provide the proofs of lemmas for the smooth case in Section 4

C.1 PROOF OF LEMMA 3

Proof. Recall that

$$\tilde{\bar{\boldsymbol{\nabla}}}_t = \int_0^1 \nabla f(\bar{\mathbf{x}}_{t-1} + s\bar{\mathbf{v}}_t)\,\mathrm{d}s.$$

We can split S_1 in equation 6 as

$$\begin{aligned}&\sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \tilde{\bar{\boldsymbol{\nabla}}}_t - \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t\rangle] \\ =&\sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \bar{\boldsymbol{\nabla}}_t - \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t\rangle] + \sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \tilde{\bar{\boldsymbol{\nabla}}}_t - \bar{\boldsymbol{\nabla}}_t, \bar{\mathbf{v}}_t\rangle],\end{aligned} \tag{13}$$

where

$$\bar{\boldsymbol{\nabla}}_t = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\nabla}_{t,i} \qquad \text{and} \qquad \boldsymbol{\nabla}_{t,i} = \int_0^1 \nabla f_i(\mathbf{x}_{t-1,i} + s\mathbf{V}_{t,i})\,\mathrm{d}s.$$

We now consider the upper bound of equation (13). Line 6-7 of Algorithm 1 and Assumption 3 imply

$$\frac{n}{b}\bar{\mathbf{g}}_t = \frac{1}{b}\sum_{j=1}^{b} \nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}; \xi_{t,i_{t,j},k_j})$$

where $i_{t,j} \sim \mathrm{Unif}(\{1,\dots,n\})$. We know that

$$\mathbb{E}[\frac{n}{b}\bar{\mathbf{g}}_t] = \mathbb{E}[\nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}; \xi_{t,i_{t,j},k_j})] = \mathbb{E}[\boldsymbol{\nabla}_{t,i_{t,j}}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\nabla}_{t,i}] = \mathbb{E}[\bar{\boldsymbol{\nabla}}_t],$$

Therefore, the first term of equation (13) equal to 0.

Note that we have

$$\begin{aligned}&\left\|\tilde{\bar{\boldsymbol{\nabla}}}_t - \bar{\boldsymbol{\nabla}}_t\right\| \\ =&\left\|\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left(\nabla f_i\big(\bar{\mathbf{x}}_{t-1} + s\bar{\mathbf{v}}_t\big) - \nabla f_i\big(\mathbf{x}_{t-1,i} + s\mathbf{v}_{t,i}\big)\right)\mathrm{d}s\right\| \\ \le&\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \left\|\nabla f_i\big(\bar{\mathbf{x}}_{t-1} + s\bar{\mathbf{v}}_t\big) - \nabla f_i\big(\mathbf{x}_{t-1,i} + s\mathbf{v}_{t,i}\big)\right\|\mathrm{d}s \\ \le&\frac{H}{n}\sum_{i=1}^{n}\int_0^1 \left\|\bar{\mathbf{x}}_{t-1} + s\bar{\mathbf{v}}_t - \mathbf{x}_{t-1,i} - s\mathbf{v}_{t,i}\right\|\mathrm{d}s \\ \le&\frac{H}{n}\sum_{i=1}^{n}\left\|\bar{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1,i}\right\| + \frac{H}{2n}\sum_{i=1}^{n}\left\|\bar{\mathbf{v}}_t - \mathbf{v}_{t,i}\right\|,\end{aligned} \tag{14}$$

where the second inequality is based on the H-smoothness of the function f_i.

