Autonomous Intelligent Systems

ORIGINAL ARTICLE Open Access

Distributed gradient-free and projection-free algorithm for stochastic constrained optimization

Jie Hou[1], Xianlin Zeng[1*] and Chen Chen[1*]

Abstract

Distributed stochastic zeroth-order optimization (DSZO), in which the objective function is allocated over multiple agents and the derivative of cost functions is unavailable, arises frequently in large-scale machine learning and reinforcement learning. This paper introduces a distributed stochastic algorithm for DSZO in a projection-free and gradient-free manner via the Frank-Wolfe framework and the stochastic zeroth-order oracle (SZO). Such a scheme is particularly useful in large-scale constrained optimization problems where calculating gradients or projection operators is impractical, costly, or when the objective function is not differentiable everywhere. Specifically, the proposed algorithm, enhanced by recursive momentum and gradient tracking techniques, guarantees convergence with just a single batch per iteration. This significant improvement over existing algorithms substantially lowers the computational complexity. Under mild conditions, we prove that the complexity bounds on SZO of the proposed algorithm are $\mathcal{O}(n/\epsilon^2)$ and $\mathcal{O}(n(2\frac{1}{\epsilon}))$ for convex and nonconvex cases, respectively. The efficacy of the algorithm is verified on black-box binary classification problems against several competing alternatives.

Keywords: Zeroth-order optimization, Projection-free method, Stochastic constrained optimization, Distributed optimization

1 Introduction

In recent years, distributed optimization has received a surge of interest in diverse areas, including autonomous vehicle control [16], multi-agent systems [31] and sensor networks [1], due to its significant advantages in aspects of data privacy, robustness, flexibility, and scalability. Distributed optimization minimizes a joint function through local computation and communication between agents in a network. Recently, much effort has been dedicated to the distributed stochastic setting [11, 19, 29, 30], where each agent's objective function is the expectation of a function with random variables that follow unknown distributions. Such situation widely exists in the machine learning [5, 19], multi-agent reinforcement learning [25, 27, 28], and unmanned systems [7, 31], to name a few. Most distributed algorithms for solving such problems require the explicit gradients of objective functions. However, the feedback available to agents is incomplete or noisy because of the environmental uncertainty in many practical applications. Hence, the real gradient feedback seems too strict in reality.

Zeroth-order optimization is a typical gradient-free method that has gained widespread concern due to its wide usage in many practical large-scale optimization tasks. In these tasks, the explicit gradient of the objective function is expensive or unavailable to obtain, and only function evaluations are accessible. For instance, the objective function of many big data problems in complex data generation processes cannot be clearly defined. Such situations include large-scale black-box adversarial attacks to deep networks [8], simulation-based model-

*Correspondence: xianlin.zeng@bit.edu.cn; xiaofan@bit.edu.cn
[1]National Key Laboratory of Autonomous Intelligent Unmanned Systems, School of Automation, Beijing Institute of Technology, Beijing, 100081, China

Springer

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Table 1 Complexity bounds for Stochastic Frank-Wolfe Optimization method to find an ϵ-optimal or ϵ-stationary point

Reference	Structure	SZO	LMO	Query-Size
ZSCG[4]	centralized *convex*	$\mathcal{O}(n/\epsilon^3)$	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(nK^2)$
	centralized *nonconvex*	$\mathcal{O}(n/\epsilon^4)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(nK)$
ZSAGMIU[4]	centralized *convex*	$\mathcal{O}(n/\epsilon^2)$	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(k^2nK)$
MOST-FW[3]	centralized *convex*	$\mathcal{O}(n/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(n)$
Acc-SZOFW*[14]	centralized *nonconvex*	$\mathcal{O}(n/\epsilon^3)$	$\mathcal{O}(1/\epsilon^3)$	$\mathcal{O}(n)$
DSGFF [22]	decentralized *convex*	$\mathcal{O}(n/\epsilon^3)$	–	$\mathcal{O}(1)$
	decentralized *nonconvex*	$\mathcal{O}(n^{\frac{4}{3}}/\epsilon^4)$	–	$\mathcal{O}(1)$
Our Work	*distributed convex*	$\mathcal{O}(n/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(n)$
	distributed nonconvex	$\mathcal{O}(n(2^{\frac{1}{\epsilon}}))$	$\mathcal{O}(2^{\frac{1}{\epsilon}})$	$\mathcal{O}(n)$

* The symbol n is the dimension of strategy variables. Symbols k and K denote the iteration number and the total number of iterations, respectively. Here Query-Size is the required function query size in estimating one zeroth-order gradient.

ing [20], and reinforcement learning [24], etc. Motivated by these applications, the design and analysis of zeroth-order algorithms become increasingly popular, including distributed zeroth-order algorithms [21, 32, 34, 35] and stochastic zeroth-order algorithms [33, 36]. Nevertheless, most zeroth-order algorithms, even in centralized settings, are designed for unconstrained optimization problems or depend on projection operators for constraint sets. The projection operations may encounter an undesirable computational burden and even become computationally prohibitive for some latent group Lassos [15], e.g., l_1 norm balls and nuclear norm balls.

Consequently, Frank-Wolfe (FW) method [10], aka conditional gradient method, has resurged because of its projection-free and computationally efficient nature. FW method avoids the projection step by accessing a linear minimization (LM) oracle, which can be effectively implemented, especially for some widespread structured constraints (see Table I in [15]). For instance, solving an LM problem over a nuclear norm ball only requires computing a single pair of singular vectors corresponding to the largest singular value, whereas projecting a point onto a nuclear norm ball demands a complete SVD decomposition. Recent years have witnessed extensive research on FW algorithms both in the centralized stochastic setting [2, 12, 18] and distributed deterministic setting [5, 6, 17]. Note that the aforementioned FW algorithms are all designed based on the first-order gradient, which cannot be directly applied to problems with only access to the value of objective functions.

FW method with stochastic zeroth-order oracle (SZO) has been recently investigated in both convex and nonconvex settings. Specifically, [4] put forth zeroth-order stochastic FW algorithms with complexity bounds[1] $\mathcal{O}(n/\epsilon^2)$ and $\mathcal{O}(n/\epsilon^4)$ on SZO for convex and nonconvex cases, respectively. However, the algorithms in [4] require a mini-batch size related to the total number of iterations and the dimension of the problem for guaranteeing convergence. Further, [14] relaxed conditions on batch sizes via the variance reduction technique called SPIDER, and demonstrated that the algorithm achieves a lower complexity bound $\mathcal{O}(n/\epsilon^3)$ on SZO for the nonconvex setting. For the convex case, [3] put forth a stochastic zeroth-order FW method, which only requires a single batch per iteration by using a momentum-based gradient tracking technique, and obtained a complexity bound $\mathcal{O}(n/\epsilon^2)$ on SZO. Subsequently, [22] further extended the centralized stochastic zeroth-order FW methods to the decentralized setting, which depends on a central coordinator, and derived that the proposed algorithm has complexity bounds $\mathcal{O}(n/\epsilon^3)$ and $\mathcal{O}(n^{\frac{4}{3}}/\epsilon^4)$ on SZO for convex and nonconvex cases, respectively. Unfortunately, there are no efficient existing zeroth-order FW works for solving distributed stochastic optimization (DSO) problems in convex or nonconvex settings.

Motivated by the above discussions, this paper dedicates to designing a novel distributed projection-free and gradient-free algorithm for DSO problems. We provide rigorous theoretical analysis on the convergence rate and complexity guarantee of the proposed algorithm, which enjoys a convergence rate comparable to centralized stochastic first-order optimization algorithms [13], filling the theoretical gap of zeroth-order FW methods in DSO problems. Table 1 provides a comparison of the algorithms proposed in the context. The following is the main contributions of our work.

- We put forth a Distributed Stochastic Zeroth-Order Frank-Wolfe algorithm (DSZO-FW) by using the gradient tracking technique, the momentum-based variance reduction technique, and the coordinate-wise gradient estimation. To our best knowledge, DSZO-FW is the first zeroth-order FW algorithm for DSO problems.

[1] The following results are normalized to find an ϵ-optimal solution for convex optimization problems and ϵ-stationary point for nonconvex optimization problems. The symbol n denotes the dimension of the strategy variable.

- We derive sufficient conditions to guarantee the convergence of DSZO-FW under mild conditions. Specifically, DSZO-FW converges only using one batch by introducing the recursive momentum technique [9]. We establish convergence rates of $\mathcal{O}(k^{-\frac{1}{2}})$ and $\mathcal{O}(1/\log_2(k))$ for the convex and nonconvex case, respectively. The guarantee of the convex case matches the previous best-known result of centralized stochastic optimization methods.
- For convex objective functions, we prove that DSZO-FW has a function query complexity of $\mathcal{O}(n/\epsilon^2)$ for finding an ϵ-optimal solution, which coincides with that of the existing centralized best results [3, 4], and is even smaller than that of the recent decentralized FW method in [22].
- For nonconvex objective functions, we derive that DSZO-FW has a function query complexity of $\mathcal{O}(n(2^{\frac{1}{\epsilon}}))$ for finding an ϵ-stationary point under time-decaying step sizes. In contrast, other works [4, 14, 22] for solving such problems rely on the step sizes related to the total number of iterations.

The remaining is structured as follows. We introduce the problem and the algorithm design in Sect. 2. The convergence performance and theoretical guarantees of the proposed algorithm is presented in Sect. 3. Section 4 takes several simulation experiments to validate the efficacy of the algorithm. Section 5 concludes the work. Appendix provides some technical proofs of the paper.

Notations The notations used in this paper are fairly standard. Specifically, we denote $\mathbb{R}$ as a set of real numbers, and $\mathbb{R}_+$ as a set of nonnegative real numbers. Symbols $\langle\cdot\rangle$ and $\lceil\cdot\rceil$ denote the inner product and the ceiling operation, respectively. In addition, $\mathbb{R}^p$ is the set of p-dimensional real vectors. Consider a vector $v \in \mathbb{R}^p$. We write $\|v\|_q$ for the l_q norm of v and $\|v\|$ for the Euclidean norm of v. We write $\mathbb{E}[\cdot]$ to denote the expectation operator; moreover, $\mathbb{E}[\cdot|\mathcal{F}_k]$ represents the conditional expectation on the σ-field $\mathcal{F}_k$. Finally, $W = [w_{ij}]_{N\times N}$ is the weighted adjacency matrix of a topology graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, 2, \ldots, N\}$ is a set containing of N agents, and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is a set of edges. For any $i, j \in \mathcal{N}$, if $(i, j) \in \mathcal{E}$, then $w_{ij} > 0$, otherwise $w_{ij} = 0$.

2 Problem statement and algorithm design

2.1 Problem statement

Consider a set of agents $\mathcal{N} = \{1, 2, \ldots, N\}$ over an undirected network $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is a set of edges. These agents aim to collaborate to find an optimal solution x^* of the problem

$$\min_{x\in\mathcal{X}} h(x), h(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\xi^i}\left[h_i\left(x, \xi^i\right)\right], \tag{1}$$

where $x \in \mathbb{R}^n$ is the strategy variable, and $\mathcal{X} \subseteq \mathbb{R}^n$ is a compact and convex set. The function $H_i(x) := \mathbb{E}_{\xi^i}[h_i(x, \xi^i)]$ is a local objective function, and $h_i : \mathcal{X} \times \mathbb{R}^p \to \mathbb{R}$ is a function involving random variable ξ^i with an unknown distribution. The randomness ξ^i can be viewed as a random sample inserted by algorithms or as measurement noise inherent in systems. *Here, we assume that the gradient of the objective function $H_i(\cdot)$ is expensive or infeasible to obtain* and agent $i \in \mathcal{N}$ is only able to access a stochastic approximation of the real objective value $h_i(x, \xi^i)$ for any given x and ξ^i.

