DIMENSION-FREE DECENTRALIZED NONSMOOTH NONCONVEX STOCHASTIC OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate decentralized nonsmooth nonconvex stochastic optimization over a network of n nodes, with the goal of finding an (δ, ϵ)-Goldstein stationary point. The best existing algorithm achieves $O(\delta^{-1}(\epsilon^{-3} + d\epsilon^{-1}))$ sample complexity and $\widetilde{O}(\gamma^{-1/2}\delta^{-1}(\epsilon^{-3} + d\epsilon^{-1}))$ communication complexity, where d is the problem dimension and γ is the spectral gap of the communication matrix. However, the polynomial dependence on d can be a major bottleneck in high-dimensional regimes. In this paper, we propose a novel algorithm that achieves $O(\delta^{-1}\epsilon^{-3})$ sample complexity and $\widetilde{O}(\gamma^{-1/2}\delta^{-1}\epsilon^{-3})$ communication complexity. The primary technique is an elegant decentralized online-to-nonconvex conversion that reduces the original problem to a decentralized online convex optimization (D-OCO) problem. A key property of our conversion is that its consensus requirements can be inherited directly from the consensus of the underlying D-OCO decisions. In particular, this property enables us to establish an explicit connection between the dimension dependence and the consensus error, which in turn shows that the polynomial dependence on d can be removed with only logarithmic additional communication.

1 INTRODUCTION

Decentralized stochastic optimization (DSO) (Ram et al., 2010; Lian et al., 2017) has become an attractive alternative to its centralized counterpart as modern learning problems increasingly involve large-scale models and data distributed across multiple devices. Given a network of n nodes and the problem dimension d, DSO can be formulated as

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

where $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i}[F_i(\mathbf{x};\xi_i)]$ is the local objective held by node i, and each node i can only access the stochastic component $F_i(\cdot;\xi_i) : \mathbb{R}^d \to \mathbb{R}$ indexed by a random variable ξ_i and communicate with its neighboring nodes. For convex objectives, the goal is typically to find a solution with a small objective gap $f(\mathbf{x}) - f(\mathbf{x}^*)$ or a small error $\|\mathbf{x} - \mathbf{x}^*\|$, where $\|\cdot\|$ denotes the Euclidean norm and $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. For nonconvex but smooth objectives, it is common to seek an ϵ-stationary point, i.e., $\|\nabla f(\mathbf{x})\| \le \epsilon$. These cases have been extensively studied, leading to numerous algorithms with theoretical guarantees (Lan et al., 2020; Koloskova et al., 2019; 2020; 2021; Lu & Sa, 2021; Yuan et al., 2022).

However, in many practical problems, such as training ReLU neural networks (Mishkin & Pilanci, 2023), the objective function is often both nonconvex and nonsmooth. Motivated by these problems, we investigate DSO with nonsmooth nonconvex objectives and aim to find an (δ, ϵ)-Goldstein stationary point.[1] Although a few recent studies (Lin et al., 2024; Sahinoglu & Shahrampour, 2024; Chen et al., 2026) have proposed algorithms for this setting, they remain unsatisfactory in terms of both sample and communication complexities. Specifically, the best existing algorithm (Chen et al., 2026) achieves $O(\delta^{-1}(\epsilon^{-3} + d\epsilon^{-1}))$ sample complexity and $\widetilde{O}(\gamma^{-1/2}\delta^{-1}(\epsilon^{-3} + d\epsilon^{-1}))$ communication complexity, where γ is the spectral gap of the communication matrix.[2] Such a polynomial

[1] This is a tractable extension of the ϵ-stationary point to nonsmooth objectives and will be defined later.

[2] The $\tilde{O}(\cdot)$ notation hides constant factors as well as polylogarithmic factors.

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.

dependence on d can limit the scalability of this algorithm in high-dimensional regimes. Thus, it is natural to ask whether the polynomial dependence on d can be further eliminated from both the sample and communication complexities of DSO with nonsmooth nonconvex objectives.

In this paper, we provide an affirmative answer to the above question. Our primary technique is a decentralized extension of the online-to-nonconvex conversion originally proposed for the non-distributed setting (Cutkosky et al., 2023). Specifically, Cutkosky et al. (2023) reduce nonsmooth nonconvex stochastic optimization to an online convex optimization (OCO) problem (Shalev-Shwartz, 2011; Hazan, 2016). The underlying OCO algorithm generates update directions using losses constructed from stochastic gradients evaluated at random interpolation points along the line segments connecting consecutive iterates. Inspired by Cutkosky et al. (2023), our conversion reduces DSO with nonsmooth nonconvex objectives to a decentralized online convex optimization (D-OCO) problem (Wan et al., 2024; 2025). The underlying D-OCO algorithm generates update directions for each node using losses constructed from local stochastic gradients evaluated at random interpolation points along the line segments connecting consecutive local iterates. Compared with Cutkosky et al. (2023), our key additional challenges are to control the consensus among the local iterates and random interpolation points across different nodes, as well as the discrepancy between the average of local gradients and the global gradient evaluated at the average random interpolation point.

To this end, we first observe that the consensus error of the local iterates is proportional to that of the local update directions generated by the D-OCO algorithm. In contrast, the consensus error of the random interpolation points does not generally inherit this property, because different nodes may use different interpolation weights. Interestingly, we show that this issue can be addressed by using a shared random interpolation weight across nodes, which requires only a common random seed. Furthermore, to control the gradient discrepancy error, we apply the classical randomized smoothing technique (Duchi et al., 2012b) to the original objective functions. This yields an error bound proportional to the product of the consensus error of the random interpolation points and the smoothness constant induced by randomized smoothing. Although such a smoothness constant scales as $\sqrt{d}$, our preceding consensus analysis allows us to cancel this factor by sufficiently tightening the consensus error of the underlying D-OCO decisions. Taken together, we reduce all additional challenges introduced by decentralization to controlling the consensus error of the underlying D-OCO decisions. Importantly, existing D-OCO algorithms such as Wan et al. (2024) can achieve the required consensus error with only logarithmic additional communication. Finally, by combining our conversion with the D-OCO algorithm of Wan et al. (2024), we obtain a novel algorithm with $O(\delta^{-1}\epsilon^{-3})$ sample complexity and $\widetilde{O}(\gamma^{-1/2}\delta^{-1}\epsilon^{-3})$ communication complexity.

Note that two existing algorithms (Sahinoglu & Shahrampour, 2024; Chen et al., 2026) can also be interpreted as decentralized online-to-nonconvex conversions and have employed the randomized smoothing technique. However, the consensus among their local iterates is guaranteed by further applying the standard or accelerated gossip (Xiao & Boyd, 2004; Liu & Morse, 2011) to the iterates themselves. This incurs additional communication outside the D-OCO algorithm and complicates the analysis. More critically, without the shared random interpolation weight used in our conversion, the consensus error of their random interpolation points contains an $O(1)$ term. Consequently, the $\sqrt{d}$ factor appearing in their gradient discrepancy error cannot be canceled regardless of how small the consensus error of the underlying D-OCO decisions is. In contrast, our conversion provides a clean and unified way to exploit the consensus property of the underlying D-OCO decisions and allows us to achieve complexity bounds without polynomial dependence on d for the first time.

2 Related Work

In this section, we briefly review related work on nonsmooth nonconvex stochastic optimization in both the non-distributed and decentralized settings.

2.1 Non-distributed Nonsmooth Nonconvex Stochastic Optimization

Early studies on non-distributed nonsmooth nonconvex stochastic optimization mainly focus on establishing asymptotic convergence guarantees (Kiwiel, 2007; Majewski et al., 2018; Davis et al., 2020; Bolte & Pauwels, 2021). The pioneering work of Zhang et al. (2020) provides the first non-

asymptotic analysis for this problem. Note that for nonsmooth objectives, Zhang et al. (2020) show that the ϵ-stationary point cannot be found in finite time. To this end, they introduce a tractable relaxation, namely the (δ, ϵ)-Goldstein stationary point, and establish the first finite-time guarantee for this notion, with $O(\delta^{-1}\epsilon^{-4})$ sample complexity. Moreover, under a deterministic generalized gradient oracle, they obtain an improved oracle complexity of $O(\delta^{-1}\epsilon^{-3})$. Since then, this stationarity criterion has attracted growing research interest, particularly in the deterministic setting (Davis et al., 2022; Tian et al., 2022; Kornowski & Shamir, 2022; Jordan et al., 2023; Tian & So, 2024). The first improvement in the stochastic setting is achieved by Cutkosky et al. (2023), who reduce the sample complexity to $O(\delta^{-1}\epsilon^{-3})$ and show the optimality of this rate. The key technique of Cutkosky et al. (2023) is a novel online-to-nonconvex conversion that employs an OCO algorithm (Hazan, 2016) to generate update directions rather than the iterates themselves. Specifically, the losses fed to the underlying OCO algorithm are constructed from stochastic gradients evaluated at random interpolation points along the line segments connecting consecutive iterates, and the final output is also formed from these interpolation points.

In addition, the more challenging zeroth-order setting, where only function values are available, has been studied recently (Lin et al., 2022; Chen et al., 2023; Kornowski & Shamir, 2024). Specifically, in this setting, the randomized smoothing technique (Duchi et al., 2012b) serves as a natural bridge from zeroth-order information to gradient information. Meanwhile, it transforms the original nonsmooth objective into a smooth surrogate, at the price of introducing a $\sqrt{d}$ factor into the smoothness constant. By applying randomized smoothing and exploiting the smoothness of the surrogate objective, Lin et al. (2022) first propose a gradient-free method with $O(d^{3/2}\delta^{-1}\epsilon^{-4})$ zeroth-order sample complexity. Later, Chen et al. (2023) further improve the zeroth-order sample complexity to $O(d^{3/2}\delta^{-1}\epsilon^{-3})$ by using a variance reduction technique (Fang et al., 2018). More interestingly, Kornowski & Shamir (2024) establish a connection between the Goldstein subdifferentials of the original and smoothed objectives, which allows them to combine randomized smoothing with the online-to-nonconvex conversion of Cutkosky et al. (2023), rather than relying on the smoothness of the surrogate objective. This yields the optimal zeroth-order sample complexity of $O(d\delta^{-1}\epsilon^{-3})$.