According to Lemma 2, we have

$$\mathbb{E}[\|\bar{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1,i}\|] \le \epsilon'. \tag{15}$$

According to Lemma 1, we have

$$\mathbb{E}[\|\bar{\mathbf{v}}_t - \mathbf{v}_{t,i}\|] \le \epsilon' \tag{16}$$

Substituting equations (15) and (16) into equation (14), we have

$$\mathbb{E}[\|\tilde{\bar{\nabla}}_t - \bar{\nabla}_t\|] \leq 2H\epsilon'. \tag{17}$$

Therefore, the second term in equation (13) holds

$$\begin{aligned}
&\sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \tilde{\bar{\nabla}}_t - \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t \rangle] \\
=&\sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \tilde{\bar{\nabla}}_t - \bar{\nabla}_t, \bar{\mathbf{v}}_t \rangle] \\
\leq&\sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\|\tilde{\bar{\nabla}}_t - \bar{\nabla}_t\| \|\bar{\mathbf{v}}_t\|] \\
\leq&2DH\epsilon' T,
\end{aligned}$$

where the second step is based on Cauchy–Schwarz inequality, and the last step is based on equation (17) and the fact $\|\bar{\mathbf{v}}_t\| \leq D$. □

C.2 Proof of Lemma 4

Proof. By setting

$$\mathbf{u}_k = -D\frac{\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})}{\left\|\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right\|},$$

we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \sum_{t=(k-1)T+1}^{kT}\frac{n}{b}\bar{\mathbf{g}}_t, \mathbf{u}_k\right\rangle\right] \\
=&\mathbb{E}\left[\left\langle \frac{1}{n}\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i}), \mathbf{u}_k\right\rangle\right] + \mathbb{E}\left[\left\langle \sum_{t=(k-1)T+1}^{kT}\left(\frac{n}{b}\bar{\mathbf{g}}_t - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right), \mathbf{u}_k\right\rangle\right] \\
\leq&-D\mathbb{E}\left[\left\|\frac{1}{n}\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right\|\right] + D\mathbb{E}\left[\left\|\sum_{t=(k-1)T+1}^{kT}\left(\frac{n}{b}\bar{\mathbf{g}}_t - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right)\right\|\right] \\
\leq&-D\mathbb{E}\left[\left\|\frac{1}{n}\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right\|\right] + D\mathbb{E}\left[\left\|\frac{1}{b}\sum_{t=(k-1)T+1}^{kT}\sum_{j=1}^{b}(\nabla f_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i}))\right\|\right] \\
&+D\mathbb{E}\left[\left\|\frac{1}{b}\sum_{t=(k-1)T+1}^{kT}\sum_{j=1}^{b}(\nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}, \xi_{t,i_{t,j},j}) - \nabla f_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}))\right\|\right]
\end{aligned}$$

where the last inequality is based on Cauchy–Schwarz inequality and the fact that $\left\|\mathbf{u}^k\right\| \leq D$. Since $\mathbb{E}_\xi[\nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}})] = \nabla f_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}))$ and $\mathbb{E}_{i_{t,j}}[\nabla f_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}})] = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})$ By Lemma 11 and Jensen's inequality, we can get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{t=(k-1)T+1}^{kT}\sum_{j=1}^{b}(\nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}) - \nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}))\right\|\right] \\
\leq&2\sqrt{2}\mathbb{E}\left[\left(\sum_{t=(k-1)T+1}^{kT}\sum_{j=1}^{b}\left\|\nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}}) - \nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}})\right\|^2\right)^{\frac{1}{2}}\right]
\end{aligned}$$

$$\leq 2\sqrt{2}\left(\mathbb{E}\left[\sum_{t=(k-1)T+1}^{kT}\sum_{j=1}^{b}\left\|\nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}})-\nabla F_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}})\right\|^2\right]\right)^{\frac{1}{2}}$$
$$\leq 2\sqrt{2}(Tb\sigma^2)^{\frac{1}{2}}.$$