2.2 Algorithm design

We propose a Distributed Stochastic Zeroth-Order Frank-Wolfe algorithm (DSZO-FW), which is summarized in Algorithm 1. To measure the convergence performance of DSZO-FW, we introduce the following two oracle complexities and a performance measure.

Algorithm 1 DSZO-FW

Input: initial conditions $x_1^i \in \mathcal{X}$, and $y_1^i = s_1^i = \hat{\nabla} h_i(\bar{x}_1^i, \xi_1^i)$ for $\forall i \in \mathcal{N}$.

1: **for all** $k = 1, 2, \ldots, K$ **do**

2: Approximate the average iterate

$$\bar{x}_k^i = \sum_{j\in\mathcal{N}_i} w_{ij} x_k^j. \tag{2}$$

3: Approximate the local gradient

$$g_k^i = (1-\beta_k) g_{k-1}^i + \hat{\nabla} h_i\left(\bar{x}_k^i, \xi_k^i\right) - (1-\beta_k)\hat{\nabla} h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right), \tag{3}$$

where $\beta_k \in (0, 1]$, and $\hat{\nabla} h_i$ is defined in (9).

4: Approximate the global gradient

$$y_k^i = \sum_{j\in\mathcal{N}_i} w_{ij} y_{k-1}^j + g_k^i - g_{k-1}^i, \tag{4}$$

$$s_k^i = \sum_{j\in\mathcal{N}_i} w_{ij} y_k^j. \tag{5}$$

5: Update the iterate

$$z_k^i \in \operatorname*{argmin}_{\phi\in\mathcal{X}} \langle s_k^i, \phi\rangle, \tag{6}$$

$$x_{k+1}^i = \bar{x}_k^i + \gamma_k\left(z_k^i - \bar{x}_k^i\right), \tag{7}$$

where $\gamma_k \in (0, 1]$ is a step size.

6: **end for**

Output: x_{k+1}^i for all $i \in \mathcal{N}$.

- Stochastic Zeroth-order Oracle (**SZO**): SZO returns a function value $h_i(x, \xi^i)$ for given $x \in \mathbb{R}^n$ and $\xi^i \in \mathbb{R}^p$.
- Linear Minimization Oracle (**LMO**): LMO solves a linear optimization problem, and returns $\operatorname{argmin}_{\phi \in \mathcal{X}} \langle s, \phi \rangle$ for given direction s and constraint set $\mathcal{X}$.
- ϵ-optimal solution: Let $x^* \in \mathcal{X}$ be an optimal solution of problem (1). If $h(x) - h(x^*) \leq \epsilon$, then $x \in \mathcal{X}$ is an ϵ-optimal solution of problem (1).

Due to the unavailability of the gradient information for objective functions, agent i estimates the gradient $\nabla h_i(x^i, \xi^i)$ by using a coordinate-wise gradient estimator [3, 14]:

$$\hat{\nabla} h_i\big(x^i, \xi^i\big) = \sum_{j=1}^{n} \frac{h_i(x^i + \rho e_j, \xi^i) - h_i(x^i - \rho e_j, \xi^i)}{2\rho} e_j, \quad (8)$$

where $\rho > 0$ denotes the element-wise smoothing parameter, and $e_j \in \mathbb{R}^n$ is a standard basis vector with $[e_j]_i = 1$ if $i = j$, otherwise $[e_j]_i = 0$. We convert the estimator (8) to the following expression at an iteration k in Algorithm 1:

$$\hat{\nabla} h_i\big(x_k^i, \xi_k^i\big) = \sum_{j=1}^{n} \frac{h_i(x_k^i + \rho_k e_j, \xi_k^i) - h_i(x_k^i - \rho_k e_j, \xi_k^i)}{2\rho_k} e_j, \quad (9)$$

where $\{\rho_k\}_{k=1}^{\infty}$ is a decreasing sequence of positive real numbers.

In Algorithm 1, each agent uses SZO rather than the gradient information and mainly executes four steps. Here, we briefly introduce the process of the ith agent's kth iteration.

- Step 1: Agent i takes a weighted average of values from its neighbors on the basis of W, and uses $\bar{x}_k^i$ to approximate the average iterate. The specific description is provided in (2).
- Step 2: Agent i estimates the gradient by using the coordinate-wise gradient estimator (9). To address the non-vanishing variance caused by the gradient estimation, the paper introduces a modified momentum-based variance reduction method, aka recursive momentum [9], into the distributed stochastic Frank-Wolfe (FW) algorithm. The specific expression is described in (3).
- Step 3: Agent i approximates the global gradient by using the gradient tracking technique, which reuses the global gradient estimation y_{k-1}^i from the previous iteration via (4) and (5).
- Step 4: To avoid projection operations, agent i updates the iterate by firstly solving a linear minimization problem (6) to obtain a conditional gradient z_k^i, and then makes a convex combination with the average iterate approximation $\bar{x}_k^i$ in (7).

Remark 1 The employment of zeroth-order gradients, also known as derivative-free optimization methods, brings forth both unique challenges and potential advantages. One of the main challenges with zeroth-order methods is their high requirement of function evaluations compared to first-order methods, leading to the gradient variance and higher computational costs. To address this issue, this paper incorporates recursive momentum techniques into a gradient-tracking distributed framework to reduce the non-vanishing variance caused by the gradient estimation. Remarkably, the proposed distributed zeroth-order algorithm can not only attenuate the noise in gradient approximation by only using single batch, but also achieve a comparable function query complexity to the existing centralized best result in convex case. The most significant advantage of using zeroth-order gradients is the ability to optimize functions without the need for gradient information, making it applicable to a wider range of problems where gradients are difficult or impossible to compute.

Remark 2 In Algorithm 1, we introduce the recursive momentum technique into the distributed zeroth-order FW method for reducing the variance caused by gradient estimates, as described in (3). Specifically, we rewrite (3) as

$$g_k^i = \beta_k \hat{\nabla} h_i\big(\bar{x}_k, \xi_k^i\big) + (1 - \beta_k)\big(\hat{\nabla} h_i\big(\bar{x}_k, \xi_k^i\big) - \hat{\nabla} h_i\big(\bar{x}_{k-1}, \xi_k^i\big) + g_{k-1}^i\big). \quad (10)$$

The second term $\hat{\nabla} h_i(\bar{x}_k, \xi_k^i) - \hat{\nabla} h_i(\bar{x}_{k-1}, \xi_k^i) + g_{k-1}^i$ plays an important role in reducing variance caused by the gradient estimation. In addition, the recursive momentum technique allows Algorithm 1 to converge with only one sample at each iteration, unlike the algorithms in [4] and [22], which require large batches. Hence, Algorithm 1 is also well-competent to large-scale finite-sum optimization problems.

Remark 3 In Algorithm 1, the FW step ((6)–(7)) circumvents the projection operation by minimizing a linear optimization subproblem (6) over a constraint set $\mathcal{X}$. When constraint sets are structural constraints such as nuclear and l_1 norm balls, (6) provides an efficient implementation or even a closed-form solution [15], resulting in a cheaper computational cost compared with the projection step. For example, if $\mathcal{X}$ is an l_1 norm ball ($\mathcal{X} := \{x | \|x\|_1 \leq d\}$), the FW step allows for a closed-form solution $z_k^i = d \cdot [0, \ldots, 0, -\operatorname{sgn}[s_k^i]_h, 0, \ldots, 0]^{\mathrm{T}}$ with $h = \operatorname{argmax}_j |[s_k^i]_j|$ in Algorithm 1. Moreover, when $\mathcal{X}$ is a nuclear norm ball, solving (6) requires computing only a single pair of singular vectors corresponding to the largest singular value, whereas computing a projection onto $\mathcal{X}$ demands a complete SVD decomposition.

3 Assumptions and convergence analysis

This section dedicates to analyzing the convergence performance of Algorithm 1. Before providing main results, several standard assumptions are required.

3.1 Assumptions and facts

Assumption 1 The network $\mathcal{G}$ is connected.

Assumption 2 The weighted adjacency matrix W is doubly stochastic.

Assumptions 1 and 2 indicate that for each round of the Step 1 in Algorithm 1, the agent takes a weighted average of the values from its neighbors according to W. In addition, these assumptions [26] also imply that the matrix W's second largest eigenvalue λ satisfies $|\lambda| < 1$. The following fact is true under Assumptions 1 and 2 [26].

Fact 1 *Let* $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x^i$ *and* $\bar{x}^i = \sum_{j=1}^{N} w_{ij}x^j$. *Then,* $(\sum_{i=1}^{N} \|\bar{x}^i - \bar{x}\|^2)^{\frac{1}{2}} \leq |\lambda|(\sum_{i=1}^{N} \|x^i - \bar{x}\|^2)^{\frac{1}{2}}$.

Fact 1 suggests that each update in the average consensus process (Step 1) incrementally aligns the iteration variables more closely with their mean value $\bar{x}$. To streamline our convergence analysis, we introduce $k_0 \in \mathbb{R}_+$ as the smallest integer such that $|\lambda| \leq [k_0/(k_0+1)]^2$. Clearly, $k_0 = \lceil (|\lambda|^{-\frac{1}{2}} - 1)^{-1} \rceil$.

Assumption 3 $H_i(\cdot)$ and $h_i(\cdot, \xi^i)$ are L-smooth functions on the constraint set $\mathcal{X}$ for all $i \in \mathcal{N}$ and $\xi^i \in \mathbb{R}^p$.

Furthermore, we posit an additional assumption regarding the constraint set $\mathcal{X}$, which forms a foundational element in the context of FW-based methods [3, 4, 14, 22].

Assumption 4 $\mathcal{X}$ is compact and convex, that is, $\|x - y\| \leq d$ for all $x, y \in \mathcal{X}$, where d is a positive constant.

Assumption 5 The variance of $\nabla h_i(x, \xi^i)$ is bounded for all $x \in \mathcal{X}$ and $i \in \mathcal{N}$. That is, there exists a constant δ such that $\mathbb{E}[\|\nabla h_i(x, \xi^i) - \nabla H_i(x)\|^2] \leq \delta^2$, where $H_i(x) = \mathbb{E}[h_i(x, \xi^i)]$.

Fact 2 (see [13]) *If Assumptions 4–5 hold, there is a positive constant* l *such that* $\mathbb{E}[\|\nabla h_i(x, \xi^i)\|^2] \leq l^2$ *and* $\mathbb{E}[\|\nabla h_i(x, \xi^i)\|] \leq l$.

Assumptions 3–5 are standard assumptions in stochastic FW methods [3, 4, 9, 13, 14, 22]. If Assumption 3 holds, the following fact is true.