2.2 Decentralized Nonsmooth Nonconvex Stochastic Optimization

For decentralized nonsmooth nonconvex stochastic optimization, Lin et al. (2024) provide the first finite-time analysis under the zeroth-order setting. They propose a decentralized gradient-free method (DGFM) and its improved variant called DGFM$^+$ by extending the algorithms of Lin et al. (2022) and Chen et al. (2023), respectively. Specifically, DGFM achieves $O(nd^{3/2}\,\mathrm{poly}(1/\gamma)\delta^{-1}\epsilon^{-4})$ zeroth-order sample complexity and $O(d^{3/2}\,\mathrm{poly}(1/\gamma)\delta^{-1}\epsilon^{-4})$ communication complexity, whereas DGFM$^+$ achieves $O(n^{1/2}d^{1/2}\delta^{-1}(n\epsilon^{-2} + d\epsilon^{-3}))$ zeroth-order sample complexity and $O(n^{1/2}d^{1/2}\,\mathrm{poly}(1/\gamma)\delta^{-1}\epsilon^{-2})$ communication complexity. As in the non-distributed case, the polynomial dependence of these complexity bounds on d is intrinsic to zeroth-order optimization. Later, Sahinoglu & Shahrampour (2024) consider the first-order setting and propose a multi-epoch decentralized online learning algorithm (ME-DOL) by extending the online-to-nonconvex conversion of Cutkosky et al. (2023). They adopt an existing D-OCO algorithm of Shahrampour & Jadbabaie (2018) to generate update directions for each node and encounter the three aforementioned challenges introduced by decentralization.

For the first challenge, i.e., controlling the consensus among the local iterates, they further apply one step of standard gossip (Xiao & Boyd, 2004) to the iterates after each update. For the second challenge, i.e., controlling the consensus of the random interpolation points, no additional mechanism is introduced. Instead, the resulting bound consists of the consensus error of the local iterates together with an additional $O(1)$ term. For the third challenge, i.e., controlling the gradient discrepancy error, they apply randomized smoothing (Duchi et al., 2012b) to the original objective functions and exploit the smoothness of the resulting surrogate functions together with the consensus bound for the random interpolation points. As a consequence of these consensus properties and the $\sqrt{d}$ factor in the smoothness constant, ME-DOL achieves only $O(n\delta^{-1}\epsilon^{-3}(n\gamma^{-2} + dn^{1/2}\gamma^{-1}))$ sample complexity and $O(\delta^{-1}\epsilon^{-3}(n\gamma^{-2} + dn^{1/2}\gamma^{-1}))$ communication complexity. Note that these bounds do not exhibit a clear advantage over those of DGFM$^+$, which operates in the zeroth-order setting.

To address this limitation, Chen et al. (2026) propose an improved algorithm called decentralized online-to-nonconvex conversion with client sampling (DOC^{2}S). Their key idea is to replace the sin-

gle standard gossip step (Xiao & Boyd, 2004) used in ME-DOL with multiple accelerated gossip steps (Liu & Morse, 2011), which provide sharper consensus guarantees with respect to the dependence on n and γ. Moreover, they incorporate a client sampling technique (Chen et al., 2022) that reduces the sample complexity by a factor of n, since only one node queries a local stochastic gradient at each iteration. Building on these two improvements, DOC^{2}S achieves $O(\delta^{-1}(\epsilon^{-3}+d\epsilon^{-1}))$ sample complexity and $\widetilde{O}(\gamma^{-1/2}\delta^{-1}(\epsilon^{-3}+d\epsilon^{-1}))$ communication complexity, thereby exhibiting a clear benefit from access to first-order information. Nonetheless, the polynomial dependence on d remains unsatisfactory.

3 PRELIMINARIES

In this section, we introduce the necessary preliminaries including the problem setup, randomized smoothing, and D-OCO.

3.1 PROBLEM SETUP

Following previous studies (Sahinoglu & Shahrampour, 2024; Chen et al., 2026), we consider a decentralized network whose communication topology is modeled by an undirected graph $\mathcal{G} = ([n], E)$, where $[n] = \{1, 2, \ldots, n\}$ denotes the set of nodes and $E \subseteq [n] \times [n]$ denotes the set of edges. Each communication round among the nodes occurs via a gossip step (Xiao & Boyd, 2004), i.e., computing a weighted average of some local variables according to a weight matrix $P \in \mathbb{R}^{n\times n}$. It is common to impose the following assumption on P, and the spectral gap of P is then given by $\gamma = 1 - \sigma_2(P) \in (0, 1]$, where $\sigma_2(P)$ denotes the second largest singular value of P.

Assumption 1. *The communication matrix $P \in \mathbb{R}^{n\times n}$ is supported on the graph $\mathcal{G} = ([n], E)$, and doubly stochastic, which satisfies: (i) $P_{ij} > 0$ only if $(i, j) \in E$ or $i = j$; (ii) $\sum_{j=1}^{n} P_{ij} = 1, \forall i \in [n]$, and $\sum_{i=1}^{n} P_{ij} = 1, \forall j \in [n]$. Moreover, P is symmetric and positive semidefinite, and $\sigma_2(P) < 1$.*

Besides, there are also several standard assumptions on the nonsmooth nonconvex objectives.

Assumption 2. *The global objective $f(\mathbf{x})$ is lower bounded, i.e., $f^\star = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$.*

Assumption 3. *For each node $i \in [n]$, the stochastic component $F_i(\mathbf{x}; \xi_i)$ is $L(\xi_i)$-Lipschitz, i.e., $|F_i(\mathbf{x}; \xi_i) - F_i(\mathbf{y}; \xi_i)| \le L(\xi_i)\|\mathbf{x} - \mathbf{y}\|, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Moreover, the random Lipschitz constant has a bounded second moment, i.e., there exists $L > 0$ such that $\mathbb{E}_{\xi_i}[L(\xi_i)^2] \le L^2, \forall\, i \in [n]$.*

Assumption 4. *Each node $i \in [n]$ has access to a local stochastic gradient oracle that returns $\nabla F_i(\mathbf{x}; \xi_i)$ for a given $\mathbf{x}$. The oracle is unbiased, i.e., $\mathbb{E}_{\xi_i}[\nabla F_i(\mathbf{x}; \xi_i)] = \nabla f_i(\mathbf{x})$. Moreover, there exists a constant $G > 0$ such that $\mathbb{E}_{\xi_i}\left[\|\nabla F_i(\mathbf{x}; \xi_i)\|^2\right] \le G^2$.*

Remark. Assumption 3 directly implies that each local objective $f_i(\mathbf{x})$ is L-Lipschitz, which will be used in our analysis. Moreover, the sample-wise Lipschitz assumption on $F_i(\mathbf{x}; \xi_i)$ ensures that it is differentiable almost everywhere, so that the stochastic gradient oracle in Assumption 4 is well-defined almost everywhere.

Under these assumptions, our goal is to find an (δ, ϵ)-Goldstein stationary point (Zhang et al., 2020) of the global objective $f(\mathbf{x})$. To be precise, for a Lipschitz function $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$, let $\partial f(\mathbf{x})$ denote its Clarke subdifferential (Clarke, 1990). The Goldstein δ-subdifferential (Goldstein, 1977) is defined as $\partial_\delta f(\mathbf{x}) = \mathrm{conv}(\cup_{\|\mathbf{y}-\mathbf{x}\|\le\delta}\partial f(\mathbf{y}))$, where $\mathrm{conv}(\cdot)$ denotes the convex hull of the given set. Then, the (δ, ϵ)-Goldstein stationary point can be formally defined as below.

Definition 1. *A point $\mathbf{x} \in \mathbb{R}^d$ is called (δ, ϵ)-Goldstein stationary for $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ if $\|\nabla f(\mathbf{x})\|_\delta \le \epsilon$, where $\|\nabla f(\mathbf{x})\|_\delta = \min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x})\}$.*

3.2 RANDOMIZED SMOOTHING

Randomized smoothing is a classical technique for constructing a smooth surrogate of a nonsmooth function (Duchi et al., 2012b). Let $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ denote the function and let $\mathcal{B}^d = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| \le 1\}$ denote the unit Euclidean ball centered at the origin of $\mathbb{R}^d$. For any smoothing radius $\delta > 0$, the smoothed version of $f(\mathbf{x})$ is defined as

$$f_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathrm{Unif}(\mathcal{B}^d)}[f(\mathbf{x} + \delta\mathbf{u})] \tag{1}$$

where $\mathrm{Unif}(\mathcal{B}^d)$ denotes the uniform distribution on $\mathcal{B}^d$. As summarized by Lin et al. (2022), the smoothed function has the following basic properties.

Lemma 1 (Proposition 2.3 of Lin et al. (2022)). *Suppose that $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz. Then, for any $\delta > 0$, its smoothed version defined in* (1) *satisfies: (i) $|f_\delta(\mathbf{x}) - f(\mathbf{x})| \leq \delta L$ for all $\mathbf{x} \in \mathbb{R}^d$; (ii) $f_\delta(\mathbf{x})$ is differentiable and L-Lipschitz; (iii) $f_\delta(\mathbf{x})$ has $(c\sqrt{d}L\delta^{-1})$-Lipschitz gradients for some constant $c > 0$.*

Lemma 1 allows us to convert the original nonsmooth problem into a smooth one and thereby exploit algorithmic and analytical techniques developed for smooth optimization. However, this alone only yields stationarity guarantees for the smoothed surrogate. To transfer such guarantees back to the original nonsmooth objective, an additional connection between their stationarity measures is required. Fortunately, Kornowski & Shamir (2024) have established the following lemma.

Lemma 2 (Lemma 4 of Kornowski & Shamir (2024)). *Suppose that $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz. Then, for any $\delta, r \geq 0$ and $\mathbf{x} \in \mathbb{R}^d$, its smoothed version defined in* (1)*, with the convention $f_0 = f$, satisfies $\partial_r f_\delta(\mathbf{x}) \subseteq \partial_{r+\delta} f(\mathbf{x})$.*

As discussed in Sahinoglu & Shahrampour (2024), for any $\delta > 0$ and $a \in (0, 1)$, Lemma 2 can be simply used to derive that

$$\|\nabla f(\mathbf{x})\|_\delta \leq \|\nabla f_{a\delta}(\mathbf{x})\|_{(1-a)\delta}, \forall \mathbf{x} \in \mathbb{R}^d. \tag{2}$$

Consequently, any $((1-a)\delta, \epsilon)$-Goldstein stationary point of $f_{a\delta}(\cdot)$ is also an (δ, ϵ)-Goldstein stationary point of $f(\cdot)$. For simplicity, we set $a = 1/2$ throughout the paper. Therefore, it suffices to find an $(\delta/2, \epsilon)$-Goldstein stationary point of the $(\delta/2)$-smoothed objective.

3.3 Decentralized Online Convex Optimization (D-OCO)

D-OCO is formulated as a collaborative game between n decentralized nodes and an adversary, and has been extensively studied in the literature (Yan et al., 2013; Shahrampour & Jadbabaie, 2018; Li et al., 2023; Wan et al., 2024; 2025). Specifically, at each round $t \in [T]$, each node i first selects a decision $\mathbf{x}_{t,i} \in \mathcal{K}$, where $\mathcal{K} \subseteq \mathbb{R}^d$ is a convex set, and then observes a convex local loss $\ell_{t,i} : \mathcal{K} \to \mathbb{R}$. Let $\ell_t(\mathbf{x}) = \sum_{j=1}^n \ell_{t,j}(\mathbf{x})$ denote the global loss at round t. The goal of each node i is to minimize its regret measured in terms of the global losses, i.e.,

$$\mathrm{Reg}_{T,i} = \sum_{t=1}^T \ell_t(\mathbf{x}_{t,i}) - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{x}).$$

As in DSO, each node i can communicate only with its immediate neighbors through gossip steps (Xiao & Boyd, 2004). Note that the D-OCO algorithm of Wan et al. (2024) has achieved a nearly optimal regret bound together with an explicit upper bound on the consensus error, i.e., $\|\mathbf{x}_{t,i} - \bar{\mathbf{x}}_t\|$ for all $i \in [n]$, where $\bar{\mathbf{x}}_t = (1/n)\sum_{j=1}^n \mathbf{x}_{t,j}$. Both guarantees will be used in our decentralized online-to-nonconvex conversion.