and

$$\mathbb{E}\left[\left\|\sum_{t=(k-1)T+1}^{kT}\sum_{j=1}^{b}(\nabla f_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}})-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i}))\right\|\right]$$
$$\leq 2\sqrt{2}\mathbb{E}\left[\left(\sum_{t=(k-1)T+1}^{kT}\sum_{j=1}^{b}\left\|\nabla f_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}})-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i}))\right\|^2\right)^{\frac{1}{2}}\right]$$
$$\leq 2\sqrt{2}\left(\mathbb{E}\left[\sum_{t=(k-1)T+1}^{kT}\sum_{j=1}^{b}\left\|\nabla f_{i_{t,j}}(\mathbf{w}_{t,i_{t,j}})-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i}))\right\|^2\right]\right)^{\frac{1}{2}}$$
$$\leq 2\sqrt{2}(Tb(2L)^2)^{\frac{1}{2}}.$$

Thus we have

$$\mathbb{E}\left[\left\langle\sum_{t=(k-1)T+1}^{kT}\frac{n}{b}\bar{\mathbf{g}}_t,\mathbf{u}_k\right\rangle\right]\leq -D\mathbb{E}\left[\left\|\frac{1}{n}\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right\|\right]+\frac{2\sqrt{2}D(\sigma+2L)\sqrt{T}}{\sqrt{b}}.$$

□

C.3 Proof of Lemma 5

Proof. Let $\tilde{\mathbf{v}}_{t+1}=\mathbf{clip}(\bar{\mathbf{v}}_t-\frac{n\eta\bar{\mathbf{g}}_t}{b},D)$. By Lemma 12 with $\mathbf{g}=\frac{n}{b}\bar{\mathbf{g}}_t$, $\mathbf{V}=\bar{\mathbf{v}}_t$, $\mathbf{u}=\mathbf{u}_k$ and $\mathbf{x}^*=\tilde{\mathbf{v}}_{t+1}$, we have

$$\langle\frac{n}{b}\bar{\mathbf{g}}_t,\tilde{\mathbf{v}}_{t+1}-\mathbf{u}_k\rangle\leq\frac{1}{2\eta}\left\|\bar{\mathbf{v}}_t-\mathbf{u}_k\right\|^2-\frac{1}{2\eta}\left\|\mathbf{u}_k-\tilde{\mathbf{v}}_{t+1}\right\|^2-\frac{1}{2\eta}\left\|\bar{\mathbf{v}}_t-\tilde{\mathbf{v}}_{t+1}\right\|^2. \tag{18}$$

Thus we have

$$\begin{aligned}\langle\frac{n}{b}\bar{\mathbf{g}}_t,\bar{\mathbf{v}}_t-\mathbf{u}_k\rangle\leq&\frac{1}{2\eta}\left(\left\|\bar{\mathbf{v}}_t-\mathbf{u}_k\right\|^2-\left\|\tilde{\mathbf{v}}_{t+1}-\mathbf{u}_k\right\|^2\right)-\frac{1}{2\eta}\left\|\bar{\mathbf{v}}_t-\tilde{\mathbf{v}}_{t+1}\right\|^2+\langle\frac{n}{b}\bar{\mathbf{g}}_t,\bar{\mathbf{v}}_t-\tilde{\mathbf{v}}_{t+1}\rangle\\ \leq&\frac{1}{2\eta}\left(\left\|\bar{\mathbf{v}}_t-\mathbf{u}_k\right\|^2-\left\|\bar{\mathbf{v}}_{t+1}-\mathbf{u}_k\right\|^2\right)-\frac{1}{2\eta}\left\|\bar{\mathbf{v}}_t-\tilde{\mathbf{v}}_{t+1}\right\|^2+\langle\frac{n}{b}\bar{\mathbf{g}}_t,\bar{\mathbf{v}}_t-\tilde{\mathbf{v}}_{t+1}\rangle\\ &+\frac{1}{2\eta}\|\tilde{\mathbf{v}}_{t+1}-\bar{\mathbf{v}}_{t+1}\|\|\tilde{\mathbf{v}}_{t+1}+\bar{\mathbf{v}}_{t+1}-2\mathbf{u}_k\|.\end{aligned}$$