Fact 3 *Define* $\hat{\nabla} H_i(x^i) := \sum_{j=1}^{n} \frac{H_i(x^i + \rho e_j) - H_i(x^i - \rho e_j)}{2\rho} e_j = \mathbb{E}[\hat{\nabla} h_i(x^i, \xi^i)]$, *where* $\hat{\nabla} h_i(x^i, \xi^i)$ *defined in* (8). *Then, for any* $x^i \in \mathcal{X}$ $(i \in \mathcal{N})$ *and* $\xi^i \in \mathbb{R}^p$,

$$\left\| \hat{\nabla} h_i(x^i, \xi^i) - \nabla h_i(x^i, \xi^i) \right\|^2 \leq nL^2\rho^2, \tag{11}$$

$$\left\| \hat{\nabla} H_i(x^i) - \nabla H_i(x^i) \right\|^2 \leq nL^2\rho^2. \tag{12}$$

Proof We first prove (11). It follows from the definition of $\hat{\nabla} h_i(x^i, \xi^i)$ and the mean value theorem to $\nabla h_i(x^i, \xi^i)$ that there exists $\alpha_j \in (0, 1)$ such that

$$\begin{aligned}
&\left\| \hat{\nabla} h_i(x^i, \xi^i) - \nabla h_i(x^i, \xi^i) \right\|^2 \\
&= \left\| \sum_{j=1}^{n} \frac{h_i(x^i + \rho e_j, \xi^i) - h_i(x^i - \rho e_j, \xi^i)}{2\rho} e_j \right. \\
&\quad \left. - \nabla h_i(x^i, \xi^i) \right\|^2 \\
&= \left\| \frac{1}{2\rho} \sum_{j=1}^{n} \left(2\rho e_j e_j^{\mathrm{T}} \nabla h_i\left(x^i + (2\alpha_j - 1)\rho e_j, \xi^i\right)\right) \right. \\
&\quad \left. - \nabla h_i(x^i, \xi^i) \right\|^2.
\end{aligned}$$

It follows from the property of the basis vector e_j and Euclidean norm that

$$\begin{aligned}
&\left\| \hat{\nabla} h_i(x^i, \xi^i) - \nabla h_i(x^i, \xi^i) \right\|^2 \\
&= \sum_{j=1}^{n} \left\| e_j e_j^{\mathrm{T}} \left(\nabla h_i\left(x^i + (2\alpha_j - 1)\rho e_j, \xi^i\right) - \nabla h_i(x^i, \xi^i)\right) \right\|^2 \\
&\leq \sum_{j=1}^{n} \left\| \nabla h_i\left(x^i + (2\alpha_j - 1)\rho e_j, \xi^i\right) - \nabla h_i(x^i, \xi^i) \right\|^2 \\
&\leq L^2 \sum_{j=1}^{n} \left\| (2\alpha_j - 1)\rho e_j \right\|^2 \\
&\leq nL^2\rho^2,
\end{aligned}$$

where we use Assumption 3 in the second inequality. We obtain Eqn. (12) in a similar way. □

Fact 4 (see [13]) *For any vectors* $v_1, \ldots, v_N \in \mathbb{R}^n$,

$$\|v_1 + \cdots + v_N\|^2 \leq N\left(\|v_1\|^2 + \cdots + \|v_N\|^2\right). \tag{13}$$

Assumptions 1–5 and Facts 2–4 are crucial to the subsequent analysis. They serve as the theoretical groundwork upon which our analysis is constructed, ensuring a rigorous foundation for the methodologies employed and the conclusions drawn.

Table 2 The nomenclature of values employed in this article

Symbol	Vaule
C_1	$k_0\sqrt{N}d$
C_2	$k_0^3 4^{k_0-1}N(60L^2(d+2C_1)^2+12(2l^2+\psi_2))$
C_3	$\max\{2\|\bar{g}_1-\hat{\nabla}h(x_1)\|^2, 156L^2(d+2C_1)^2+24\delta^2\}$
C_4	$\max\{\sqrt{3}(h(\bar{x}_1)-h(x^*)), 2Ld^2+2d\sqrt{18L^2(d+2C_1)^2+12C_2+6C_3}\}$
Γ	$(L^2(d+C_1)^2+C_2+C_3)^{\frac{1}{2}}+Ld^2$
γ_k, β_k	$2/(k+2), 2/(k+1)$
ρ_k	$0<\rho_k\le(d+2C_1)/(sqrtn(k+2))$
c	$c\ge\sum_{k=1}^{2^m}(4d/(k+2)^{\frac{3}{2}})$

3.2 Convergence analysis

For the convenience of analysis, we define

$$\bar{x}_k := \frac{1}{N}\sum_{i=1}^{N} x_k^i, \qquad \bar{g}_k := \frac{1}{N}\sum_{i=1}^{N} g_k^i,$$

$$\bar{p}_k := \frac{1}{N}\sum_{i=1}^{N} \nabla H_i(\bar{x}_k^i).$$

The following lemma estimates the tracking error for the average iterate in Algorithm 1, and we provide the proof in Appendix 1.2.

Lemma 1 *Let $\gamma_k = \frac{2}{k+2}$. If Assumptions 1, 2 and 4 hold, then, for any $i\in\mathcal{N}$ and $k\ge 1$, $\|\bar{x}_k^i-\bar{x}_k\|\le\frac{2C_1}{k+2}$ and $\|\bar{x}_{k+1}^i-\bar{x}_k^i\|\le\frac{2(d+2C_1)}{k+2}$, where C_1 is defined in Table 2 and $k\ge 1$.*

Lemma 1 shows that the averaged iterate estimation $\bar{x}_k^i$ approximates to the real average value $\bar{x}_k$ at a rate of $\mathcal{O}(1/k)$.

We provide the performance of the averaged gradient tracking for Algorithm 1 in the following lemma. Appendix 1.4 presents the proof of Lemma 2.

Lemma 2 *Suppose Assumptions 1–5 hold. If $\beta_k=\frac{2}{k+1}$, $\gamma_k=\frac{2}{k+2}$ and $0<\rho_k\le\frac{d+2C_1}{\sqrt{n}(k+2)}$, then*

$$\mathbb{E}\big[\big\|\bar{g}_k-s_k^i\big\|^2\big]\le\frac{4C_2}{(k+2)^2}, \tag{14}$$

where C_2 is defined in Table 2 and $k\ge 1$.

Lemma 2 establishes that $\mathbb{E}[\|\bar{g}_k-s_k^i\|^2]=\mathcal{O}(1/k^2)$, which implies that $\|\bar{g}_k-s_k^i\|$ converges to zero as $k\to+\infty$ in expectation.

The following lemma plays an important role in the convergence analysis of Algorithm 1.

Lemma 3 *Define $\hat{\nabla}\bar{h}_k := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_k[\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)]$. If Assumptions 1–5 hold, the following two relations are established.*

1) *For any $k\ge 1$, it holds that*

$$\begin{aligned}\mathbb{E}\big[\|\bar{g}_k-\hat{\nabla}\bar{h}_k\|^2\big]&\le(1-\beta_k)^2\mathbb{E}\big[\|\bar{g}_{k-1}-\hat{\nabla}\bar{h}_{k-1}\|^2\big]+60nL^2\rho_{k-1}^2\\&\quad+6\delta^2\beta_k^2+24L^2\gamma_{k-1}^2(d+2C_1)^2.\end{aligned}\tag{15}$$

2) *If $\beta_k=\frac{2}{k+1}$, $\gamma_k=\frac{2}{k+2}$, and $\rho_k\le\frac{d+2C_1}{\sqrt{n}(k+2)}$, then for any $k\ge 1$,*

$$\mathbb{E}\big[\|\bar{g}_k-\bar{p}_k\|^2\big]\le\frac{2C_3+2L^2(d+2C_1)^2}{k+2},\tag{16}$$

where C_3 and C_1 are defined in Table 2.

The proof of Lemma 3 is provided in Appendix 1.5.

Lemma 3 shows that the variable $\bar{g}_k$ tracks the real average gradient $\bar{p}_k$ with an average error bounded by $\mathcal{O}(\frac{C_3+L^2(d+C_1)^2}{k+2})$. That is, the expected error of the approximation in stochastic gradient diminishes as the number of iterations increases. Making use of Lemmas 2 and 3, the following lemma is established.

Lemma 4 *Choose $\beta_k=\frac{2}{k+1}$, $\gamma_k=\frac{2}{k+2}$, and $0<\rho_k\le\frac{d+2C_1}{\sqrt{n}(k+2)}$. If Assumptions 1–5 hold, then, for any $k\ge 1$ and $i\in\mathcal{N}$,*

$$\begin{aligned}\mathbb{E}\big[\big\|\nabla h(\bar{x}_k)-s_k^i\big\|^2\big]&\le\frac{18L^2(d+2C_1)^2+12C_2+6C_3}{k+2}.\end{aligned}\tag{17}$$

The proof is presented in Appendix 1.6.

The following two theorems establish convergence rates of Algorithm 1 for convex and nonconvex objectives, respectively.

Theorem 1 (*Convex objective*) *Let Assumptions 1–5 hold. Choose $\beta_k=\frac{2}{k+1}$, $\gamma_k=\frac{2}{k+2}$, and $0<\rho_k\le\frac{d+2C_1}{\sqrt{n}(k+2)}$. If $h_i(\cdot,\xi^i)$ is convex for any $i\in\mathcal{N}$ and ξ^i, then*

$$\mathbb{E}\big[h(\bar{x}_{k+1})\big]-h(x^*)\le\frac{C_4}{(k+3)^{\frac{1}{2}}},\quad\forall k\ge 1,$$

where C_4 is defined in Table 2.

The proof of Theorem 1 is presented in Appendix 1.7.

Theorem 1 indicates that the convergence rate of Algorithm 1 is $\mathcal{O}(1/k^{\frac{1}{2}})$. The result can be directly translated into finding an ϵ-optimal solution to problem (1). The numbers of calls to SZO and LMO for ϵ-optimal solutions are $\mathcal{O}(\frac{nC_4^2}{\epsilon^2})$ and $\mathcal{O}(\frac{C_4^2}{\epsilon^2})$, respectively.

For the nonconvex case, we introduce a convergence criterion used for standard FW methods, aka FW-gap

[4, 13, 14, 22], which is

$$p_k = \max_{x\in\mathcal{X}} \langle \nabla h(\bar{x}_k), \bar{x}_k - x \rangle. \tag{18}$$

Based on the convergence measure (18), we establish the following theorem for problem (1) with nonconvex objective functions.

Theorem 2 (*Nonconvex objective*) *Suppose Assumptions 1–5 hold. Choose* $\beta_k = \frac{2}{k+1}$, $\gamma_k = \frac{2}{k+2}$, *and* $0 < \rho_k \le \frac{d+2C_1}{\sqrt{n}(k+2)}$. *Then,*

$$\mathbb{E}\Big[\min_{k\in\{1,\dots,K\}} p_k\Big] \le \frac{1}{\log_2(K)-1}\big(h(\bar{x}_1) - h(\bar{x}_{K+1}) + 4Ld + c\sqrt{18L^2(d+2C_1)^2 + 12C_2 + 6C_3}\big),$$

where $c \in \mathbb{R}$ *satisfies* $\sum_{k=1}^{2^m} (4d/(k+2)^{\frac{3}{2}}) \le c$.

The proof of Theorem 2 is presented in Appendix 1.8.

Theorem 2 shows that Algorithm 1 converges to a stationary point at a rate of $\mathcal{O}(1/\log_2(K))$ when the objective function is nonconvex. The total number of calls to SZO and LMO are $\mathcal{O}(2^{\frac{\Gamma}{\epsilon}} d)$ and $\mathcal{O}(2^{\frac{\Gamma}{\epsilon}})$ for finding an ϵ-stationary point, respectively.

Remark 4 Table 1 shows that both the number of calls and the function query-size to SZO of Algorithm 1 are significantly less than those in ZSCG and ZSAGMIU [4], at the cost of a larger complexity bound on LMO. In addition, Algorithm 1 has the same complexity bounds for both SZO and LMO as those in the recently proposed centralized method MOST-FW [3]. Compared with the existing decentralized zeroth-order FW method DSGFF [22], which requires a central coordinator, the fully distributed Algorithm 1 has a lower complexity bound of SZO in the convex case and a weaker dimensional dependency of SZO in the nonconvex case.

Remark 5 It is worth noting that the step sizes we use are monotone decreasing, different from the existing zeroth-order nonconvex FW methods [4, 14, 22]. The step sizes mentioned in these references depend on the total iteration number K and the dimension of the variable.