4 Our Decentralized Online-to-Nonconvex Conversion

In this section, we first introduce the procedure of our decentralized online-to-nonconvex conversion (D-ONC), and then show how to instantiate it with the D-OCO algorithm of Wan et al. (2024) to obtain dimension-free sample and communication complexities. Omitted proofs can be found in the appendix.

4.1 Detailed Procedure

To aid understanding, we begin with the special case of a single node, which can be handled directly by the online-to-nonconvex conversion of Cutkosky et al. (2023). Specifically, this conversion proceeds over K epochs, each consisting of T iterations. Let $\mathbf{x}_{T,1}^0 = \mathbf{0} \in \mathbb{R}^d$ denote an initial point. During each epoch k, it is natural to maintain the iterate $\mathbf{x}_{t,1}^k$ by setting $\mathbf{x}_{0,1}^k = \mathbf{x}_{T,1}^{k-1}$ and performing the update $\mathbf{x}_{t,1}^k = \mathbf{x}_{t-1,1}^k + \Delta_{t,1}^k$, where $\Delta_{t,1}^k$ is the update direction. Critically, Cutkosky et al.

(2023) propose to generate the update direction via an OCO algorithm. The key insight is that if $f_1(\cdot)$ is well-behaved, one can construct a stochastic gradient $\mathbf{g}_{t,1}^k = \nabla F_1(\mathbf{w}_{t,1}^k; \xi_{t,1}^k)$ such that

$$f_1(\mathbf{x}_{t,1}^k) - f_1(\mathbf{x}_{t-1,1}^k) = \mathbb{E}[\langle \mathbf{g}_{t,1}^k, \Delta_{t,1}^k \rangle] \tag{3}$$

by setting $\mathbf{w}_{t,1}^k = \mathbf{x}_{t-1,1}^k + s_{t,1}^k \Delta_{t,1}^k$, where $s_{t,1}^k \sim \mathrm{Unif}[0,1]$. This gives rise to an OCO problem: $\Delta_{t,1}^k$ should incur a small linear loss $\langle \mathbf{g}_{t,1}^k, \Delta_{t,1}^k \rangle$, while $\mathbf{g}_{t,1}^k$ is revealed only after $\Delta_{t,1}^k$ has been chosen. Thus, it is natural to generate $\{\Delta_{t,1}^k\}_{t\in[T]}$ by running an OCO algorithm on the sequence of linear losses $\{\ell_{t,1}^k(\mathbf{x}) = \langle \mathbf{g}_{t,1}^k, \mathbf{x} \rangle\}_{t\in[T]}$. In addition, following Cutkosky et al. (2023), the final solution should be $\mathbf{w}_1 \sim \mathrm{Unif}\{\mathbf{w}_1^1, \ldots, \mathbf{w}_1^K\}$, where $\mathbf{w}_1^k = (1/T)\sum_{t=1}^T \mathbf{w}_{t,1}^k$ for each epoch k.

Then, we extend the above construction to the general case with $n > 1$ nodes. The key idea is conceptually simple: each node i maintains its own local copies $\mathbf{x}_{t,i}^k$, $\Delta_{t,i}^k$, $\mathbf{w}_{t,i}^k$, and $\mathbf{g}_{t,i}^k$ of the corresponding quantities introduced above. Specifically, for any node $i \in [n]$, we initialize $\mathbf{x}_{T,i}^0 = \mathbf{0} \in \mathbb{R}^d$, and maintain the iterate $\mathbf{x}_{t,i}^k$ during each epoch $k \in [K]$ by setting $\mathbf{x}_{0,i}^k = \mathbf{x}_{T,i}^{k-1}$ and performing the update $\mathbf{x}_{t,i}^k = \mathbf{x}_{t-1,i}^k + \Delta_{t,i}^k$ with some update direction $\Delta_{t,i}^k$. Note that although the update of $\mathbf{x}_{t,i}^k$ follows the same form as in the single-node case, the generation of $\Delta_{t,i}^k$ is fundamentally different due to the mismatch between the local objective $f_i(\mathbf{x})$ and the global objective $f(\mathbf{x})$. To correct this mismatch, we first introduce a virtual global variable $\bar{\mathbf{x}}_t^k = (1/n)\sum_{j=1}^n \mathbf{x}_{t,j}^k$ and exploit the Lipschitz continuity of each local objective f_i to obtain

$$\begin{aligned} f(\bar{\mathbf{x}}_t^k) - f(\bar{\mathbf{x}}_{t-1}^k) =& \frac{1}{n}\sum_{j=1}^n \left(f_j(\mathbf{x}_{t,j}^k) - f_j(\mathbf{x}_{t-1,j}^k) + f_j(\bar{\mathbf{x}}_t^k) - f_j(\mathbf{x}_{t,j}^k) + f_j(\mathbf{x}_{t-1,j}^k) - f_j(\bar{\mathbf{x}}_{t-1}^k) \right) \\ =& \frac{1}{n}\sum_{j=1}^n \mathbb{E}[\langle \mathbf{g}_{t,j}^k, \Delta_{t,j}^k \rangle] + O\left(\max_{j\in[n]} \|\mathbf{x}_{t,j}^k - \bar{\mathbf{x}}_t^k\| + \max_{j\in[n]} \|\mathbf{x}_{t-1,j}^k - \bar{\mathbf{x}}_{t-1}^k\| \right) \end{aligned} \tag{4}$$

as long as $\mathbf{g}_{t,i}^k$ satisfies the node-wise counterpart of (3) for every $i \in [n]$ (i.e., replacing the index 1 with i). If only considering the first term on the right-hand side of (4), one can simply run an independent OCO algorithm at each node $i \in [n]$ to generate $\Delta_{t,i}^k$. However, controlling the second term on the right-hand side of (4) further requires sufficient consensus among the local iterates, which cannot be guaranteed by independent OCO algorithms. To address this issue, we define another virtual global variable $\bar{\Delta}_t^k = (1/n)\sum_{j=1}^n \Delta_{t,j}^k$, and obtain

$$\frac{1}{n}\sum_{j=1}^n \mathbb{E}[\langle \mathbf{g}_{t,j}^k, \Delta_{t,j}^k \rangle] = \frac{1}{n}\sum_{j=1}^n \mathbb{E}[\langle \mathbf{g}_{t,j}^k, \bar{\Delta}_t^k \rangle] + O\left(\max_{j\in[n]} \|\Delta_{t,j}^k - \bar{\Delta}_t^k\| \right). \tag{5}$$

This motivates us to generate $\{\Delta_{t,i}^k\}_{t\in[T],i\in[n]}$ by running a D-OCO algorithm on the sequence of local linear losses $\{\ell_{t,i}^k(\mathbf{x}) = \langle \mathbf{g}_{t,i}^k, \mathbf{x} \rangle\}_{t\in[T],i\in[n]}$. Indeed, the two terms on the right-hand side of (5) can be controlled by the regret and consensus error bounds of the D-OCO algorithm, respectively. More interestingly, by combining this observation with the update rule of $\mathbf{x}_{t,i}^k$, we notice that the second term on the right-hand side of (4) can be controlled simultaneously as long as the consensus error bound of the D-OCO algorithm is sufficiently small.

The remaining problem is how to construct the stochastic gradient $\mathbf{g}_{t,i}^k$. Inspired by the single-node case, a naive idea is to independently sample $s_{t,i}^k \sim \mathrm{Unif}[0,1]$, and set $\mathbf{w}_{t,i}^k = \mathbf{x}_{t-1,i}^k + s_{t,i}^k \Delta_{t,i}^k$ and $\mathbf{g}_{t,i}^k = \nabla F_i(\mathbf{w}_{t,i}^k; \xi_{t,i}^k)$. Moreover, the final solution $\mathbf{w}_1$ in the single-node case can be naturally extended to $\bar{\mathbf{w}} \sim \mathrm{Unif}\{\bar{\mathbf{w}}^1, \ldots, \bar{\mathbf{w}}^K\}$, where $\bar{\mathbf{w}}^k = (1/(Tn))\sum_{t=1}^T \sum_{i=1}^n \mathbf{w}_{t,i}^k$ for each epoch k. However, establishing (δ,ϵ)-Goldstein stationarity requires relating the average of local gradients to gradients of the global objective evaluated at common points across nodes. Specifically, let $\bar{\mathbf{w}}_t^k = (1/n)\sum_{i=1}^n \mathbf{w}_{t,i}^k$. In the analysis, we can directly control $(1/(nT))\sum_{t=1}^T \sum_{i=1}^n \nabla f_i(\mathbf{w}_{t,i}^k)$, whereas the relevant quantity for (δ,ϵ)-Goldstein stationarity is $(1/T)\sum_{t=1}^T \nabla f(\bar{\mathbf{w}}_t^k)$, provided that the points $\bar{\mathbf{w}}_t^k$ remain sufficiently close to the output $\bar{\mathbf{w}}^k$. These two averages can differ because the local gradients are evaluated at different points. For nonsmooth objectives, this discrepancy cannot generally be controlled by the distances between these points. We therefore apply randomized

Algorithm 1 Decentralized Online-to-Nonconvex Conversion (D-ONC)