Note that we have

$$\langle\frac{n}{b}\bar{\mathbf{g}}_t,\bar{\mathbf{v}}_t-\tilde{\mathbf{v}}_{t+1}\rangle\leq\|\frac{n}{b}\bar{\mathbf{g}}_t\|\|\bar{\mathbf{v}}_t-\tilde{\mathbf{v}}_{t+1}\|\leq\frac{\eta}{2}\|\frac{n}{b}\bar{\mathbf{g}}_t\|^2+\frac{1}{2\eta}\left\|\bar{\mathbf{v}}_t-\tilde{\mathbf{v}}_{t+1}\right\|^2.$$

By Lemma 10, we also have

$$\mathbb{E}[\|\tilde{\mathbf{v}}_{t+1}-\bar{\mathbf{v}}_{t+1}\|]\leq\mathbb{E}\left[\left\|\mathbf{V}_{t+\frac{1}{2}}-\mathbf{1}\bar{\mathbf{v}}_{t+\frac{1}{2}}^\top\right\|_F\right]\leq\epsilon'.$$

Combine all above together, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=(k-1)T+1}^{kT}\langle\frac{n}{b}\bar{\mathbf{g}}_t,\bar{\mathbf{v}}_t-\mathbf{u}_k\rangle\right]\\ \leq&\frac{\|\bar{\mathbf{v}}_0-\mathbf{u}_k\|^2}{2\eta}+\sum_{t=(k-1)T+1}^{kT}\mathbb{E}[\frac{\eta}{2}\|\frac{n}{b}\bar{\mathbf{g}}_t\|^2]+\sum_{t=(k-1)T+1}^{kT}\frac{4D\mathbb{E}[\|\tilde{\mathbf{v}}_{t+1}-\bar{\mathbf{v}}_{t+1}\|]}{2\eta}\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{D^2}{2\eta} + \frac{T\eta(L^2+\sigma^2/b)}{2} + \frac{2TD\epsilon'}{\eta} \\
&\leq D\sqrt{T(L^2+\sigma^2/b)} + \epsilon' T\sqrt{T(L^2+\sigma^2/b)},
\end{aligned}
$$

where the last step comes from the choice of $\eta = \frac{D}{\sqrt{T(L^2+\sigma^2/b)}}$. □

C.4 Proof of Lemma 7

Proof. According to the update rule of $\mathbf{x}_t$ in Algorithm 1 (line 10), we have:

$$\bar{\mathbf{x}}_t = \bar{\mathbf{x}}_{t-1} + \bar{\mathbf{v}}_t.$$

where the last step is based on the doubly stochastic assumption of matrix $\mathbf{P}$ (Assumption 4).

Recall that

$$\tilde{\boldsymbol{\nabla}}_t = \int_0^1 \nabla f(\bar{\mathbf{x}}_{t-1} + s\bar{\mathbf{v}}_t)\mathrm{d}s$$

then the continuity of the function f means

$$f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t-1}) = \int_0^1 \langle \nabla f(\bar{\mathbf{x}}_{t-1} + s\bar{\mathbf{v}}_t), \bar{\mathbf{v}}_t\rangle \mathrm{d}s = \langle \tilde{\boldsymbol{\nabla}}_t, \bar{\mathbf{v}}_t\rangle.$$

Summing over t and taking expectation yields

$$
\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}_{kT}) - f(\bar{\mathbf{x}}_{(k-1)T})] &= \sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \tilde{\boldsymbol{\nabla}}_t, \bar{\mathbf{v}}_t\rangle] \\
&= \underbrace{\sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \tilde{\boldsymbol{\nabla}}_t - \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t\rangle]}_{S_1} + \underbrace{\sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \frac{n}{b}\bar{\mathbf{g}}_t, \mathbf{u}_k\rangle]}_{S_2} + \underbrace{\sum_{t=(k-1)T+1}^{kT} \mathbb{E}[\langle \frac{n}{b}\bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t - \mathbf{u}_k\rangle]}_{S_3}.
\end{aligned}
$$

hold for all $k \in [K]$ and $\mathbf{u}_k \in \mathbb{R}^d$.