4 Numerical simulations

In this section, we apply Algorithm 1 (DSZO-FW) to solve a black-box distributed stochastic binary classification problem with convex and nonconvex objectives, respectively. To solve such problems, DSZO-FW is applied over a connected network $\mathcal{G}$ with $N = 5$ agents and a doubly stochastic adjacency matrix W. The communication graph is a ring topology, and each agent only accesses its own objective function h_i. We construct matrix W by using maximum-degree weights. Specifically, the maximum degree of ring topology is $d_{\max} = 2$. For any edge (i,j) in the graph, the weight w_{ij} is set as $w_{ij} = 1/(1 + d_{\max})$ for all $i \neq j$. The diagonal elements w_{ii} are then set to make the rows sum up to 1, which typically results in $w_{ii} = 1 - \sum_{j\in\mathcal{N}_i} w_{ij}$, where $\mathcal{N}_i$ denotes the set of neighbors of node i. We set the constraint set to an l_1-norm ball such that $\mathcal{X} = \{x \mid \|x\|_1 \le d\}$. Here we assume $d = 5$.

For better evaluating the performance of DSZO-FW, we compare it against centralized algorithms ZSCG [4], SGFFW [23], and MOST-FW [3] as baselines. In the experiments, we use three public datasets[2] (*covtype.binary*, *a9a* and *w8a*) and suppose that each iteration randomly obtains only 1% of data. Because a large batch size m_k (related to the dimension and the total number of iterations) required by ZSCG exceeds the total number of samples in these three datasets, we regard ZSCG as a deterministic algorithm in the experiment, which uses full data to compute the function value. We evaluate these four algorithms according to the FW-gap, which is defined in (18).

4.1 Black-box binary classification with convex objectives

This subsection dedicates to verifying the theoretical results of DSZO-FW in the convex case. Our goal is to find an optimal solution $x \in \mathbb{R}^n$ by solving the following stochastic binary classification problem:

$$\min_{x\in\mathcal{X}} h(x), \quad h(x) := \frac{1}{N}\sum_{i=1}^{N} h_i(x),$$

$$h_i(x) := \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbb{E}_{a_{ij},b_{ij}}\big[\ln\big(1 + \exp\big(-b_{ij}\langle a_{ij}, x\rangle\big)\big)\big],$$

where $(a_{ij}, b_{ij})_{j=1}^{m_i}$ are m_i (feature, label) pairs randomly obtained by agent i from the dataset. For benchmark, we set step sizes of these four algorithms to the same values as their theoretical results in the convex setting, i.e., $\alpha_k = 6/(k+5)$ for ZSCG [4]; $\rho_k = 4/(k+8)^{\frac{2}{3}}$, $\gamma_k = 2/(k+8)$ and $c_k = 2/(n^{\frac{1}{2}}(k+8)^{\frac{1}{3}})$ for SGFFW [23]; $\gamma_k = 1/k$, $\eta_k = 2/(k+1)$, $\mu_k = 0$ and $\rho_k = d/\sqrt{n}(k+1)$ for MOST-FW [3]; $\beta_k = 2/(k+1)$, $\gamma_k = 2/(k+2)$ and $\rho_k = d/\sqrt{n}(k+2)$ for DSZO-FW.

Figure 1 shows the convergence performance of these four algorithms on a convex binary classification problem. We observe that DSZO-FW and MOST-FW perform a smaller FW-gap than ZSCG and SGFFW, especially on dataset *w8a*, although they use less data than ZSCG. This dedicates that the local gradient estimate via the recursive momentum technique might be a better candidate for approximating the gradient. We observe the periodic vibrate on the curves of these four algorithms, especially on

[2] Available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

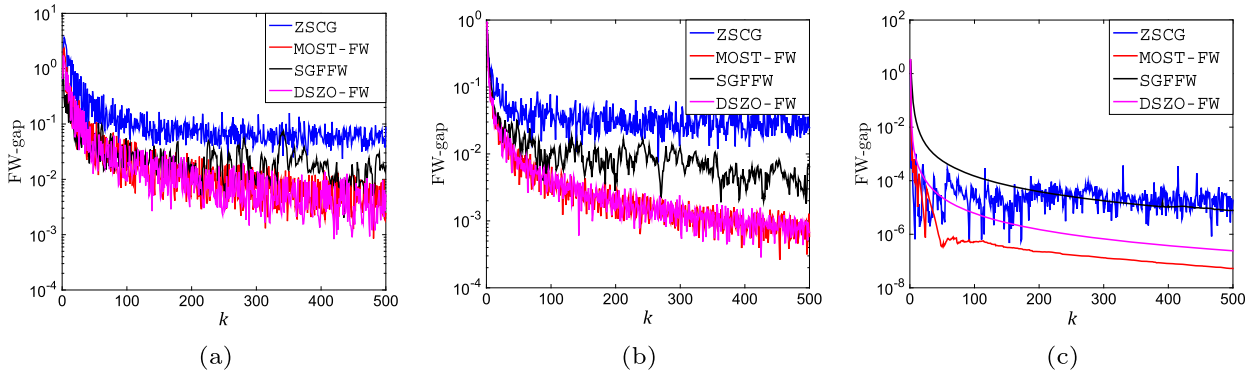

Figure 1 The comparison between ZSCG, MOST-FW, SGFFW and DSZO-FW on convex black-box binary classification. (a) *a9a* (b) *w8a* (c) *covtype.binary*

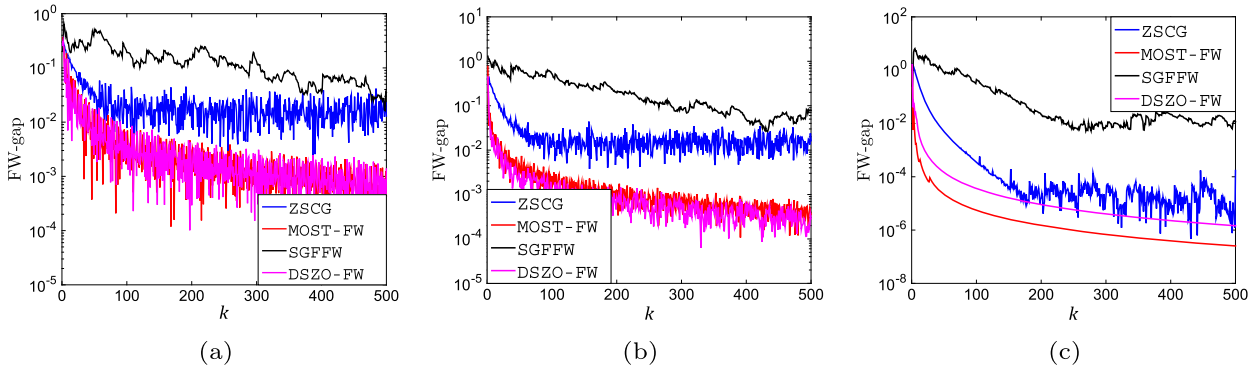

Figure 2 The comparison between ZSCG, MOST-FW, SGFFW and DSZO-FW on nonconvex black-box binary classification. (a) *a9a* (b) *w8a* (c) *covtype.binary*

datasets *a*9*a* and *w*8*a*. We intuitively believe that this phenomenon occurs due to the imprecise estimation of the gradient estimator and the gradient variance reduced period via the variance reduction technique.

4.2 Black-box binary classification with nonconvex objectives

In this subsection, we dedicate to verifying the theoretical results of DSZO-FW in the nonconvex case. Consider the following stochastic binary classification problem with nonconvex objective functions:

$$\min_{x\in\mathcal{X}} h(x), \quad h(x) := \frac{1}{N}\sum_{i=1}^{N} h_i(x),$$

$$h_i(x) := \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbb{E}_{a_{ij},b_{ij}}\left[\frac{1}{1+\exp(b_{ij}\langle a_{ij},x\rangle)}\right],$$

where $(a_{ij}, b_{ij})_{j=1}^{m_i}$ are m_i (feature, label) pairs randomly obtained by agent i from the dataset. For benchmark, we set step sizes of these four algorithms to the same values as their theoretical results in the nonconvex setting, i.e., $\alpha_k = 1/T^{\frac{1}{2}}$ for ZSCG [4]; $\gamma_k = 1/T^{\frac{3}{4}}$, $\rho_k = 4/((k+8)^{\frac{2}{3}}(1+n)^{\frac{1}{3}})$, and $c_k = 2/(n^{\frac{3}{2}}(k+8)^{\frac{1}{3}})$ for SGFFW [23]; $\gamma_k = 1/k$, $\eta_k = 2/(k+1)$, $\mu_k = 0$, and $\rho_k = d/\sqrt{n}(k+1)$ for MOST-FW [3]; $\beta_k = 2/(k+1)$, $\rho_k = d/\sqrt{n}(k+2)$, and $\gamma_k = 2/(k+2)$ for DSZO-FW. Note that MOST-FW is not proven to be convergent for the nonconvex case. We implement the algorithm only for comparison purposes.

Figure 2 shows the convergence performance measured by FW-gap of these four algorithms on a nonconvex binary classification problem. The results show that DSZO-FW converges faster than ZSCG and SGFFW in both three datasets. In contrast, DSZO-FW has a comparable convergence performance to MOST-FW on datasets *a*9*a* and *w*8*a*, demonstrating the efficacy of the variance reduction technique used in DSZO-FW and MOST-FW. Similar to Fig. 1, the periodic vibrate on the curves of these four algorithms also appears, especially on datasets *a*9*a* and *w*8*a*. We infer that this phenomenon occurs because the vari-

ance of the gradient estimator is too high in these two cases.

5 Conclusions

This paper proposed a novel algorithm in a projection-free and gradient-free manner for distributed stochastic optimization problems accessing only the stochastic zeroth-order oracle (SZO). The proposed algorithm only requires a single batch size to guarantee convergence using recursive momentum and gradient tracking techniques. We proved that the proposed algorithm has the comparable complexity bound $\mathcal{O}(n/\epsilon^2)$ on SZO as that of the centralized best results for the convex case. For the nonconvex case, the algorithm has a complexity bound $\mathcal{O}(n/(2^{\frac{1}{\epsilon}}))$ on SZO under mild conditions. The efficacy of the proposed algorithm is demonstrated through simulation experiments on multiple datasets. Our future works include extending the algorithm to stochastic nonsmooth optimization problems and introducing variance reduction techniques to obtain a better convergence performance.

Appendix

1.1 Technical lemmas for Lemma 1

We first provide some technical lemmas before proving Lemma 1.

Lemma 5 (*Lemma 2, [2]*) *Let* $\{\Pi_k\}$ *be a sequence of real numbers such that*

$$\Pi_k = \left(1 - \frac{A_1}{(k+t_0)^{a_1}}\right)\Pi_{k-1} + \frac{A_2}{(k+t_0)^{a_2}},$$

for some $a_1 \in [0,1]$ *satisfying* $a_1 \le a_2 \le 2a_1$, $A_1 > 1$ *and* $A_2 \ge 0$. *Then* Π_k *converges to zero at a rate of*

$$\Pi_k \le \frac{A}{(k+t_0+1)^{a_2-a_1}},$$

where $A = \max\{\Pi_0(t_0+1)^{a_2-a_1}, \frac{A_2}{A_1-1}\}$.

Lemma 6 *For all* $k = 1,2,\ldots,K$, *if Assumption 1 holds, we have the following relationships*:
(*a*) $\frac{1}{N}\sum_{i=1}^N y_{k+1}^i = \bar{g}_{k+1}$;
(*b*) $\bar{x}_{k+1} = (1-\gamma_k)\bar{x}_k + \gamma_k \bar{z}_k$, *where* $\bar{z}_k = \frac{1}{N}\sum_{i=1}^N z_k^i$.