Input: Number of epochs K, epoch length T, stationarity radius δ, and a D-OCO algorithm $\mathcal{A}$ with a decision set $\mathcal{K}$
Set $\mathbf{x}_{T,i}^0 = \mathbf{0}$ for all $i \in [n]$
for $k = 1, \ldots, K$ **do**
  Restart $\mathcal{A}$ over $\mathcal{K}$ and set $\mathbf{x}_{0,i}^k = \mathbf{x}_{T,i}^{k-1}$ for all $i \in [n]$
  for $t = 1, \ldots, T$ **do**
    Obtain $\Delta_{t,i}^k$ from $\mathcal{A}$ and set $\mathbf{x}_{t,i}^k = \mathbf{x}_{t-1,i}^k + \Delta_{t,i}^k$ for all $i \in [n]$
    Draw shared $s_t^k \sim \mathrm{Unif}[0,1]$ and set $\mathbf{w}_{t,i}^k = \mathbf{x}_{t-1,i}^k + s_t^k \Delta_{t,i}^k$ for all $i \in [n]$
    Draw $i_t^k \sim \mathrm{Unif}([n])$, $\mathbf{u}_{t,i_t^k}^k \sim \mathrm{Unif}(\mathcal{B}^d)$, and $\xi_{t,i_t^k}^k$
    Set $\mathbf{g}_{t,i_t^k}^k = n\nabla F_{i_t^k}\left(\mathbf{w}_{t,i_t^k}^k + (\delta/2)\mathbf{u}_{t,i_t^k}^k; \xi_{t,i_t^k}^k\right)$ and $\mathbf{g}_{t,i}^k = \mathbf{0}$ for all $i \neq i_t^k$
    Send $\ell_{t,i}^k(\mathbf{x}) = \langle \mathbf{g}_{t,i}^k, \mathbf{x} \rangle$ to the local node i of $\mathcal{A}$, for all $i \in [n]$
  end for
  Set $\bar{\mathbf{w}}^k = (nT)^{-1} \sum_{t=1}^T \sum_{i=1}^n \mathbf{w}_{t,i}^k$
end for
return $\bar{\mathbf{w}} \sim \mathrm{Unif}\{\bar{\mathbf{w}}^1, \ldots, \bar{\mathbf{w}}^K\}$

smoothing and carry out the analysis with the smoothed objectives. Their smoothness allows us to bound the corresponding gradient discrepancy in terms of the consensus errors $\|\mathbf{w}_{t,i}^k - \bar{\mathbf{w}}_t^k\|$. Now, a critical limitation of the previous naive idea becomes apparent: the independent samples $\{s_{t,i}^k\}_{i\in[n]}$ introduce an $O(1)$ term in the resulting consensus bound for $\{\mathbf{w}_{t,i}^k\}_{i\in[n]}$, even when the consensus errors of $\{\mathbf{x}_{t-1,i}^k\}_{i\in[n]}$ and $\{\Delta_{t,i}^k\}_{i\in[n]}$ are arbitrarily small. To address this limitation, we instead use a shared random interpolation weight $s_t^k \sim \mathrm{Unif}[0,1]$ across all nodes, and set $\mathbf{w}_{t,i}^k = \mathbf{x}_{t-1,i}^k + s_t^k \Delta_{t,i}^k$. Despite its simplicity, this change allows the consensus error of $\{\mathbf{w}_{t,i}^k\}_{i\in[n]}$ to be directly controlled by those of $\{\mathbf{x}_{t-1,i}^k\}_{i\in[n]}$ and $\{\Delta_{t,i}^k\}_{i\in[n]}$, and hence ultimately by the consensus error of the underlying D-OCO algorithm according to the above discussions. In addition, to reduce the total sample complexity, we also incorporate the client sampling technique (Chen et al., 2022; 2026), i.e., only one node $i_t^k \sim \mathrm{Unif}([n])$ is sampled at each iteration t of epoch k and queries its local stochastic gradient. To keep the unbiasedness of $\mathbf{g}_{t,i}^k$ (and recalling the randomized smoothing technique), we construct it as

$$\mathbf{g}_{t,i}^k = \begin{cases} n\nabla F_i\left(\mathbf{w}_{t,i}^k + \delta' \mathbf{u}_{t,i}^k; \xi_{t,i}^k\right), & i = i_t^k \\ \mathbf{0}, & i \neq i_t^k \end{cases}$$

where $\mathbf{u}_{t,i}^k \sim \mathrm{Unif}(\mathcal{B}^d)$ and $\delta' = \delta/2$ is the smoothing radius.

Based on the above discussions, the detailed procedure of our conversion is summarized in Algorithm 1, where the D-OCO algorithm is denoted as $\mathcal{A}$ and its decision set $\mathcal{K}$ will be specified later.

4.2 Theoretical Guarantees

To demonstrate the power of our conversion, we first establish a general reduction guarantee in terms of the regret and consensus error of the D-OCO algorithm. Let $\mathcal{R}_T$ denote a uniform deterministic upper bound on the expected regret of all nodes over all epochs, and let $\mathcal{C}_{T,K}$ denote a uniform deterministic upper bound on the consensus error of all nodes over all iterations and epochs, i.e.,

$$\mathcal{R}_T \geq \max_{i,k} \mathbb{E}\left[\sum_{t=1}^T \sum_{j=1}^n \ell_{t,j}^k(\Delta_{t,i}^k) - \min_{\mathbf{u}\in\mathcal{K}} \sum_{t=1}^T \sum_{j=1}^n \ell_{t,j}^k(\mathbf{u})\right], \quad \mathcal{C}_{T,K} \geq \max_{i,t,k} \mathbb{E}[\|\Delta_{t,i}^k - \bar{\Delta}_t^k\|]. \tag{6}$$

The final solution of our conversion satisfies the following theorem.

Theorem 1. *Suppose that Assumptions 2–4 hold. Following the definitions in* (6)*, for any $\delta > 0$, Algorithm 1 with $\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq D\}$ and $D = \delta/(2T)$ satisfies*

$$\mathbb{E}\|\nabla f(\bar{\mathbf{w}})\|_\delta \leq \frac{\mathcal{R}_T}{nDT} + \frac{G}{\sqrt{T}} + \frac{2(\Delta_f + L\delta)}{\delta K} + \frac{L(K+1)\mathcal{C}_{T,K}}{D} + \frac{c\sqrt{d}L(KT+1)\mathcal{C}_{T,K}}{\delta} \tag{7}$$

where $\Delta_f = f(\mathbf{0}) - f^\star$ and $c > 0$ is the numerical constant in Lemma 1.

Algorithm 2 AD-FTRL

Input: Decision set $\mathcal{K}$, learning rate η, gossip matrix P, gossip rounds R, and mixing ratio θ
Initialize $\mathbf{q}_{0,i} = \mathbf{h}_{0,i} = \mathbf{0}$ for every $i \in [n]$
for $t = 1, \ldots, T$ **do**
  Output $\Delta_{t,i} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{K}} \{\langle \mathbf{q}_{t-1,i}, \mathbf{x} \rangle + \|\mathbf{x}\|^2/(2\eta)\}$
  Receive $\mathbf{g}_{t,i}$ and set $\mathbf{q}_{t,i}^0 = \mathbf{q}_{t-1,i} + \mathbf{g}_{t,i}$ and $\mathbf{q}_{t,i}^{-1} = \mathbf{h}_{t-1,i} + \mathbf{g}_{t,i}$ for all $i \in [n]$
  for $r = 0, \ldots, R-1$ **do**
    Set $\mathbf{q}_{t,i}^{r+1} = (1+\theta) \sum_{j=1}^n P_{ij} \mathbf{q}_{t,j}^r - \theta \mathbf{q}_{t,i}^{r-1}$ for all $i \in [n]$
  end for
  Set $\mathbf{q}_{t,i} = \mathbf{q}_{t,i}^R$ and $\mathbf{h}_{t,i} = \mathbf{q}_{t,i}^{R-1}$ for all $i \in [n]$
end for

Remark. Although previous studies have already provided decentralized online-to-nonconvex conversions, their algorithms and analyses are limited to specific choices of D-OCO algorithms. To the best of our knowledge, Theorem 1 provides the first general decentralized online-to-nonconvex conversion that can be instantiated with any D-OCO algorithm satisfying the required regret and consensus guarantees. In fact, Theorem 1 reduces all challenges introduced by decentralization to the properties of the underlying D-OCO algorithm. More importantly, the problem dimension d only appears in the last term of the above bound and is multiplied by the consensus error bound $\mathcal{C}_{T,K}$. This allows us to remove the polynomial dependence on d by sufficiently tightening the consensus error of the underlying D-OCO algorithm.

Now, we proceed to instantiate our conversion with the D-OCO algorithm of Wan et al. (2024). It is called accelerated decentralized follow-the-regularized-leader (AD-FTRL), and its detailed procedure is summarized in Algorithm 2, where the superscript k in the output $\Delta_{t,i}$ and feedback $\mathbf{g}_{t,i}$ is omitted for brevity. The key idea of this algorithm is to approximate the cumulative average gradient via multiple accelerated gossip steps per round (Liu & Morse, 2011), i.e., by maintaining $\mathbf{q}_{t,i}$, and generate each local decision by applying the classical follow-the-regularized-leader (FTRL) algorithm (Hazan, 2016) with the approximated gradient information, i.e., Step 4 of Algorithm 2. Moreover, there exists a minor yet important difference between Algorithm 2 and the version originally provided by Wan et al. (2024). To be precise, in the standard D-OCO setting, the number of communication steps is commonly restricted to one per round. To implement accelerated gossip under this communication constraint, Wan et al. (2024) further employ a blocking update mechanism, i.e., local decisions are updated only once every several rounds. As a result, both the regret and consensus error bounds have a polynomial dependence on γ^{-1}, which is undesirable for our application. Fortunately, the DSO setting does not impose such a communication constraint. Therefore, we remove the blocking update mechanism and establish the following guarantee.

Proposition 1. *Suppose that Assumptions 1 and 4 hold. For any $\zeta > 0$, if Algorithm 2 invoked in Algorithm 1 is run with $\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq D\}$, $\eta = D/(G\sqrt{T})$ and*

$$R = \left\lceil \frac{\sqrt{2} \log(1 + 2\sqrt{14}\, n / \min\{1, \zeta\sqrt{T}\})}{(\sqrt{2}-1)\sqrt{\gamma}} \right\rceil, \quad \theta = \frac{1}{1 + \sqrt{1 - \sigma_2^2(P)}} \tag{8}$$

then the two bounds in (6) *can be chosen as $\mathcal{R}_T = 2nDG\sqrt{T}$ and $\mathcal{C}_{T,K} = \zeta D$.*

Remark. First, as discussed above, the regret and consensus error bounds of Algorithm 2 do not depend on the spectral gap γ. In contrast, it only affects the number of communication rounds in our application. Second, the consensus error of Algorithm 2 can be sufficiently tightened with only logarithmic additional communication. This is an important property that has been overlooked by previous decentralized online-to-nonconvex conversions.

By combining Theorem 1 with Proposition 1, we have the following corollary.