Notice that S_1, S_2 and S_3 can be bounded by Lemma 3, Lemma 4 and Lemma 5, respectively. Thus we have

$$
\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}_{kT}) - f(\bar{\mathbf{x}}_{(k-1)T})] \leq & 2DH\epsilon' T - D\mathbb{E}\left[\left\|\frac{1}{n}\sum_{t=(k-1)T+1}^{kT}\sum_{i=1}^{n}\nabla f_i(\mathbf{w}_{t,i})\right\|\right] \\
& + \frac{2\sqrt{2}D(\sigma+2L)\sqrt{T}}{\sqrt{b}} + D\sqrt{T(L^2+\sigma^2/b)} + \epsilon' T\sqrt{T(L^2+\sigma^2/b)}
\end{aligned} \tag{19}
$$

Next, we target to bound the distance from $\{\mathbf{w}_i^{t,k}\}$ to $\mathbf{w}_i^{\mathrm{out}}$. For all $i, j \in [n]$, $k \in [K]$, and $t_1, t_2 \in [(k-1)T+1, kT]$ such that $t_1 > t_2$, we have

$$\|\mathbf{w}_{t_1,i} - \mathbf{w}_{t_2,j}\| \leq \|\mathbf{w}_{t_1,i} - \bar{\mathbf{w}}_{t_1}\| + \|\bar{\mathbf{w}}_{t_1} - \bar{\mathbf{w}}_{t_2}\| + \|\bar{\mathbf{w}}_{t_2} - \mathbf{w}_{t_2,j}\|. \tag{20}$$

The update rule of $\mathbf{w}_{t,i}$ implies

$$
\begin{aligned}
\|\mathbf{w}_{t,i} - \bar{\mathbf{w}}_t\| &= \left\|\mathbf{x}_{t-1,i} + s_t\mathbf{v}_{t,i} - \bar{\mathbf{x}}_{t-1} - \frac{1}{n}\sum_{i=1}^{n} s_t\mathbf{v}_{t,i}\right\| \\
&\leq \|\mathbf{x}_{t-1,i} - \bar{\mathbf{x}}_{t-1}\| + s_t\|\mathbf{v}_{t,i} - \bar{\mathbf{v}}\| \\
&\leq 2\epsilon'
\end{aligned} \tag{21}
$$

for all $t \in [T]$ and $\epsilon' \leq D/2$, where the last inequality is based on the fact $s_t \leq 1$ consensus error bounds. Consequently, we have

$$\begin{aligned}
&\|\bar{\mathbf{w}}_{t_1} - \bar{\mathbf{w}}_{t_2}\| \leq \sum_{t=t_2}^{t_1-1} \|\bar{\mathbf{w}}_{t+1} - \bar{\mathbf{w}}_t\| \leq \sum_{t=1}^{T} \left\|\bar{\mathbf{w}}^{k,t+1} - \bar{\mathbf{w}}^{k,t}\right\| \\
=&\sum_{t=1}^{T-1} \left\| \bar{\mathbf{x}}_t + \frac{1}{n}\sum_{i=1}^{n} s_{t+1}\mathbf{v}_{t+1,i} - \bar{\mathbf{x}}_{t-1} - \frac{1}{n}\sum_{i=1}^{n} s_t \mathbf{v}_{t,i} \right\| \\
=&\sum_{t=1}^{T-1} \left\| \bar{\mathbf{x}}_{t-1} + \mathbf{v}_{t,i} + \frac{1}{n}\sum_{i=1}^{n} s_{t+1}\mathbf{v}_{t+1,i} - \bar{\mathbf{x}}_{t-1} - \frac{1}{n}\sum_{i=1}^{n} s_t \mathbf{v}_{t,i} \right\| \\
\leq&\sum_{t=1}^{T} \left(\|\mathbf{v}_{t,i}\| + \left\| \frac{1}{n}\sum_{i=1}^{n} s_{t+1}\mathbf{v}_{t+1,i} \right\| + \left\| \frac{1}{n}\sum_{i=1}^{n} s_t \mathbf{v}_{t,i} \right\| \right) \\
\leq&3DT,
\end{aligned} \tag{22}$$