Proof (a) It follows from (4) of Algorithm 1 that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N y_{k+1}^i &= \frac{1}{N}\sum_{i=1}^N\left(g_{k+1}^i - g_k^i + \sum_{j=1}^N w_{ij}y_k^j\right)\\ &= \bar{g}_{k+1} - \bar{g}_k + \frac{1}{N}\sum_{i=1}^N y_k^i\\ &= \frac{1}{N}\sum_{i=1}^N g_1^i - \bar{g}_1 + \bar{g}_{k+1} = \bar{g}_{k+1},\end{aligned}$$

where the fact that the matrix W is doubly stochastic is used in the second equality. Hence, $\frac{1}{N}\sum_{i=1}^N y_{k+1}^i = \bar{g}_{k+1}$.

(b) According to the definitions of $\bar{x}_k$ and x_k^i,

$$\begin{aligned}\bar{x}_{k+1} &= \frac{1}{N}\sum_{i=1}^N\left[\gamma_k z_k^i + (1-\gamma_k)\sum_{j=1}^N w_{ij}x_k^j\right]\\ &= \frac{(1-\gamma_k)}{N}\sum_{i=1}^N x_k^i + \frac{\gamma_k}{N}\sum_{i=1}^N z_k^i\\ &= \gamma_k\bar{z}_k + (1-\gamma_k)\bar{x}_k,\end{aligned}$$

where the fact that the matrix W is doubly stochastic is used in the first equality. The proof is completed. □

1.2 Proof of Lemma 1

Proof In the first step, we prove that $\|\bar{x}_k^i - \bar{x}_k\| \le C_1\gamma_k$.

We derive that $\|\bar{x}_k^i - \bar{x}_k\| \le \max_{i\in\mathcal{N}}\|\bar{x}_k^i - \bar{x}_k\| \le (\sum_{i=1}^N \|\bar{x}_k^i - \bar{x}_k\|^2)^{\frac{1}{2}}$ from the properties of Euclidean norm. Next, we make a proof on the following inequality by using induction on k,

$$\left(\sum_{i=1}^N\left\|\bar{x}_k^i - \bar{x}_k\right\|^2\right)^{\frac{1}{2}} \le \frac{2C_1}{k+2} = \frac{2k_0\sqrt{N}d}{k+2} = C_1\gamma_k. \tag{19}$$

It can be observed that (19) holds for $k = 1$ to $k = k_0 - 2$.

We assume that (19) holds for some $k \ge k_0 - 2$ in the induction step. It follows from Lemma 6 (b) and (7) that

$$\begin{aligned}&\sum_{i=1}^N\left\|\bar{x}_{k+1}^i - \bar{x}_{k+1}\right\|^2\\ &= \sum_{i=1}^N\Bigg\|\sum_{j=1}^N w_{ij}(1-\gamma_k)\bar{x}_k^j\\ &\quad + \sum_{j=1}^N w_{ij}\gamma_k z_k^j - (1-\gamma_k)\bar{x}_k - \gamma_k\bar{z}_k\Bigg\|^2\\ &= \sum_{i=1}^N\Bigg\|\sum_{j=1}^N w_{ij}\left[(1-\gamma_k)\sum_{h=1}^N w_{jh}x_k^h + \gamma_k z_k^j\right]\\ &\quad - \frac{1}{N}\sum_{j=1}^N\left[(1-\gamma_k)\sum_{h=1}^N w_{jh}x_k^h + \gamma_k z_k^j\right]\Bigg\|^2\\ &\le |\lambda|^2\sum_{i=1}^N\left\|(1-\gamma_k)\left(\bar{x}_k^i - \bar{x}_k\right) + \gamma_k\left(z_k^i - \bar{z}_k\right)\right\|^2,\end{aligned} \tag{20}$$

where λ is the second largest eigenvalue of W, and we use Fact 1 in the last inequality. Next, we provide an upper bound on $\sum_{i=1}^N \|(1-\gamma_k)(\bar{x}_k^i - \bar{x}_k) + \gamma_k(z_k^i - \bar{z}_k)\|^2$.

$$\begin{aligned}
&\sum_{i=1}^N \left\|\gamma_k\left(z_k^i - \bar{z}_k\right) + (1-\gamma_k)\left(\bar{x}_k^i - \bar{x}_k\right)\right\|^2 \\
&\overset{(a)}{\leq} \sum_{i=1}^N \left[\gamma_k^2 d^2 + (1-\gamma_k)^2 \left\|\bar{x}_k^i - \bar{x}_k\right\|^2 \right. \\
&\quad \left. + 2\gamma_k(1-\gamma_k) D \left\|\bar{x}_k^i - \bar{x}_k\right\|\right] \\
&\overset{(b)}{\leq} \sum_{i=1}^N \left(\left\|\gamma_k x_k^i - \bar{x}_k\right\|^2 + 2\gamma_k d \left\|\bar{x}_k^i - \bar{x}_k\right\| + \gamma_k^2 d^2\right) \\
&\overset{(c-)}{\leq} C_1^2\gamma_k^2 + n\gamma_k^2 d^2 + 2\gamma_k d\sqrt{N}\sqrt{\sum_{i=1}^N \left\|\bar{x}_k^i - \bar{x}_k\right\|^2} \\
&\leq C_1^2\gamma_k^2 + N\gamma_k^2 d^2 + 2d\gamma_k^2\sqrt{N}C_1 \\
&= \gamma_k^2\left(C_1 + C_1 k_0^{-1}\right)^2 = \left(\frac{k_0+1}{k_0} C_1\gamma_k\right)^2, \qquad (21)
\end{aligned}$$

where (a) holds because of Assumption 4; (b) is due to $1-\gamma_k \leq 1$; (c) follows from $\sum_{i=1}^N \|\bar{x}_k^i - \bar{x}_k\| \leq \sqrt{N}\sqrt{\sum_{i=1}^N \|\bar{x}_k^i - \bar{x}_k\|^2}$ and the induction hypothesis (19). Substituting (21), $\lambda \leq (\frac{k_0}{k_0+1})^2$ and $\gamma_k = \frac{2}{k+2}$ into (20), it has

$$\begin{aligned}
\sum_{i=1}^N \left\|\bar{x}_{k+1}^i - \bar{x}_{k+1}\right\|^2 &\leq \left(\frac{2(k_0+1)}{k_0(k+2)}\left(\frac{k_0}{k_0+1}\right)^2 C_1\right)^2 \\
&\leq \left(\frac{2(k+2)}{(k+2+1)(k+2)} C_1\right)^2 \\
&= C_1^2\gamma_{k+1}^2, \qquad (22)
\end{aligned}$$

where we use the monotonically increasing property of function $g(x) = x/(1+x)$ with respect to x over $[0,\infty)$ in the second inequality. We obtain $\sum_{i=1}^N \|\bar{x}_{k+1}^i - \bar{x}_{k+1}\|_2 \leq C_1\gamma_{k+1}$ by (22). That is, (19) holds for the iteration $k+1$. Hence, $\|\bar{x}_k^i - \bar{x}_k\| \leq 2C_1/(k+2)$ for all $k \geq 1$.

Next, we prove that $\|\bar{x}_{k+1}^i - \bar{x}_k^i\| \leq \frac{2(d+2C_1)}{k+2}$. From the definition of $\bar{x}_k^i$, we have

$$\begin{aligned}
&\left\|\bar{x}_{k+1}^i - \bar{x}_k^i\right\| \\
&\leq \sum_{j=1}^N w_{ij}\left(\left\|x_{k+1}^j - \bar{x}_k^j\right\| + \left\|\bar{x}_k^j - \bar{x}_k^i\right\|\right) \\
&\overset{(a)}{=} \sum_{j=1}^N w_{ij}\left(\left\|\gamma_k z_k^j - \gamma_k \bar{x}_k^j\right\| + \left\|\bar{x}_k^j - \bar{x}_k + \bar{x}_k - \bar{x}_k^i\right\|\right) \\
&\overset{(b)}{\leq} \sum_{j=1}^N w_{ij}\left(\left\|\bar{x}_k^j - \bar{x}_k\right\| + \left\|\bar{x}_k^i - \bar{x}_k\right\|\right) \\
&\quad + \sum_{j=1}^N w_{ij}\left\|\gamma_k\left(z_k^j - \bar{x}_k^j\right)\right\| \\
&\overset{(c-)}{\leq} \sum_{j=1}^N w_{ij}(\gamma_k d + 2C_1\gamma_k) = \frac{2(d+2C_1)}{k+2},
\end{aligned}$$

where (a) holds for (7); (b) is due to the triangle inequality; (c) follows from Assumption 4. □

1.3 Technique lemmas

The following two Lemmas provide the bounds of $\mathbb{E}[\|g_k^i\|]$ and $\mathbb{E}[\|g_k^i\|^2]$ in Algorithm 1.

Lemma 7 *Choose* $\beta_k = \frac{2}{k+1}$, $\gamma_k = \frac{2}{k+2}$, *and* $\rho_k \leq \frac{d+2C_1}{\sqrt{n}(k+2)}$. *If Assumptions 1–5 hold, then, for any* $k \geq 1$ *and* $i \in \mathcal{N}$,

$$\mathbb{E}\left[\left\|g_k^i\right\|\right] \leq \psi_1, \qquad (23)$$

where $\psi_1 = \max\{\|g_1^i\|, 2l + 5L(d+2C_1)\}$ *and* $C_1 = k_0\sqrt{N}d$.

Proof Obviously, (23) holds for $k=1$. We discuss the case when $k>1$ in the next step. It follows from the update (3) that

$$\begin{aligned}
&\mathbb{E}\left[\left\|g_k^i\right\|\right] \\
&= \mathbb{E}\left[\left\|(1-\beta_k)g_{k-1}^i + \hat{\nabla}h_i\left(\bar{x}_k^i, \xi_k^i\right)\right.\right. \\
&\quad \left.\left. - (1-\beta_k)\hat{\nabla}h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right)\right\|\right] \\
&\leq (1-\beta_k)\mathbb{E}\left[\left\|g_{k-1}^i\right\|\right] \\
&\quad + \beta_k\mathbb{E}\left[\left\|\hat{\nabla}h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right) - \nabla h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right)\right.\right. \\
&\quad \left.\left. + \nabla h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right)\right\|\right] \\
&\quad + \mathbb{E}\left[\left\|\hat{\nabla}h_i\left(\bar{x}_k^i, \xi_k^i\right) - \hat{\nabla}h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right)\right\|\right] \\
&\overset{(a)}{\leq} (1-\beta_k)\mathbb{E}\left[\left\|g_{k-1}^i\right\|\right] + \beta_k l + \sqrt{n}L\rho_{k-1}\beta_k \\
&\quad + \mathbb{E}\left[\left\|\hat{\nabla}h_i\left(\bar{x}_k^i, \xi_k^i\right) - \nabla h_i\left(\bar{x}_k^i, \xi_k^i\right)\right.\right. \\
&\quad + \nabla h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right) - \hat{\nabla}h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right) + \nabla h_i\left(\bar{x}_k^i, \xi_k^i\right) \\
&\quad \left.\left. - \nabla h_i\left(\bar{x}_{k-1}^i, \xi_k^i\right)\right\|\right] \\
&\overset{(b)}{\leq} (1-\beta_k)\mathbb{E}\left[\left\|g_{k-1}^i\right\|\right] + \beta_k l + \sqrt{n}L\rho_{k-1}\beta_k + \sqrt{n}L\rho_k \\
&\quad + \sqrt{n}L\rho_{k-1} + L\mathbb{E}\left[\left\|\bar{x}_k^i - \bar{x}_{k-1}^i\right\|\right] \\
&\overset{(c-)}{\leq} (1-\beta_k)\mathbb{E}\left[\left\|g_{k-1}^i\right\|\right] + \beta_k l + \sqrt{n}L\rho_{k-1} \\
&\quad + 2\sqrt{n}L\rho_{k-1} + \gamma_{k-1}L(d+2C_1),
\end{aligned}$$

where we use Facts 2 and 3 in (a); (b) holds by Fact 3 and the smoothness of h_i in Assumption 3; (c) holds because of the fact $\rho_k \le \rho_{k-1}$, $\beta_k < 1$, and Lemma 1. Taking $\beta_k = \frac{2}{k+1}$, $\gamma_k = \frac{2}{k+2}$, and $\rho_k \le \frac{d+2C_1}{\sqrt{n}(k+2)}$, we yield that

$$\mathbb{E}\big[\|g_k^i\|\big] \le \left(1 - \frac{2}{k+1}\right)\mathbb{E}\big[\|g_{k-1}^i\|\big] + \frac{2l + 5L(d+2C_1)}{k+1}.$$