Corollary 1. *Under Assumptions 1–4, for any $\delta, \epsilon \in (0, 1]$, Algorithm 1 can ensure $\mathbb{E}\|\nabla f(\bar{\mathbf{w}})\|_\delta \leq \epsilon$ by setting $T = \lceil \max\{1, 64G^2\epsilon^{-2}\} \rceil$, $K = \lceil \max\{1, 8(\Delta_f + L\delta)\delta^{-1}\epsilon^{-1}\} \rceil$, and invoking Algorithm 2 with the same input as in Proposition 1, where $\zeta = \min\{1/2, \epsilon/(4L(K+1)(1+c\sqrt{d}))\}$ and $D = \delta/(2T)$. Accordingly, $O(\delta^{-1}\epsilon^{-3})$ stochastic gradient queries and $\widetilde{O}(\gamma^{-1/2}\delta^{-1}\epsilon^{-3})$ communication rounds are required in total.*

Proof. Let $A = \Delta_f + L\delta$ for brevity. By combining (7) with Proposition 1 and $D = \delta/(2T)$, we have

$$\mathbb{E}\|\nabla f(\bar{\mathbf{w}})\|_\delta \leq \frac{3G}{\sqrt{T}} + \frac{2A}{\delta K} + L(K+1)\zeta + \frac{c\sqrt{d}L(KT+1)}{2T}\zeta. \tag{9}$$

By the definitions of T and K, we always have

$$T \geq 1, \qquad T \geq \frac{64G^2}{\epsilon^2}, \qquad K \geq 1, \qquad K \geq \frac{8A}{\delta\epsilon}$$

which implies that

$$\frac{3G}{\sqrt{T}} \leq \frac{3\epsilon}{8}, \qquad \frac{2A}{\delta K} \leq \frac{\epsilon}{4}.$$

In particular, if $64G^2/\epsilon^2 \leq 1$, then $T = 1$ and $G \leq \epsilon/8$, which gives the same first bound. If $8A/(\delta\epsilon) \leq 1$, then $K = 1$ and $2A/\delta \leq \epsilon/4$, which gives the same second bound.

For the remaining terms in (9), $T \geq 1$ implies $(KT+1)/T \leq K+1$. Moreover, the definition of ζ ensures $\zeta \leq \epsilon/[4L(K+1)(1+c\sqrt{d})]$. Therefore,

$$L(K+1)\zeta + \frac{c\sqrt{d}L(KT+1)}{2T}\zeta \leq L(K+1)(1+c\sqrt{d})\zeta \leq \frac{\epsilon}{4}.$$

By combining (9) with the three bounds, we have

$$\mathbb{E}\|\nabla f(\bar{\mathbf{w}})\|_\delta \leq \frac{3\epsilon}{8} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{7\epsilon}{8} \leq \epsilon.$$

To bound the total complexities, we notice that

$$T \leq 1 + \frac{64G^2}{\epsilon^2} \leq \frac{1+64G^2}{\epsilon^2}, \qquad K \leq 1 + \frac{8A}{\delta\epsilon} \leq \frac{1+8(\Delta_f + L)}{\delta\epsilon}$$

where we used $\delta, \epsilon \in (0, 1]$. Moreover, due to $\sqrt{T} \geq 1$ and $\sqrt{T} \geq 8G/\epsilon$, we have

$$\frac{1}{\zeta\sqrt{T}} = \max\left\{\frac{2}{\sqrt{T}}, \frac{4L(K+1)(1+c\sqrt{d})}{\epsilon\sqrt{T}}\right\} \leq \max\left\{2, \frac{L(K+1)(1+c\sqrt{d})}{2G}\right\} = O\left(\frac{\sqrt{d}}{\delta\epsilon}\right).$$

By combining this inequality with (8), we have $R = O(\gamma^{-1/2}\log(1 + n\sqrt{d}\delta^{-1}\epsilon^{-1}))$.

Finally, it is easy to verify that the algorithm uses $KT = O(\delta^{-1}\epsilon^{-3})$ stochastic gradient queries and $RKT = \widetilde{O}(\gamma^{-1/2}\delta^{-1}\epsilon^{-3})$ communication rounds. □

Remark. From Corollary 1, we have established the first dimension-free $O(\delta^{-1}\epsilon^{-3})$ sample complexity and the first nearly dimension-free $\widetilde{O}(\gamma^{-1/2}\delta^{-1}\epsilon^{-3})$ communication complexity for finding an (δ, ϵ)-Goldstein stationary point of DSO with nonsmooth nonconvex objectives in expectation. Moreover, the $O(\delta^{-1}\epsilon^{-3})$ sample complexity has been proved to be optimal even in the non-distributed setting (Cutkosky et al., 2023).

5 Conclusion and Future Work

This paper proposes a decentralized online-to-nonconvex conversion for nonsmooth nonconvex stochastic optimization. Although it builds on existing techniques, including randomized smoothing and client sampling, its key property is that consensus among both the local iterates and the random interpolation points can be inherited from the update directions generated by a D-OCO algorithm, with a shared random interpolation weight ensuring consensus among the interpolation points. This yields a general reduction guarantee in which the dimension-dependent term is controlled by the consensus error of the D-OCO subroutine. By instantiating the conversion with AD-FTRL, an existing D-OCO algorithm whose accelerated gossip steps sufficiently reduce this error, we obtain $O(\delta^{-1}\epsilon^{-3})$ sample complexity and $\widetilde{O}(\gamma^{-1/2}\delta^{-1}\epsilon^{-3})$ communication complexity for finding an (δ, ϵ)-Goldstein stationary point in expectation. These results eliminate the polynomial dimension dependence in existing decentralized first-order guarantees, with only logarithmic dependence on the dimension in the communication cost. Nonetheless, several questions remain open. First, it is unclear whether our communication complexity is nearly optimal. Second, it may be possible to generalize our conversion to the zeroth-order setting and achieve similar improvements.

AI USE STATEMENT

The research problem, core proof ideas, and initial proof drafts are developed by the authors. We used generative AI tools to assist with revising and expanding these drafts, checking derivations, and polishing the writing. We reviewed all AI-assisted content and take full responsibility for the final text and mathematical claims.

REFERENCES

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 5219–5233, 2023.

Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, pp. 1–32, 2022.

Xinyan Chen, Weiguo Gao, and Luo Luo. Decentralized nonsmooth nonconvex optimization with client sampling. *arXiv:2601.19381*, 2026.

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 *Proceedings of the 40th International Conference on Machine Learning*, pp. 6643–6670, 2023.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

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 35*, pp. 6692–6703, 2022.

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57 (3):592–606, 2012a.

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

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems 31*, pp. 689–699, 2018.

A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13 (1):14–22, 1977.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2016.

Michael I. Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *Proceedings of the 36th Conference on Learning Theory*, pp. 4570–4597, 2023.

Krzysztof C. Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3478–3487, 2019.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5381–5393, 2020.

Anastasia Koloskova, Tao Lin, and Sebastian U. Stich. An improved analysis of gradient tracking for decentralized machine learning. In *Advances in Neural Information Processing Systems 34*, pp. 11422–11435, 2021.

Guy Kornowski and Ohad Shamir. On the complexity of finding small subgradients in nonsmooth optimization. *arXiv:2209.10346*, 2022.

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.

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

Xiuxian Li, Lihua Xie, and Na Li. A survey on distributed online optimization and online games. *Annual Reviews in Control*, 56(100904):1–24, 2023.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems 30*, pp. 5330–5340, 2017.

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 35*, 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 *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 17477–17486, 2024.

Ji Liu and A. Stephen Morse. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165, 2011.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7111–7123, 2021.

Szymon Majewski, Błażej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv:1805.01916*, 2018.

Aaron Mishkin and Mert Pilanci. Optimal sets and solution paths of relu networks. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 24888–24924, 2023.

S. Sundhar Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147:516–545, 2010.

Emre Sahinoglu and Shahin Shahrampour. An online optimization perspective on first-order and zero-order decentralized nonsmooth nonconvex stochastic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 43043–43059, 2024.

Shahin Shahrampour and Ali Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3):714–725, 2018.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, 208(1–2):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 *Proceedings of the 39th International Conference on Machine Learning*, pp. 21360–21379, 2022.

Yuanyu Wan, Tong Wei, Mingli Song, and Lijun Zhang. Nearly optimal regret for decentralized online convex optimization. In *Proceedings of the 37th Annual Conference on Learning Theory*, pp. 4862–4888, 2024.

Yuanyu Wan, Tong Wei, Bo Xue, Mingli Song, and Lijun Zhang. Optimal and efficient algorithms for decentralized online convex optimization. *Journal of Machine Learning Research*, 26(135): 1–43, 2025.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems and Control Letters*, 53(1):65–78, 2004.

Feng Yan, Shreyas Sundaram, S.V.N. Vishwanathan, and Yuan Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2013.

Kun Yuan, Xinmeng Huang, Yiming Chen, Xiaohan Zhang, Yingya Zhang, and Pan Pan. Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. In *Advances in Neural Information Processing Systems 35*, pp. 36382–36395, 2022.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11173–11182, 2020.

A Proof of Theorem 1

Since we have used the randomized smoothing technique, according to (1), we first define the smoothed versions of the global objective $f(\mathbf{x})$ and $\{f_i(\mathbf{x})\}_{i\in[n]}$ as follows

$$f_{i,\delta'}(\mathbf{x}) = \mathbb{E}_{\mathbf{u}\sim\mathrm{Unif}(\mathcal{B}^d)}[f_i(\mathbf{x}+\delta'\mathbf{u})], \forall i \in [n], \quad f_{\delta'}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_{i,\delta'}(\mathbf{x}) \tag{10}$$

where $\delta' = \delta/2$. By combining the L-Lipschitzness of each $f_i(\mathbf{x})$ derived from Assumption 3 with Lemma 1, we can derive three nice properties for each $f_{i,\delta'}(\mathbf{x})$. From the second property of Lemma 1, $f_{i,\delta'}(\mathbf{x})$ is differentiable and L-Lipschitz. Then, for any $i \in [n]$, $t \in [T]$, and $k \in [K]$, the fundamental theorem of calculus gives

$$\begin{aligned} f_{i,\delta'}(\mathbf{x}_{t,i}^k) - f_{i,\delta'}(\mathbf{x}_{t-1,i}^k) &= \int_0^1 \left\langle \nabla f_{i,\delta'}(\mathbf{x}_{t-1,i}^k + s(\mathbf{x}_{t,i}^k - \mathbf{x}_{t-1,i}^k)), \mathbf{x}_{t,i}^k - \mathbf{x}_{t-1,i}^k \right\rangle ds \\ &= \int_0^1 \left\langle \nabla f_{i,\delta'}(\mathbf{x}_{t-1,i}^k + s\Delta_{t,i}^k), \Delta_{t,i}^k \right\rangle ds \end{aligned} \tag{11}$$

where the second equality is due to the update rule $\mathbf{x}_{t,i}^k = \mathbf{x}_{t-1,i}^k + \Delta_{t,i}^k$.