where we use the update rule of $\mathbf{w}_{t,i}$ and $\mathbf{x}_i^{k,t}$ and the setting $s_t \in [0, 1]$.

Substituting equations (21) and (22) into equation (20), we have

$$\left\| \mathbf{w}_i^{k,t_1} - \mathbf{w}_j^{k,t_2} \right\| \leq 2\epsilon' + 3DT \leq \delta, \tag{23}$$

where the last inequality is based on taking

$$D = \frac{\delta}{4T}, \qquad T > 6, \qquad \text{and} \qquad \epsilon' \leq \frac{D}{3T}. \tag{24}$$

Taking the average on inequality 19 over $k = 1, \ldots, K$ and dividing DT, we achieve

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \left\| \frac{1}{nT} \sum_{t=(k-1)T+1}^{kT} \sum_{i=1}^{n} \nabla f_i(\mathbf{w}_{t,i}) \right\| \right] \\
\leq&\frac{\mathbb{E}[f(\bar{\mathbf{x}}^{1,0}) - f(\bar{\mathbf{x}}^{K,T})]}{DKT} + 2H\epsilon' + \frac{4\sigma}{\sqrt{Tb}} + \frac{8L}{\sqrt{T}} + \epsilon'\sqrt{T(L^2 + \sigma^2/b)}/D.
\end{aligned} \tag{25}$$

Now we start to show the desired approximate stationary point can be achieved at each client. According to Lemma 6 with $r = 2\epsilon'$, we have

$$\left\| \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \nabla f(\mathbf{w}_i^{k,t}) \right\| \leq \left\| \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \nabla f_i(\mathbf{w}_i^{k,t}) \right\| + 2H\epsilon'. \tag{26}$$

Additionally, we have

$$\left\| \hat{\mathbf{w}}_i^k - \mathbf{w}_j^{k,t} \right\| = \left\| \frac{1}{T}\sum_{\tau=1}^{T} \mathbf{w}_i^{k,\tau} - \mathbf{w}_j^{k,t} \right\| \leq \left\| \frac{1}{T}\sum_{\tau=1}^{T} \left(\mathbf{w}_i^{k,\tau} - \mathbf{w}_j^{k,t} \right) \right\| \leq \delta \tag{27}$$

for all $i, j \in [n]$, $k \in [K]$, and $t \in [T]$, where the first step is based on the setting of $\hat{\mathbf{w}}_i^k$ (line 11 of Algorithm 1) and the last step is based on equation (23). Combing above results, we have

$$\begin{aligned}
&\mathbb{E}[\left\|\nabla f(\mathbf{w}_i^{\text{out}})\right\|_\delta] \\
=&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \left\|\nabla f(\hat{\mathbf{w}}_i^k)\right\|_\delta \right] \\
\leq&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \left\| \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \nabla f(\mathbf{w}_i^{k,t}) \right\| \right] \\
\leq&\frac{\mathbb{E}[f(\bar{\mathbf{x}}^{1,0}) - f(\bar{\mathbf{x}}^{K,T})]}{DKT} + 2H\epsilon' + \frac{4\sigma}{\sqrt{Tb}} + \frac{8L}{\sqrt{T}} + \epsilon'\sqrt{T(L^2 + \sigma^2/b)}/D.,
\end{aligned}$$