Using Lemma 5 with $t_0 = 1$, $a_1 = a_2 = 1$, $A_1 = 2$, and $A_2 = 2nl + 4L(d + 2C_1)$, we yield that $\mathbb{E}[\|g_k^i\|] \le \psi_1 = \max\{\|g_1^i\|, 2l + 5L(d + 2C_1)\}$. □

Lemma 8 *Suppose Assumptions 1–5 hold. Choose* $\gamma_k = \frac{2}{k+2}$, $\beta_k = \frac{2}{k+1}$, *and* $\rho_k \le \frac{d+2C_1}{\sqrt{n}(k+2)}$. *Then, for any* $k \ge 1$ *and* $i \in \mathcal{N}$,

$$\mathbb{E}\big[\|g_k^i\|^2\big] \le \psi_2, \tag{24}$$

where $\psi_2 = \max\{\|g_1^i\|^2, 10L\psi_1(d + 2C_1) + 28L^2(d + 2C_1)^2 + 8l^2 + 4l\psi_1\}$.

Proof Obviously, (24) holds for $k = 1$. Next, we consider the case when $k > 1$. It follows from the update (3) that

$$\begin{aligned}
\|g_k^i\|^2 &\le (1-\beta_k)^2\|g_{k-1}^i\|^2 + 2(1-\beta_k)\|g_{k-1}^i\| \\
&\quad \times \big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - (1-\beta_k)\hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\| \\
&\quad + \big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - (1-\beta_k)\hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\|^2 \\
&\le (1-\beta_k)^2\|g_{k-1}^i\|^2 + 2(1-\beta_k)\|g_{k-1}^i\| \\
&\quad \times \big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - \hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i) \\
&\quad + \beta_k\hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\| + \big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) \\
&\quad - \hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i) + \beta_k\hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\|^2 \\
&\le (1-\beta_k)\|g_{k-1}^i\|^2 \\
&\quad + 2\|g_{k-1}^i\|\big(\big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - \hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\| \\
&\quad + \beta_k\big\|\hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\|\big) \\
&\quad + 2\big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - \hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\|^2 \\
&\quad + 2\beta_k^2\big\|\hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\|^2,
\end{aligned} \tag{25}$$

where we use the fact $1 - \beta_k \le 1$ and the triangle inequality in the last inequality. Next, we concentrate on the term $\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - \hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\|$ on the RHS of (25). Introducing $\nabla h_i(\bar{x}_k^i, \xi_k^i) - \nabla h_i(\bar{x}_{k-1}^i, \xi_k^i)$, we have

$$\begin{aligned}
&\big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - \hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\| \\
&= \big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - \nabla h_i(\bar{x}_k^i, \xi_k^i) + \nabla h_i(\bar{x}_{k-1}^i, \xi_k^i) \\
&\quad - \hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i) + \nabla h_i(\bar{x}_k^i, \xi_k^i) - \nabla h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\| \\
&\le \big\|\hat{\nabla}h_i(\bar{x}_k^i, \xi_k^i) - \nabla h_i(\bar{x}_k^i, \xi_k^i)\big\| + \big\|\nabla h_i(\bar{x}_{k-1}^i, \xi_k^i) \\
&\quad - \hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\| + \big\|\nabla h_i(\bar{x}_k^i, \xi_k^i) - \nabla h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\| \\
&\le \sqrt{n}L\rho_k + \sqrt{n}L\rho_{k-1} + L\big\|\bar{x}_k^i - \bar{x}_{k-1}^i\big\| \\
&\le 2\sqrt{n}L\rho_{k-1} + L(d + 2C_1)\gamma_{k-1},
\end{aligned} \tag{26}$$

where we use the triangle inequality, Fact 3, and Lemma 1 to obtain the result. Substituting (26) into (25), and taking the conditional expectation on $\mathcal{F}_k$, we therefore have that

$$\begin{aligned}
\mathbb{E}_k\big[\|g_k^i\|^2\big] &\le (1-\beta_k)\|g_{k-1}^i\|^2 \\
&\quad + \big(4\sqrt{n}L\rho_{k-1} + 2L(d+2C_1)\gamma_{k-1}\big)\|g_{k-1}^i\| \\
&\quad + 2\big(2\sqrt{n}L\rho_{k-1} + L(d+2C_1)\gamma_{k-1}\big)^2 \\
&\quad + 2\beta_k^2\mathbb{E}_k\big[\big\|\hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i) \\
&\quad - \nabla h_i(\bar{x}_{k-1}^i, \xi_k^i) + \nabla h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\|^2\big] \\
&\quad + 2\beta_k\|g_{k-1}^i\|\mathbb{E}_k\big[\big\|\hat{\nabla}h_i(\bar{x}_{k-1}^i, \xi_k^i) \\
&\quad - \nabla h_i(\bar{x}_{k-1}^i, \xi_k^i) + \nabla h_i(\bar{x}_{k-1}^i, \xi_k^i)\big\|\big] \\
&\le (1-\beta_k)\|g_{k-1}^i\|^2 \\
&\quad + \big(4\sqrt{n}L\rho_{k-1} + 2L(d+2C_1)\gamma_{k-1}\big)\|g_{k-1}^i\| \\
&\quad + 16nL^2\rho_{k-1}^2 + 4L^2(d+2C_1)^2\gamma_{k-1}^2 \\
&\quad + 4nL^2\rho_{k-1}^2\beta_k^2 + 4l^2\beta_k^2 \\
&\quad + 2\sqrt{n}L\|g_{k-1}^i\|\beta_k\rho_{k-1} + 2l\|g_{k-1}^i\|\beta_k,
\end{aligned} \tag{27}$$

where the last inequality is due to (13), Fact 2, and Fact 3. Taking the full expectation on both sides of (27) and taking $\beta_k = \frac{2}{k+1}$, $\gamma_k = \frac{2}{k+2}$, $\rho_k \le \frac{d+2C_1}{\sqrt{n}(k+2)}$, it follows from (23) that

$$\begin{aligned}
\mathbb{E}\big[\|g_k^i\|^2\big] &\le (1-\beta_k)\mathbb{E}\big[\|g_{k-1}^i\|^2\big] + 4\sqrt{n}L\psi_1\rho_{k-1} \\
&\quad + 2L(d+2C_1)\psi_1\gamma_{k-1} + 8nL^2\rho_{k-1}^2 \\
&\quad + 4L^2(d+2C_1)^2\gamma_{k-1} + 4nL^2\rho_{k-1}^2 + 4l^2\beta_k \\
&\quad + 2\sqrt{n}L\psi_1\rho_{k-1} + 2l\psi_1\beta_k \\
&\le \left(1 - \frac{2}{k+1}\right)\mathbb{E}\big[\|g_{k-1}^i\|^2\big] + \frac{10L\psi_1(d+2C_1)}{k+1} \\
&\quad + \frac{28L^2(d+2C_1)^2 + 8l^2 + 4l\psi_1}{k+1}.
\end{aligned}$$

Using Lemma 5 with $t_0 = 1$, $a_1 = a_2 = 1$, $A_1 = 2$, and $A_2 = 10L\psi_1(d + 2C_1) + 28L^2(d + 2C_1)^2 + 8l^2 + 4l\psi_1$, we prove the result. □

1.4 Proof of Lemma 2

Proof To obtain the result in (14), we prove that

$$\mathbb{E}\left[\sum_{i=1}^{N}\left\|s_k^i-\bar{g}_k\right\|^2\right]\le C_2\gamma_k^2 \tag{28}$$

by using induction on k. Firstly, we prove that (28) holds if $1\le k\le k_0-2$. It follows from the updates (4) and (5) that $s_k^i=y_{k+1}^i-g_{k+1}^i+g_k^i$. We have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\left\|s_k^i-\bar{g}_k\right\|^2\right]\\
&=\mathbb{E}\left[\sum_{i=1}^{N}\left\|y_{k+1}^i-g_{k+1}^i+g_k^i-\bar{g}_k\right\|^2\right]\\
&\le\mathbb{E}\left[\sum_{i=1}^{N}\left(4\left\|y_{k+1}^i\right\|^2+4\left\|g_{k+1}^i\right\|^2+4\left\|g_k^i\right\|^2\right.\right.\\
&\quad\left.\left.+4\left\|\frac{1}{N}\sum_{j=1}^{N}g_k^j\right\|^2\right)\right]\\
&\le 12N\psi_2+4\sum_{i=1}^{N}\mathbb{E}\left[\left\|y_{k+1}^i\right\|^2\right],
\end{aligned}\tag{29}$$

where we use (13) in the first inequality and (24) in the last inequality. Next, we focus on the second term of the RHS of (29). It follows from the update (4) that

$$\begin{aligned}
\mathbb{E}\left[\left\|y_{k+1}^i\right\|^2\right]&=\mathbb{E}\left[\left\|\sum_{j=1}^{N}w_{ij}y_k^j+g_{k+1}^i-g_k^i\right\|^2\right]\\
&\overset{(a)}{\le}\mathbb{E}\left[3\sum_{j=1}^{N}w_{ij}\left\|y_k^j\right\|^2+3\left\|g_{k+1}^i\right\|^2+3\left\|g_k^i\right\|^2\right]\\
&\overset{(b)}{\le}3\sum_{j=1}^{N}w_{ij}\mathbb{E}\left[\left\|y_k^j\right\|^2\right]+6\psi_2\\
&\overset{(c-)}{\le}3^{k-2}\left(18l^2+L^2(d+2C_1)^2\right)+6k3^{k-1}\psi_2,
\end{aligned}$$

where (a) holds because of (13) and the Jensen's inequality; (b) follows from (24); (c) is due to the fact that $\mathbb{E}[\|y_1^j\|^2]=\mathbb{E}[\|\hat{\nabla}h_i(\bar{x}_1^i,\xi_1^i)-\nabla h_i(\bar{x}_1^i,\xi_1^i)+\nabla h_i(\bar{x}_1^i,\xi_1^i)\|^2]\le 2l^2+L^2(d+2C_1)^2/9$. We therefore yield that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\left\|\bar{g}_k-s_k^i\right\|^2\right]\\
&\le 12N\psi_2+4\left(3^{k-2}\right)\left(18l^2+L^2(d+2C_1)^2\right)N\\
&\quad+24k3^{k-1}N\psi_2\\
&\le 12N\psi_2+\left(4^{k_0-3}\right)\left(18l^2+L^2(d+2C_1)^2\right)N\\
&\quad+8(k_0-2)3^{k_0-2}N\psi_2\\
&\le 12N\psi_2+4^{k_0-1}N\left(2l^2+L^2(d+2C_1)^2+2(k_0-2)\psi_2\right)
\end{aligned}$$

for $k<k_0-2$. Obviously, (28) is true when $k\in\{1,k_0-2\}$.