Moreover, according to the construction of each $\mathbf{g}_{t,i}^k$ and Assumption 4, it is easy to verify that

$$\mathbb{E}[\langle \mathbf{g}_{t,i}^k, \Delta_{t,i}^k \rangle] = \mathbb{E}[\langle \nabla f_i(\mathbf{w}_{t,i}^k + \delta'\mathbf{u}_{t,i}^k), \Delta_{t,i}^k \rangle] = \mathbb{E}[\langle \nabla f_{i,\delta'}(\mathbf{x}_{t-1,i}^k + s_t^k \Delta_{t,i}^k), \Delta_{t,i}^k \rangle] \tag{12}$$

where the second equality is due to the definition in (10) and $\mathbf{w}_{t,i}^k = \mathbf{x}_{t-1,i}^k + s_t^k \Delta_{t,i}^k$. By combining (11), (12), and $s_t^k \sim \mathrm{Unif}[0,1]$, we have

$$\mathbb{E}[\langle \mathbf{g}_{t,i}^k, \Delta_{t,i}^k \rangle] = \mathbb{E}[f_{i,\delta'}(\mathbf{x}_{t,i}^k) - f_{i,\delta'}(\mathbf{x}_{t-1,i}^k)]. \tag{13}$$

Summing both sides of (13) over $t \in [T]$, we have

$$\sum_{t=1}^{T} \mathbb{E}[\langle \mathbf{g}_{t,i}^k, \Delta_{t,i}^k \rangle] = \mathbb{E}\left[f_{i,\delta'}(\mathbf{x}_{T,i}^k) - f_{i,\delta'}(\mathbf{x}_{0,i}^k)\right] = \mathbb{E}\left[f_{i,\delta'}(\mathbf{x}_{T,i}^k) - f_{i,\delta'}(\mathbf{x}_{T,i}^{k-1})\right] \tag{14}$$

where the last equality is due to $\mathbf{x}_{0,i}^k = \mathbf{x}_{T,i}^{k-1}$.

Note that the above analysis is only a natural extension of the non-distributed analysis of Cutkosky et al. (2023) to each local node. Our main technical innovation is to establish a connection between these local guarantees and the global objective. Specifically, by defining $\bar{\mathbf{x}}_t^k = (1/n)\sum_{j=1}^n \mathbf{x}_{t,j}^k$, the L-Lipschitzness of $f_{i,\delta'}(\mathbf{x})$ yields

$$\begin{aligned}
&\mathbb{E}\left[f_{\delta'}(\bar{\mathbf{x}}_T^k) - f_{\delta'}(\bar{\mathbf{x}}_T^{k-1})\right] \\
=&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(f_{i,\delta'}(\mathbf{x}_{T,i}^k) - f_{i,\delta'}(\mathbf{x}_{T,i}^{k-1}) + f_{i,\delta'}(\bar{\mathbf{x}}_T^k) - f_{i,\delta'}(\mathbf{x}_{T,i}^k) + f_{i,\delta'}(\mathbf{x}_{T,i}^{k-1}) - f_{i,\delta'}(\bar{\mathbf{x}}_T^{k-1})\right)\right] \\
\leq&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(f_{i,\delta'}(\mathbf{x}_{T,i}^k) - f_{i,\delta'}(\mathbf{x}_{T,i}^{k-1}) + L\|\bar{\mathbf{x}}_T^k - \mathbf{x}_{T,i}^k\| + L\|\bar{\mathbf{x}}_T^{k-1} - \mathbf{x}_{T,i}^{k-1}\|\right)\right] \\
=&\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \Delta_{t,i}^k\rangle\right] + \mathbb{E}\left[\frac{L}{n}\sum_{i=1}^n(\|\bar{\mathbf{x}}_T^k - \mathbf{x}_{T,i}^k\| + \|\bar{\mathbf{x}}_T^{k-1} - \mathbf{x}_{T,i}^{k-1}\|)\right]
\end{aligned} \tag{15}$$

where the last equality is due to (14).

To bound the first term on the right-hand side of (15), we make the following decomposition

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \Delta_{t,i}^k\rangle\right] = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \Delta_{t,i}^k - \mathbf{v}^k\rangle\right] + \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \mathbf{v}^k\rangle\right] \tag{16}$$

for any $\mathbf{v}^k \in \mathbb{R}^d$. With $\bar{\Delta}_t^k = (1/n)\sum_{j=1}^n \Delta_{t,j}^k$, we have

$$\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \Delta_{t,i}^k - \mathbf{v}^k\rangle\right] = \mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \bar{\Delta}_t^k - \mathbf{v}^k\rangle\right] + \mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \Delta_{t,i}^k - \bar{\Delta}_t^k\rangle\right]. \tag{17}$$

If $\mathbf{v}^k \in \mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq D\}$, we can utilize the regret bound of the D-OCO algorithm to derive that

$$\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \bar{\Delta}_t^k - \mathbf{v}^k\rangle\right] = \frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \Delta_{t,j}^k - \mathbf{v}^k\rangle\right] \leq \mathcal{R}_T \tag{18}$$

where $\mathcal{R}_T$ is defined in (6).

By further recalling $\mathcal{C}_{T,K}$ is defined in (6), we also utilize the consensus error bound of the D-OCO algorithm to derive that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \Delta_{t,i}^k - \bar{\Delta}_t^k\rangle\right] =&\mathbb{E}\left[\sum_{t=1}^T\sum_{i=1}^n\langle\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k), \Delta_{t,i}^k - \bar{\Delta}_t^k\rangle\right] \\
\leq&L\sum_{t=1}^T\sum_{i=1}^n\mathbb{E}[\|\Delta_{t,i}^k - \bar{\Delta}_t^k\|] \leq nLT\mathcal{C}_{T,K}
\end{aligned} \tag{19}$$

where the first inequality is due to the L-Lipschitzness of $f_{i,\delta'}(\mathbf{x})$. By combining (16), (17), (18), and (19), for $\mathbf{v}^k \in \mathcal{K}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \Delta_{t,i}^k\rangle\right] \leq \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\langle\mathbf{g}_{t,i}^k, \mathbf{v}^k\rangle\right] + \frac{\mathcal{R}_T}{n} + LT\mathcal{C}_{T,K} \\
=&\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\langle\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k), \mathbf{v}^k\rangle\right] + \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\left\langle\mathbf{g}_{t,i}^k - \nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k), \mathbf{v}^k\right\rangle\right] + \frac{\mathcal{R}_T}{n} + LT\mathcal{C}_{T,K}.
\end{aligned} \tag{20}$$

By choosing

$$\mathbf{v}^k = -D\frac{\sum_{t=1}^T\sum_{i=1}^n\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k)}{\|\sum_{t=1}^T\sum_{i=1}^n\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k)\|}$$

with $\mathbf{v}^k = \mathbf{0}$ if the denominator is zero, we have $\mathbf{v}^k \in \mathcal{K}$ and

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle \nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k), \mathbf{v}^k\rangle = -\frac{D}{n}\left\|\sum_{t=1}^{T}\sum_{i=1}^{n}\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k)\right\|. \tag{21}$$

Moreover, let $\mathbf{z}_t^k = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}_{t,i}^k - \nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k))$ for brevity. Due to $\mathbf{v}^k \in \mathcal{K}$, it is not hard to verify that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle \mathbf{z}_t^k, \mathbf{v}^k\rangle\right] \leq & D\mathbb{E}\left\|\sum_{t=1}^{T}\mathbf{z}_t^k\right\| \leq D\left(\mathbb{E}\left\|\sum_{t=1}^{T}\mathbf{z}_t^k\right\|^2\right)^{1/2} = D\left(\sum_{t=1}^{T}\mathbb{E}\|\mathbf{z}_t^k\|^2\right)^{1/2} \\ = & D\left(\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\sum_i \mathbf{g}_{t,i}^k\right\|^2 - \left\|\frac{1}{n}\sum_i \nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k)\right\|^2\right]\right)^{1/2} \\ \leq & D\left(\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{g}_{t,i_t^k}^k\right\|^2\right]\right)^{1/2} \leq DG\sqrt{T}\end{aligned} \tag{22}$$

where the second inequality is due to Jensen's inequality, and the last inequality is due to the construction of $\mathbf{g}_{t,i_t^k}^k$ and Assumption 4.

By combining (20), (21), and (22) with (15), we have

$$\begin{aligned}\frac{D}{n}\mathbb{E}\left\|\sum_{t=1}^{T}\sum_{i=1}^{n}\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k)\right\| \leq & \frac{\mathcal{R}_T}{n} + DG\sqrt{T} + LT\mathcal{C}_{T,K} + \mathbb{E}[f_{\delta'}(\bar{\mathbf{x}}_T^{k-1}) - f_{\delta'}(\bar{\mathbf{x}}_T^k)] \\ & + \frac{L}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{x}_{T,i}^k - \bar{\mathbf{x}}_T^k\| + \|\mathbf{x}_{T,i}^{k-1} - \bar{\mathbf{x}}_T^{k-1}\|\right].\end{aligned} \tag{23}$$

Taking the average of both sides over $k \in [K]$, we have

$$\begin{aligned}\frac{D}{nK}\sum_{k=1}^{K}\mathbb{E}\left\|\sum_{t=1}^{T}\sum_{i=1}^{n}\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k)\right\| \leq & \frac{\mathcal{R}_T}{n} + DG\sqrt{T} + LT\mathcal{C}_{T,K} + \frac{1}{K}\mathbb{E}[f_{\delta'}(\bar{\mathbf{x}}_T^0) - f_{\delta'}(\bar{\mathbf{x}}_T^K)] \\ & + \frac{L}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{x}_{T,i}^k - \bar{\mathbf{x}}_T^k\| + \|\mathbf{x}_{T,i}^{k-1} - \bar{\mathbf{x}}_T^{k-1}\|\right].\end{aligned} \tag{24}$$

Due to $\bar{\mathbf{x}}_T^0 = \mathbf{0}$ and the first property of Lemma 1, we have

$$f_{\delta'}(\bar{\mathbf{x}}_T^0) - f_{\delta'}(\bar{\mathbf{x}}_T^k) = f_{\delta'}(\mathbf{0}) - f_{\delta'}(\bar{\mathbf{x}}_T^k) \leq f(\mathbf{0}) - f(\bar{\mathbf{x}}_T^k) + 2L\delta' \leq \Delta_f + L\delta. \tag{25}$$

To bound $\|\mathbf{x}_{T,i}^k - \bar{\mathbf{x}}_T^k\|$, we exploit the consensus error bound of the D-OCO algorithm to establish the following lemma.

Lemma 3. *For every $i \in [n]$, $t \in [T]$, and $k \in [K]$, Algorithm 1 ensures*

$$\mathbb{E}\|\mathbf{x}_{t,i}^k - \bar{\mathbf{x}}_t^k\| \leq ((k-1)T + t)\mathcal{C}_{T,K}.$$

From Lemma 3, it is easy to verify that

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{x}_{T,i}^k - \bar{\mathbf{x}}_T^k\| + \|\mathbf{x}_{T,i}^{k-1} - \bar{\mathbf{x}}_T^{k-1}\|\right] \leq nT\sum_{k=1}^{K}(2k-1)\mathcal{C}_{T,K} = nTK^2\mathcal{C}_{T,K}. \tag{26}$$

By combining (24), (25), and (26), we have

$$\frac{D}{nK}\sum_{k=1}^{K}\mathbb{E}\left\|\sum_{t=1}^{T}\sum_{i=1}^{n}\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k)\right\| \leq \frac{\mathcal{R}_T}{n} + DG\sqrt{T} + LT\mathcal{C}_{T,K} + \frac{\Delta_f + L\delta}{K} + LKT\mathcal{C}_{T,K}. \tag{27}$$

Finally, we introduce the following lemma, which relates the above bound on these local gradients to the stationarity of the original objective.