where the first step is based on the setting of $\mathbf{w}_i^{\text{out}}$ (last line of Algorithm 1), the second step is based on the definition of $\|\nabla f(\cdot)\|_\delta$ (Definition 3), and the last step is based on using equations (25) and (26). By setting $T = \mathcal{O}((L^2 + \sigma^2/b)\epsilon^{-2})$, $K = \mathcal{O}(\delta^{-1}\epsilon^{-1}\Delta)$, $D = \mathcal{O}(\delta/T)$, $R = \mathcal{O}(\gamma^{-1/2}\log(\epsilon'^{-1})$, and $\eta = D/\sqrt{T(L^2 + \sigma^2/b)}$ and $\epsilon' = D/(T(L^2 + \sigma^2)) + H$, We can guarantee

$$\mathbb{E}\left[\left\|\nabla f(\mathbf{w}_i^{\text{out}})\right\|_\delta\right] \leq \epsilon.$$

□

D Proofs for the Nonsmooth Case

This section follows the proof of Lemma 7 to achieve the results in the nonsmooth case.

D.1 Proof of Theorem 1

Proof. Recall that the setting of Theorem 1 takes $\mu = \delta/2$, then Proposition 1 means the function $f_{\delta/2}$ is $\delta L/2$-Lipschitz and $2c_0 L\sqrt{d}/\delta$-smooth. In the view of minimizing the smooth function $f_{\delta/2}$ by Algorithm 1, we can follow the first step in the derivation of equation (19) (in the proof of Lemma 7) by replacing δ, f, σ, and H by $\delta/2$, $f_{\delta/2}$, $\sqrt{L^2 + \sigma^2}$, and $2c_0 L\sqrt{d}/\delta$, respectively. This implies that if we set $\mu = \delta/2$, $T = \mathcal{O}((L^2 + \sigma^2/b)\epsilon^{-2})$, $K = \mathcal{O}(\delta^{-1}\epsilon^{-1}\Delta)$, $D = \mathcal{O}(\delta/T)$, $R = \tilde{\mathcal{O}}(\gamma^{-1/2})$, $\epsilon' = D/(T(L^2 + \sigma^2)) + 2c_0 L\sqrt{d}/\delta$ and $\eta = D/\sqrt{T(L^2 + \sigma^2/b)}$, we can get following results by using Lemma 8

$$\mathbb{E}\left[\left\|\nabla f(\mathbf{w}_i^{\text{out}})\right\|_\delta\right] \leq \mathbb{E}[\|\nabla f_{\delta/2}(\mathbf{w}_i^{\text{out}})\|_{\delta/2}] \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$$

for all $i \in [n]$. Hence, we can achieve the desired (δ, ϵ)-Goldstein stationary point on each client with $T = \mathcal{O}(\epsilon^{-2})$. □

D.2 Proof of Corollary 2

Proof. According to the proof of Theorem 1, we achieve an (δ, ϵ)-Goldstein stationary point of the objective with $KT = \mathcal{O}(\delta^{-1}\epsilon^{-3})$. Since b stochastic gradients are sampled according to a multinomial distribution in each round, with probability at least $1 - \delta'$, the client receiving the largest number of samples receives at most

$$\frac{b}{n} + \log\frac{n}{\delta'}$$

stochastic gradients. Thus the computation rounds should be

$$\tilde{\mathcal{O}}\left(\frac{(bL^2 + \sigma^2)\Delta}{n\delta\epsilon^3}\right).$$

The overall stochastic first-order oracle complexity is

$$\mathcal{O}\left(\frac{(bL^2 + \sigma^2)\Delta}{\delta\epsilon^3}\right).$$

We perform R communication rounds each time and $R = \tilde{\mathcal{O}}(\gamma^{-1/2})$ from Lemma 1. Thus the communication rounds is

$$\tilde{\mathcal{O}}\left(\frac{(L^2 + b^{-1}\sigma^2)\Delta}{\sqrt{\gamma}\delta\epsilon^3}\right)$$

. □