For induction step, we assume that (28) is true for some $k\ge k_0-2$. For convenience of analysis, we define $\Delta g_{k+1}^i:=g_{k+1}^i-g_k^i$ and $\Delta\hat{g}_{k+1}:=\bar{g}_{k+1}-\bar{g}_k$. According to the update (4), we have that $y_{k+1}^i=\Delta g_{k+1}^i+s_k^i$. Further, it follows from Fact 1 and Lemma 6 (a) that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\left\|s_{k+1}^i-\bar{g}_{k+1}\right\|^2\right]\\
&\le\mathbb{E}\left[|\lambda|^2\sum_{i=1}^{N}\left\|y_{k+1}^i-\bar{g}_{k+1}\right\|^2\right]\\
&=\mathbb{E}\left[|\lambda|^2\sum_{i=1}^{N}\left\|\Delta g_{k+1}^i+s_k^i-\bar{g}_{k+1}\right\|^2\right].
\end{aligned}\tag{30}$$

Next, we focus on the RHS of (30). It follows from the definitions of Δg_{k+1}^i and $\Delta\hat{g}_{k+1}$ that

$$\begin{aligned}
&\sum_{i=1}^{N}\left\|\Delta g_{k+1}^i+s_k^i-\bar{g}_{k+1}\right\|^2\\
&=\sum_{i=1}^{N}\left\|s_k^i-\bar{g}_k+\Delta g_{k+1}^i-\Delta\hat{g}_{k+1}\right\|^2\\
&\le\sum_{i=1}^{N}\left(\left\|s_k^i-\bar{g}_k\right\|^2+\left\|\Delta g_{k+1}^i-\Delta\hat{g}_{k+1}\right\|^2\right.\\
&\quad\left.+2\left\|s_k^i-\bar{g}_k\right\|\left\|\Delta g_{k+1}^i-\Delta\hat{g}_{k+1}\right\|\right).
\end{aligned}\tag{31}$$

Recall the definition of Δg_{k+1}^i and the update (3). We bound $\mathbb{E}[\|\Delta g_{k+1}^i\|^2]$ as

$$\begin{aligned}
&\mathbb{E}\left[\left\|\Delta g_{k+1}^i\right\|^2\right]\\
&=\mathbb{E}\left[\left\|g_{k+1}^i-g_k^i\right\|^2\right]\\
&=\mathbb{E}\left[\left\|\beta_{k+1}\left(\hat{\nabla}h_i\left(\bar{x}_k^i,\xi_{k+1}^i\right)-g_k^i\right)+\hat{\nabla}h_i\left(\bar{x}_{k+1}^i,\xi_{k+1}^i\right)\right.\right.\\
&\quad\left.\left.-\hat{\nabla}h_i\left(\bar{x}_k^i,\xi_{k+1}^i\right)\right\|^2\right]\\
&\le 3\beta_{k+1}^2\mathbb{E}\left[\left\|g_k^i\right\|^2\right]+3\beta_{k+1}^2\mathbb{E}\left[\left\|\hat{\nabla}h_i\left(\bar{x}_k^i,\xi_{k+1}^i\right)\right.\right.\\
&\quad\left.\left.-\nabla h_i\left(\bar{x}_k^i,\xi_{k+1}^i\right)+\nabla h_i\left(\bar{x}_k^i,\xi_{k+1}^i\right)\right\|^2\right]\\
&\quad+3\mathbb{E}\left[\left\|\hat{\nabla}h_i\left(\bar{x}_{k+1}^i,\xi_{k+1}^i\right)-\nabla h_i\left(\bar{x}_{k+1}^i,\xi_{k+1}^i\right)\right.\right.\\
&\quad\left.\left.+\nabla h_i\left(\bar{x}_k^i,\xi_{k+1}^i\right)-\hat{\nabla}h_i\left(\bar{x}_k^i,\xi_{k+1}^i\right)\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+\nabla h_i(\bar{x}_{k+1}^i,\xi_{k+1}^i)-\nabla h_i(\bar{x}_k^i,\xi_{k+1}^i)\|^2]\\
&\overset{(a)}{\le} 9\mathbb{E}[\|\hat{\nabla} h_i(\bar{x}_{k+1}^i,\xi_{k+1}^i)-\nabla h_i(\bar{x}_{k+1}^i,\xi_{k+1}^i)\|^2]\\
&\quad+9\mathbb{E}[\|\nabla h_i(\bar{x}_k^i,\xi_{k+1}^i)-\hat{\nabla} h_i(\bar{x}_k^i,\xi_{k+1}^i)\|^2]\\
&\quad+9\mathbb{E}[\|\nabla h_i(\bar{x}_{k+1}^i,\xi_{k+1}^i)-\nabla h_i(\bar{x}_k^i,\xi_{k+1}^i)\|^2]\\
&\quad+\big(6l^2+3\psi_2\big)\beta_{k+1}^2+6nL^2\rho_k^2\beta_{k+1}^2\\
&\overset{(b)}{\le} 9nL^2\rho_{k+1}^2+15nL^2\rho_k^2+\big(6l^2+3\psi_2\big)\beta_{k+1}^2\\
&\quad+9L^2\|\bar{x}_{k+1}^i-\bar{x}_k^i\|^2\\
&\overset{(c-)}{\le} 24nL^2\rho_k^2+\big(6l^2+3\psi_2\big)\beta_{k+1}^2\\
&\quad+9L^2(d+2C_1)^2\gamma_k^2, \qquad (32)
\end{aligned}
$$

where (a) follows from (13), (24), and Fact 2; (b) holds by (11), the smoothness of h_i, and the fact that $\beta_{k+1}<1$; (c) holds because of the fact that $\rho_{k+1}\le\rho_k$ and Lemma 1. Furthermore, we use the result in (32) to yield that

$$
\begin{aligned}
&\mathbb{E}[\|\Delta g_{k+1}^i-\Delta\hat{g}_{k+1}\|^2]\\
&=\mathbb{E}\left[\left\|\Delta g_{k+1}^i-\frac{1}{N}\sum_{i=1}^N\Delta g_{k+1}^i\right\|^2\right]\\
&=\mathbb{E}\left[\left\|\left(1-\frac{1}{N}\right)\Delta g_{k+1}^i-\frac{1}{N}\sum_{j\neq i}\Delta g_{k+1}^j\right\|^2\right]\\
&\le 2\left(1-\frac{1}{N}\right)\mathbb{E}[\|\Delta g_{k+1}^i\|^2]+\frac{2}{N}\sum_{j\neq i}\mathbb{E}[\|\Delta g_{k+1}^j\|^2]\\
&\le 4\left(1-\frac{1}{N}\right)\big(24nL^2\rho_k^2+\big(6l^2+3\psi_2\big)\beta_{k+1}^2\\
&\quad+9L^2(d+2C_1)^2\gamma_k^2\big)\\
&\le\frac{4(60L^2(d+2C_1)^2+12(2l^2+\psi_2))}{(k+2)^2}\\
&=\big(60L^2(d+2C_1)^2+12\big(2l^2+\psi_2\big)\big)\gamma_k^2, \qquad (33)
\end{aligned}
$$

where we use the fact that $1-\frac{1}{N}\le 1$ and the choice of ρ_k, β_k, γ_k in the last inequality. Taking full expectation on the RHS and LHS of (31), and then substituting (33) into the result, we obtain

$$
\begin{aligned}
&\sum_{i=1}^N\mathbb{E}[\|\Delta g_{k+1}^i+s_k^i-\bar{g}_{k+1}\|^2]\\
&\overset{(a)}{\le}\mathbb{E}\left[\sum_{i=1}^N\|s_k^i-\bar{g}_k\|^2\right]\\
&\quad+N\big(60L^2(d+2C_1)^2+12\big(2l^2+\psi_2\big)\big)\gamma_k^2\\
&\quad+2\sum_{i=1}^N\big(\mathbb{E}[\|s_k^i-\bar{g}_k\|^2]\big)^{\frac{1}{2}}\big(\mathbb{E}[\|\Delta g_{k+1}^i-\Delta\hat{g}_{k+1}\|^2]\big)^{\frac{1}{2}}\\
&\overset{(b)}{\le} C_2\gamma_k^2+N\big(60L^2(d+2C_1)^2+12\big(2l^2+\psi_2\big)\big)\gamma_k^2\\
&\quad+2N\gamma_k^2\sqrt{\big(60L^2(d+2C_1)^2+12\big(2l^2+\psi_2\big)\big)C_2}\\
&\le\gamma_k^2\big(C_2+N^2\big(60L^2(d+2C_1)^2+12\big(2l^2+\psi_2\big)\big)\\
&\quad+2N\sqrt{\big(60L^2(d+2C_1)^2+12\big(2l^2+\psi_2\big)\big)C_2}\big)\\
&\le\gamma_k^2\left(\sqrt{C_2}+\frac{\sqrt{C_2}}{k_0}\right)^2=\gamma_k^2\left(\frac{k_0+1}{k_0}\sqrt{C_2}\right)^2, \qquad (34)
\end{aligned}
$$

where (a) is due to the Hölder's inequality; (b) follows from the induction hypothesis. Hence, (30) is written as

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^N\|s_{k+1}^i-\bar{g}_{k+1}\|^2\right]&\le|\lambda|^2\gamma_k^2\left(\frac{k_0+1}{k_0}\sqrt{C_2}\right)^2\\
&\le\left(\frac{2k_0}{(k_0+1)(k+2)}\sqrt{C_2}\right)^2\\
&\le\left(\frac{2}{k+3}\sqrt{C_2}\right)^2=C_2\gamma_{k+1}^2,
\end{aligned}
$$

where we use the relations $|\lambda|\le[k_0/(k_0+1)]^2$, $\gamma_k=2/(k+2)$, and the monotonically increasing property of function $g(x)=x/(1+x)$ with respect to x over $[0,+\infty)$. That is, (28) holds when $k\leftarrow k+1$. The required result is obtained. □

1.5 Proof of Lemma 3

Proof 1) According to the definition of $\bar{g}_k$ and the update (3), we have

$$
\begin{aligned}
&\bar{g}_k-\hat{\nabla}\bar{h}_k\\
&=\frac{1}{N}\sum_{i=1}^N g_k^i-\hat{\nabla}\bar{h}_k\\
&=(1-\beta_k)\bar{g}_{k-1}+\frac{1}{N}\sum_{i=1}^N\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)\\
&\quad-\hat{\nabla}\bar{h}_k-(1-\beta_k)\frac{1}{N}\sum_{i=1}^N\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i).
\end{aligned}
$$

Introducing $(1-\beta_k)\hat{\nabla}\bar{h}_{k-1}$ into the RHS of the above equality and rearranging, we arrive at

$$
\begin{aligned}
&\bar{g}_k-\hat{\nabla}\bar{h}_k\\
&=(1-\beta_k)(\bar{g}_{k-1}-\hat{\nabla}\bar{h}_{k-1})
\end{aligned}
$$

$$+\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)-\hat{\nabla}\bar{h}_k$$
$$+(1-\beta_k)\left(\hat{\nabla}\bar{h}_{k-1}-\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i)\right). \quad (35)$$