Lemma 4. *Suppose that Assumption 3 holds. Following the definitions in* (6)*, for any $\delta > 0$, Algorithm 1 with $\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq D\}$ and $D = \delta/(2T)$ satisfies*

$$\mathbb{E}[\|\nabla f(\bar{\mathbf{w}})\|_\delta] \leq \frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K}\left\|\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k)\right\|\right] + \frac{c\sqrt{d}L(KT+1)\mathcal{C}_{T,K}}{\delta}$$

where $\delta' = \delta/2$, and $c > 0$ is the numerical constant in Lemma 1.

It is easy to complete this proof by combining Lemma 4 with (27).

B Proof of Lemma 3

For any epoch $k \in [K]$ and iteration $t \in [T]$, the update rule in Algorithm 1 and the definitions of $\bar{\mathbf{x}}_t^k$ and $\bar{\Delta}_t^k$ give

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{t,i}^k - \bar{\mathbf{x}}_t^k\|\right] &= \mathbb{E}\left[\left\|\mathbf{x}_{t-1,i}^k + \Delta_{t,i}^k - \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_{t-1,j}^k + \Delta_{t,j}^k)\right\|\right] \\ &= \mathbb{E}\left[\left\|(\mathbf{x}_{t-1,i}^k - \bar{\mathbf{x}}_{t-1}^k) + (\Delta_{t,i}^k - \bar{\Delta}_t^k)\right\|\right] \\ &\leq \mathbb{E}\left[\|\mathbf{x}_{t-1,i}^k - \bar{\mathbf{x}}_{t-1}^k\|\right] + \mathbb{E}\left[\|\Delta_{t,i}^k - \bar{\Delta}_t^k\|\right].\end{aligned}$$

Applying this inequality recursively within epoch k, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{t,i}^k - \bar{\mathbf{x}}_t^k\|\right] &\leq \mathbb{E}\left[\|\mathbf{x}_{0,i}^k - \bar{\mathbf{x}}_0^k\|\right] + \sum_{s=1}^{t}\mathbb{E}\left[\|\Delta_{s,i}^k - \bar{\Delta}_s^k\|\right] \\ &\leq \mathbb{E}\left[\|\mathbf{x}_{0,i}^k - \bar{\mathbf{x}}_0^k\|\right] + t\mathcal{C}_{T,K}\end{aligned} \tag{28}$$

where the last inequality is due to the definition of $\mathcal{C}_{T,K}$. For any $k \geq 2$, from (28), it is easy to verify that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{0,i}^k - \bar{\mathbf{x}}_0^k\|\right] = \mathbb{E}\left[\|\mathbf{x}_{T,i}^{k-1} - \bar{\mathbf{x}}_T^{k-1}\|\right] &\leq \mathbb{E}\left[\|\mathbf{x}_{0,i}^{k-1} - \bar{\mathbf{x}}_0^{k-1}\|\right] + T\mathcal{C}_{T,K} \\ &\leq \mathbb{E}\left[\|\mathbf{x}_{0,i}^1 - \bar{\mathbf{x}}_0^1\|\right] + (k-1)T\mathcal{C}_{T,K} = (k-1)T\mathcal{C}_{T,K}\end{aligned} \tag{29}$$

where the last equality is due to the common initialization $\mathbf{x}_{0,i}^1 = \mathbf{x}_{T,i}^0 = \mathbf{0}$ for all $i \in [n]$. Moreover, for $k = 1$, we have

$$\mathbb{E}\left[\|\mathbf{x}_{0,i}^k - \bar{\mathbf{x}}_0^k\|\right] = \mathbb{E}\left[\|\mathbf{x}_{0,i}^1 - \bar{\mathbf{x}}_0^1\|\right] = 0 = (k-1)T\mathcal{C}_{T,K} \tag{30}$$

By substituting (29) and (30) into (28), we finally have

$$\mathbb{E}\left[\|\mathbf{x}_{t,i}^k - \bar{\mathbf{x}}_t^k\|\right] \leq (k-1)T\mathcal{C}_{T,K} + t\mathcal{C}_{T,K} = ((k-1)T + t)\mathcal{C}_{T,K}.$$

C Proof of Lemma 4

From (2) derived from Lemma 2, we only need to derive an upper bound on $\mathbb{E}[\|\nabla f_{\delta'}(\bar{\mathbf{w}})\|_{\delta'}]$, where $\delta' = \delta/2$. Due to $\bar{\mathbf{w}} \sim \mathrm{Unif}\{\bar{\mathbf{w}}^1, \ldots, \bar{\mathbf{w}}^K\}$, it is easy to verify that

$$\mathbb{E}[\|\nabla f_{\delta'}(\bar{\mathbf{w}})\|_{\delta'}] = \frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K}\|\nabla f_{\delta'}(\bar{\mathbf{w}}^k)\|_{\delta'}\right]. \tag{31}$$

Let $\bar{\mathbf{w}}_t^k = (1/n)\sum_{i=1}^{n}\mathbf{w}_{t,i}^k$. To utilize (31), we need to bound $\|\bar{\mathbf{w}}^k - \bar{\mathbf{w}}_t^k\|$. Specifically, for any $1 \leq t' < t \leq T$, we have

$$\bar{\mathbf{w}}_t^k - \bar{\mathbf{w}}_{t'}^k = (\bar{\mathbf{x}}_{t-1}^k + s_t^k\bar{\Delta}_t^k) - (\bar{\mathbf{x}}_{t'-1}^k + s_{t'}^k\bar{\Delta}_{t'}^k) = (1 - s_{t'}^k)\bar{\Delta}_{t'}^k + \sum_{r=t'+1}^{t-1}\bar{\Delta}_r^k + s_t^k\bar{\Delta}_t^k \tag{32}$$

where the first equality is due to $\mathbf{w}_{t,i}^k = \mathbf{x}_{t-1,i}^k + s_t^k\Delta_{t,i}^k$ and the second one is due to $\bar{\mathbf{x}}_t^k = \bar{\mathbf{x}}_{t-1}^k + \bar{\Delta}_t^k$.

From (32), it is natural to derive that

$$\|\bar{\mathbf{w}}_t^k - \bar{\mathbf{w}}_{t'}^k\| \leq (1 - s_{t'}^k)\|\bar{\Delta}_{t'}^k\| + \sum_{r=t'+1}^{t-1} \|\bar{\Delta}_r^k\| + s_t^k\|\bar{\Delta}_t^k\| \leq (t - t' + 1)D \leq TD \tag{33}$$

where the second inequality is due to $\Delta_{t,i}^k \in \mathcal{K} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq D\}, \forall i \in [n], t \in [T], k \in [K]$.

By combining the definition of $\bar{\mathbf{w}}^k$ with (33), we have

$$\|\bar{\mathbf{w}}_t^k - \bar{\mathbf{w}}^k\| = \left\| \frac{1}{T} \sum_{t'=1}^{T} (\bar{\mathbf{w}}_t^k - \bar{\mathbf{w}}_{t'}^k) \right\| \leq \frac{1}{T} \sum_{t'=1}^{T} \|\bar{\mathbf{w}}_t^k - \bar{\mathbf{w}}_{t'}^k\| \leq TD = \delta' \tag{34}$$

where the last inequality is due to $D = \delta/(2T)$. Recall that $\|\nabla f_{\delta'}(\bar{\mathbf{w}}^k)\|_{\delta'} = \min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_{\delta'} f_{\delta'}(\bar{\mathbf{w}}^k)\}$. According to the definition of $\partial_{\delta'} f_{\delta'}(\bar{\mathbf{w}}^k)$, we only need to consider the Clarke subdifferential of $f_{\delta'}$ at $\{\bar{\mathbf{w}}_t^k\}_{t\in[T]}$. Moreover, due to the third property of Lemma 1, both $f_{i,\delta'}$ and their average $f_{\delta'}$ have $(c\sqrt{d}L/\delta')$-Lipschitz gradients for some constant $c > 0$. Therefore, the Clarke subdifferential of $f_{\delta'}$ at any point consists of its gradient alone. From (34), we have

$$\frac{1}{T} \sum_{t=1}^{T} \nabla f_{\delta'}(\bar{\mathbf{w}}_t^k) \in \text{conv}(\{\nabla f_{\delta'}(\bar{\mathbf{w}}_t^k) : t \in [T]\}) \subseteq \text{conv}\left(\bigcup_{\|\mathbf{y} - \bar{\mathbf{w}}^k\| \leq \delta'} \partial f_{\delta'}(\mathbf{y}) \right) = \partial_{\delta'} f_{\delta'}(\bar{\mathbf{w}}^k). \tag{35}$$

By combining (31) with (35), we have

$$\mathbb{E}[\|\nabla f_{\delta'}(\bar{\mathbf{w}})\|_{\delta'}] \leq \frac{1}{K} \mathbb{E}\left[\sum_{k=1}^{K} \left\| \frac{1}{T} \sum_{t=1}^{T} \nabla f_{\delta'}(\bar{\mathbf{w}}_t^k) \right\| \right]. \tag{36}$$

Moreover, for any $i \in [n]$, $t \in [T]$, and $k \in [K]$, we notice that

$$\begin{aligned} \mathbb{E}[\|\mathbf{w}_{t,i}^k - \bar{\mathbf{w}}_t^k\|] =& \mathbb{E}\left[\left\| \mathbf{x}_{t-1,i}^k + s_t^k \Delta_{t,i}^k - \frac{1}{n} \sum_{j=1}^{n} (\mathbf{x}_{t-1,j}^k + s_t^k \Delta_{t,j}^k) \right\| \right] \\ =& \mathbb{E}\left[\left\| \mathbf{x}_{t-1,i}^k - \bar{\mathbf{x}}_{t-1}^k + s_t^k \Delta_{t,i}^k - s_t^k \bar{\Delta}_t^k \right\| \right] \\ \leq& \mathbb{E}\left[\|\mathbf{x}_{t-1,i}^k - \bar{\mathbf{x}}_{t-1}^k\| \right] + \mathbb{E}\left[\|\Delta_{t,i}^k - \bar{\Delta}_t^k\| \right] \\ \leq& ((k-1)T + t)\mathcal{C}_{T,K} \end{aligned} \tag{37}$$

where the last inequality is due to Lemma 3 and the definition of $\mathcal{C}_{T,K}$.

Therefore, we have

$$\begin{aligned} &\mathbb{E}\left[\left\| \nabla f_{\delta'}(\bar{\mathbf{w}}_t^k) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k) \right\| \right] = \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \left(\nabla f_{i,\delta'}(\bar{\mathbf{w}}_t^k) - \nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k) \right) \right\| \right] \\ \leq& \frac{c\sqrt{d}L}{n\delta'} \sum_{i=1}^{n} \mathbb{E}\left[\left\| \bar{\mathbf{w}}_t^k - \mathbf{w}_{t,i}^k \right\| \right] \leq \frac{c\sqrt{d}L((k-1)T + t)\mathcal{C}_{T,K}}{\delta'} \end{aligned} \tag{38}$$

where the last inequality is due to (37).

By combining (36) with (38), we have

$$\begin{aligned} \mathbb{E}[\|\nabla f_{\delta'}(\bar{\mathbf{w}})\|_{\delta'}] \leq& \frac{1}{K} \mathbb{E}\left[\sum_{k=1}^{K} \left\| \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k) \right\| \right] + \frac{c\sqrt{d}L\mathcal{C}_{T,K}}{KT\delta'} \sum_{k=1}^{K} \sum_{t=1}^{T} ((k-1)T + t) \\ \leq& \frac{1}{K} \mathbb{E}\left[\sum_{k=1}^{K} \left\| \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \nabla f_{i,\delta'}(\mathbf{w}_{t,i}^k) \right\| \right] + \frac{c\sqrt{d}L(KT+1)\mathcal{C}_{T,K}}{\delta} \end{aligned} \tag{39}$$

Finally, we can complete this proof by combining (2) with (39).

D PROOF OF PROPOSITION 1

The original analysis of Wan et al. (2024) assumes a deterministic bound on the local gradient norms, i.e., $\|\mathbf{g}_{t,i}\|$. In our setting, Assumption 4 only provides a second-moment bound, so their guarantee cannot be applied directly. Fortunately, it is not hard to extend their analysis to our setting by using an expected bound regarding the local gradient norms.

For brevity, we fix an epoch k and omit the epoch superscript. Let $\mathbf{v}_t$ be the unscaled stochastic gradient queried at round t of the epoch k, i.e.,

$$\mathbf{v}_t = \nabla F_{i_t}(\mathbf{w}_{t,i_t} + (\delta/2)\mathbf{u}_{t,i_t}; \xi_{t,i_t}).$$

Due to client sampling, $\mathbf{g}_{t,i_t} = n\mathbf{v}_t$ and $\mathbf{g}_{t,i} = \mathbf{0}$ for $i \neq i_t$, so $\bar{\mathbf{g}}_t = (1/n)\sum_{i=1}^n \mathbf{g}_{t,i} = \mathbf{v}_t$. From Assumption 4, it is easy to verify that

$$\mathbb{E}\|\bar{\mathbf{g}}_t\|^2 = \mathbb{E}\|\mathbf{v}_t\|^2 \leq G^2, \quad \mathbb{E}\|\mathbf{g}_{t,i}\|^2 = n^2\mathbb{E}[\mathbf{1}_{\{i_t=i\}}\|\mathbf{v}_t\|^2] \leq nG^2, i \in [n]. \tag{40}$$

Let $\mathbf{s}_t = \sum_{s=1}^t \bar{\mathbf{g}}_s$ with $\mathbf{s}_0 = \mathbf{0}$. Following Wan et al. (2024), we define the virtual global FTRL decision for any $t \in [T+1]$ by

$$\mathbf{y}_t = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}} \left\{ \langle \mathbf{s}_{t-1}, \mathbf{x}\rangle + \frac{\|\mathbf{x}\|^2}{2\eta} \right\} = \Pi_{\mathcal{K}}(-\eta\mathbf{s}_{t-1}),$$

where $\Pi_{\mathcal{K}}(\cdot)$ denotes Euclidean projection onto $\mathcal{K}$. In Algorithm 2, the local decision can be similarly rewritten as $\Delta_{t,i} = \Pi_{\mathcal{K}}(-\eta\mathbf{q}_{t-1,i})$. Then, the difference between the local and virtual decisions can be bounded by combining $\|\mathbf{q}_{t-1,i} - \mathbf{s}_{t-1}\|$ with the following lemma.

Lemma 5 (Lemma 5 in Duchi et al. (2012a)). *For any $\eta > 0$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, it holds that*

$$\|\Pi_{\mathcal{K}}(-\eta\mathbf{u}) - \Pi_{\mathcal{K}}(-\eta\mathbf{v})\| \leq \eta\|\mathbf{u} - \mathbf{v}\|.$$

Let $c_0 = 1 - 1/\sqrt{2}$, $\rho = 1 - c_0\sqrt{\gamma} \in (0,1)$, and $\beta = \rho^R \in (0,1)$. Following the proof of Lemma 2 in Wan et al. (2024), for any $t \geq 2$, it is not hard to verify that

$$\|\mathbf{q}_{t-1,i} - \mathbf{s}_{t-1}\| \leq \sqrt{14}\sum_{\tau=1}^{t-1} \rho^{(t-\tau)R} \sqrt{\sum_{j=1}^n \|\mathbf{g}_{\tau,j}\|^2}. \tag{41}$$

Taking expectations in (41), for any $t \geq 2$, we obtain

$$\begin{aligned}
\mathbb{E}\|\mathbf{q}_{t-1,i} - \mathbf{s}_{t-1}\| &\leq \sqrt{14}\sum_{\tau=1}^{t-1} \rho^{(t-\tau)R}\mathbb{E}\left[\left(\sum_{j=1}^n \|\mathbf{g}_{\tau,j}\|^2\right)^{1/2}\right] \\
&\leq \sqrt{14}\sum_{\tau=1}^{t-1} \rho^{(t-\tau)R}\left(\sum_{j=1}^n \mathbb{E}\|\mathbf{g}_{\tau,j}\|^2\right)^{1/2} \\
&\leq \sqrt{14}\, nG\sum_{\tau=1}^{t-1} \beta^{t-\tau} \leq \frac{\sqrt{14}\, nG\beta}{1-\beta}
\end{aligned} \tag{42}$$

where the second inequality is due to Jensen's inequality and the third one is due to (40). For $t = 1$, the same bound holds trivially since $\mathbf{q}_{0,i} = \mathbf{s}_0 = \mathbf{0}$.

By combining (42) with Lemma 5, for any $i \in [n]$ and $t \in [T]$, we have

$$\mathbb{E}\|\Delta_{t,i} - \mathbf{y}_t\| \leq \eta\mathbb{E}\|\mathbf{q}_{t-1,i} - \mathbf{s}_{t-1}\| \leq \frac{\sqrt{14}\eta nG\beta}{1-\beta}. \tag{43}$$

Due to $\bar{\Delta}_t = (1/n)\sum_{j=1}^n \Delta_{t,j}$ and (43), it is easy to verify that

$$\mathbb{E}\|\Delta_{t,i} - \bar{\Delta}_t\| \leq \mathbb{E}\|\Delta_{t,i} - \mathbf{y}_t\| + \frac{1}{n}\sum_{j=1}^n \mathbb{E}\|\Delta_{t,j} - \mathbf{y}_t\| \leq \frac{2\sqrt{14}\eta nG\beta}{1-\beta}. \tag{44}$$

By the choice of R in (8), we have

$$\beta = \rho^R \leq \exp(-c_0\sqrt{\gamma}R) \leq \frac{\min\{1, \zeta\sqrt{T}\}}{\min\{1, \zeta\sqrt{T}\} + 2\sqrt{14}\, n}. \tag{45}$$

By substituting this bound and $\eta = D/(G\sqrt{T})$ into (44), we have

$$\max_{i,t} \mathbb{E}\|\Delta_{t,i} - \bar{\Delta}_t\| \leq \frac{2\sqrt{14}\, nD}{\sqrt{T}} \frac{\beta}{1-\beta} \leq \frac{D}{\sqrt{T}} \min\{1, \zeta\sqrt{T}\} \leq \zeta D.$$

Note that such an upper bound holds for every epoch $k \in [K]$. Thus, we can choose $\mathcal{C}_{T,K} = \zeta D$, as stated in Proposition 1.

It remains to bound the regret. Let $\mathrm{Reg}_{T,i}$ denote the realized regret inside the expectation in (6), with the epoch index suppressed, i.e.,

$$\mathrm{Reg}_{T,i} = \sum_{t=1}^{T}\sum_{j=1}^{n} \langle \mathbf{g}_{t,j}, \Delta_{t,i}\rangle - \min_{\mathbf{u}\in\mathcal{K}} \sum_{t=1}^{T}\sum_{j=1}^{n} \langle \mathbf{g}_{t,j}, \mathbf{u}\rangle.$$

From the proof of Theorem 1 in Wan et al. (2024), for every $\mathbf{u} \in \mathcal{K}$, the virtual global FTRL decision ensures that

$$\sum_{t=1}^{T} \langle \bar{\mathbf{g}}_t, \mathbf{y}_t - \mathbf{u}\rangle \leq \frac{\|\mathbf{u}\|^2}{2\eta} + \eta \sum_{t=1}^{T} \|\bar{\mathbf{g}}_t\|^2 \leq \frac{D^2}{2\eta} + \eta \sum_{t=1}^{T} \|\bar{\mathbf{g}}_t\|^2. \tag{46}$$

Let $\mathbf{u}^* \in \operatorname{argmin}_{\mathbf{u}\in\mathcal{K}} \sum_{t=1}^{T}\sum_{j=1}^{n} \langle \mathbf{g}_{t,j}, \mathbf{u}\rangle$. From (46), it is not hard to verify that

$$\frac{1}{n}\mathrm{Reg}_{T,i} = \sum_{t=1}^{T} \langle \bar{\mathbf{g}}_t, \Delta_{t,i} - \mathbf{u}^*\rangle \leq \frac{D^2}{2\eta} + \eta \sum_{t=1}^{T} \|\bar{\mathbf{g}}_t\|^2 + \sum_{t=1}^{T} \langle \bar{\mathbf{g}}_t, \Delta_{t,i} - \mathbf{y}_t\rangle. \tag{47}$$

Let $B_R = \sqrt{14}\, nG\beta/(1-\beta)$ for brevity. From (40) and Jensen's inequality, we have

$$\mathbb{E}\langle \bar{\mathbf{g}}_t, \Delta_{t,i} - \mathbf{y}_t\rangle \leq G\, \mathbb{E}\|\Delta_{t,i} - \mathbf{y}_t\| \leq \eta G B_R$$

where the second inequality is due to (43).

Then, taking expectations on both sides of (47) and combining (40), it is not hard to derive that

$$\frac{1}{n}\mathbb{E}[\mathrm{Reg}_{T,i}] \leq \frac{D^2}{2\eta} + \eta T G^2 + \eta T G B_R.$$

From (45), we have

$$\frac{B_R}{G} \leq \frac{\min\{1, \zeta\sqrt{T}\}}{2}.$$

By combining the above two inequalities with $\eta = D/(G\sqrt{T})$, we have

$$\mathbb{E}[\mathrm{Reg}_{T,i}] \leq nDG\sqrt{T}\left(\frac{3}{2} + \frac{\min\{1, \zeta\sqrt{T}\}}{2}\right) \leq 2nDG\sqrt{T}$$

Note that such an upper bound holds for every epoch $k \in [K]$. Thus, we can choose $\mathcal{R}_T = 2nDG\sqrt{T}$, as stated in Proposition 1.