Taking the squared-norm on RHS and LHS of (35) and then taking conditional expectation on $\mathcal{F}_k$, we obtain

$$\begin{aligned}
&\mathbb{E}_k\big[\|\bar{g}_k-\hat{\nabla}\bar{h}_k\|^2\big]\\
&=(1-\beta_k)^2\|\bar{g}_{k-1}-\hat{\nabla}\bar{h}_{k-1}\|^2\\
&\quad+2(1-\beta_k)(\bar{g}_{k-1}-\hat{\nabla}\bar{h}_{k-1})\\
&\quad\times\left(\mathbb{E}_k\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)-\hat{\nabla}\bar{h}_k\right]\right.\\
&\quad\left.+(1-\beta_k)\mathbb{E}_k\left[\hat{\nabla}\bar{h}_{k-1}-\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i)\right]\right)\\
&\quad+\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)-\hat{\nabla}\bar{h}_k+(1-\beta_k)\right.\right.\\
&\quad\left.\left.\times\left(\hat{\nabla}\bar{h}_{k-1}-\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i)\right)\right\|^2\right]\\
&=(1-\beta_k)^2\|\bar{g}_{k-1}-\hat{\nabla}\bar{h}_{k-1}\|^2\\
&\quad+\mathbb{E}_k\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)-\hat{\nabla}\bar{h}_k+(1-\beta_k)\right.\right.\\
&\quad\left.\left.\times\left(\hat{\nabla}\bar{h}_{k-1}-\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i)\right)\right\|^2\right], \quad (36)
\end{aligned}$$

where the last equality holds due to the fact that

$$\mathbb{E}_k\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)-\hat{\nabla}\bar{h}_k\right]=0$$

and

$$\mathbb{E}_k\left[\hat{\nabla}\bar{h}_{k-1}-\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i)\right]=0.$$

Next, we focus on the last term of the RHS of (36) and bounding it separately. For convenience, we define $\mathbb{U}_k:=\beta_k(\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)-\hat{\nabla}\bar{h}_k)$ and $\mathbb{V}_k:=\frac{(1-\beta_k)}{N}\sum_{i=1}^{N}(\hat{\nabla}h_i\times(\bar{x}_k^i,\xi_k^i))-\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i))$. Hence, the second term of the RHS of (36) is modified as $\mathbb{E}_k[\|\mathbb{U}_k+\mathbb{V}_k-\mathbb{E}_k[\mathbb{V}_k]\|^2]$ and bounded by

$$\begin{aligned}
&\mathbb{E}_k\big[\big\|\mathbb{U}_k+\mathbb{V}_k-\mathbb{E}_k[\mathbb{V}_k]\big\|^2\big]\\
&\le 2\mathbb{E}_k\big[\|\mathbb{U}_k\|^2\big]+2\mathbb{E}_k\big[\big\|\mathbb{V}_k-\mathbb{E}_k[\mathbb{V}_k]\big\|^2\big]\\
&\le 2\mathbb{E}_k\big[\|\mathbb{U}_k\|^2\big]+2\mathbb{E}_k\big[2\|\mathbb{V}_k\|^2+2\mathbb{E}_k\big[\|\mathbb{V}_k\|^2\big]\big]\\
&\le 2\mathbb{E}_k\big[\|\mathbb{U}_k\|^2\big]+8\mathbb{E}_k\big[\|\mathbb{V}_k\|^2\big], \quad (37)
\end{aligned}$$

where we use (13) and the Jensen's inequality.

Next, we will derive the bounds of $\mathbb{E}_k[\|\mathbb{U}_k\|^2]$ and $\mathbb{E}_k[\|\mathbb{V}_k\|^2]$. Obviously, $\hat{\nabla}\bar{h}_t=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_k[\hat{\nabla}h_i(\bar{x}_t^i,\xi_k^i)]=\frac{1}{N}\times\sum_{i=1}^{N}\hat{\nabla}H_i(\bar{x}_t^i)$. It follows from the definitions of $\mathbb{U}_k$ and $\hat{\nabla}\bar{h}_k$ that

$$\begin{aligned}
&\mathbb{E}_k\big[\|\mathbb{U}_k\|^2\big]\\
&\le\frac{\beta_k^2}{N}\sum_{i=1}^{N}\mathbb{E}_k\big[\big\|\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)-\hat{\nabla}H_i(\bar{x}_k^i)\big\|^2\big]\\
&=\frac{\beta_k^2}{N}\sum_{i=1}^{N}\mathbb{E}_k\big[\big\|\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)+\nabla h_i(\bar{x}_k^i,\xi_k^i)\\
&\quad+\nabla H_i(\bar{x}_k^i)-\hat{\nabla}H_i(\bar{x}_k^i)\\
&\quad-\nabla h_i(\bar{x}_k^i,\xi_k^i)-\nabla H_i(\bar{x}_k^i)\big\|^2\big]\\
&\overset{(a)}{\le}\frac{\beta_k^2}{N}\sum_{i=1}^{N}\big(3\mathbb{E}_k\big[\big\|\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)-\nabla h_i(\bar{x}_k^i,\xi_k^i)\big\|^2\big]\\
&\quad+3\mathbb{E}_k\big[\big\|\nabla H_i(\bar{x}_k^i)-\hat{\nabla}H_i(\bar{x}_k^i)\big\|^2\big]\\
&\quad+3\mathbb{E}_k\big[\big\|\nabla h_i(\bar{x}_k^i,\xi_k^i)-\nabla H_i(\bar{x}_k^i)\big\|^2\big]\big)\\
&\overset{(b)}{\le}6nL^2\rho_k^2+3\delta^2\beta_k^2, \quad (38)
\end{aligned}$$

where (a) holds because of using the inequality (13); (b) follows from Fact 3 and the fact that $\beta_k\le 1$. Similarly, it follows from (13), Fact 3, and the smoothness of h_i that

$$\begin{aligned}
&\mathbb{E}_k\big[\|\mathbb{V}_k\|^2\big]\\
&\le\frac{(1-\beta_k)^2}{N}\sum_{i=1}^{N}\mathbb{E}_k\big[\big\|\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i)-\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)\big\|^2\big]\\
&=\frac{(1-\beta_k)^2}{N}\sum_{i=1}^{N}\mathbb{E}_k\big[\big\|\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i)-\nabla h_i(\bar{x}_{k-1}^i,\xi_k^i)\\
&\quad+\nabla h_i(\bar{x}_{k-1}^i,\xi_k^i)\\
&\quad-\nabla h_i(\bar{x}_k^i,\xi_k^i)+\nabla h_i(\bar{x}_k^i,\xi_k^i)-\hat{\nabla}h_i(\bar{x}_k^i,\xi_k^i)\big\|^2\big]\\
&\le\frac{(1-\beta_k)^2}{N}\sum_{i=1}^{N}\big(3\mathbb{E}_k\big[\big\|\hat{\nabla}h_i(\bar{x}_{k-1}^i,\xi_k^i)\\
&\quad-\nabla h_i(\bar{x}_{k-1}^i,\xi_k^i)\big\|^2\big]+3\mathbb{E}_k\big[\big\|\nabla h_i(\bar{x}_k^i,\xi_k^i)
\end{aligned}$$

$$
\begin{aligned}
&- \hat{\nabla} h_i(\bar{x}_k^i, \xi_k^i)\|^2] \\
&+ 3\mathbb{E}_k[\|\nabla h_i(\bar{x}_{k-1}^i, \xi_k^i) - \nabla h_i(\bar{x}_k^i, \xi_k^i)\|^2]) \\
&\le \frac{(1-\beta_k)^2}{N} \sum_{i=1}^{N} \big(3nL^2\rho_{k-1}^2 + 3nL^2\rho_k^2 \\
&+ 3L^2\|\bar{x}_{k-1}^i - \bar{x}_k^i\|^2\big) \\
&\le 6nL^2\rho_{k-1}^2 + 3L^2\gamma_{k-1}^2(d+2C_1)^2, \qquad (39)
\end{aligned}
$$

where we use Lemma 1 and the fact $\rho_k \le \rho_{k-1}$, and drop the factor $(1-\beta_k)^2$ in the last inequality. Substituting (38) and (39) into (37), we obtain

$$
\begin{aligned}
&\mathbb{E}_k[\|\mathbb{U}_k + \mathbb{V}_k - \mathbb{E}_k[\mathbb{V}_k]\|^2] \\
&\le 12nL^2\rho_k^2 + 6\delta^2\beta_k^2 + 48nL^2\rho_{k-1}^2 \\
&\quad + 24L^2\gamma_{k-1}^2(d+2C_1)^2 \\
&\le 60nL^2\rho_{k-1}^2 + 6\delta^2\beta_k^2 + 24L^2\gamma_{k-1}^2(d+2C_1)^2.
\end{aligned}
$$

Substituting the above result into (36), we yield Eqn. (15).

2) Taking $\beta_k = \frac{2}{k+1}$, $\gamma_k = \frac{2}{k+2}$ and $\rho_k \le \frac{d+2C_1}{\sqrt{n}(k+2)}$, we rewrite (15) as

$$
\begin{aligned}
&\mathbb{E}[\|\bar{g}_k - \hat{\nabla}\bar{h}_k\|^2] \\
&\le \left(1 - \frac{2}{k+1}\right)\mathbb{E}[\|\bar{g}_{k-1} - \hat{\nabla}\bar{h}_{k-1}\|^2] + \frac{60L^2(d+2C_1)^2}{(k+1)^2} \\
&\quad + \frac{24\delta^2}{(k+1)^2} + \frac{96L^2(d+2C_1)^2}{(k+1)^2} \\
&= \left(1 - \frac{2}{k+1}\right)\mathbb{E}[\|\bar{g}_{k-1} - \hat{\nabla}\bar{h}_{k-1}\|^2] \\
&\quad + \frac{156L^2(d+2C_1)^2 + 24\delta^2}{(k+1)^2}.
\end{aligned}
$$

Using Lemma 5 with $a_1 = 1$, $t_0 = 1$, $a_2 = 2$, $A_1 = 2$ and $A_2 = 156L^2(d+2C_1)^2 + 24\delta^2$, we yield that $\mathbb{E}[\|\bar{g}_k - \hat{\nabla}\bar{h}_k\|^2] \le \frac{C_3}{k+2}$, where $C_3 := \max\{2\|\bar{g}_1 - \hat{\nabla}h(x_1)\|^2, 156L^2(d+2C_1)^2 + 24\delta^2\}$. Focusing on $\mathbb{E}[\|\bar{g}_k - \bar{p}_k\|^2]$ and introducing $\frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}H_i(\bar{x}_k^i)$, according to the definition of $\bar{p}_k$ and the relation $\hat{\nabla}\bar{h}_k = \frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}H_i(\bar{x}_k^i)$, we have

$$
\begin{aligned}
&\mathbb{E}[\|\bar{g}_k - \bar{p}_k\|^2] \\
&= \mathbb{E}\left[\left\|\bar{g}_k - \hat{\nabla}\bar{h}_k + \frac{1}{N}\sum_{i=1}^{N}\hat{\nabla}H_i(\bar{x}_k^i)\right.\right. \\
&\quad \left.\left. - \frac{1}{N}\sum_{i=1}^{N}\nabla H_i(\bar{x}_k^i)\right\|^2\right] \\
&\le 2\mathbb{E}[\|\bar{g}_k - \hat{\nabla}\bar{h}_k\|^2] \\
&\quad + \frac{2}{N}\sum_{i=1}^{N}\mathbb{E}[\|\hat{\nabla}H_i(\bar{x}_k^i) - \nabla H_i(\bar{x}_k^i)\|^2] \\
&\le \frac{2C_3 + 2L^2(d+2C_1)^2}{k+2},
\end{aligned}
$$

where we use (13) in the first inequality, and the last inequality holds due to (12). □

1.6 Proof of Lemma 4

Proof Focusing on the LHS of (17), adding and subtracting the term $(\bar{p}_k + \bar{g}_k)$ into $\|\nabla h(\bar{x}_k) - s_k^i\|^2$, we have

$$
\begin{aligned}
&\mathbb{E}[\|\nabla h(\bar{x}_k) - s_k^i\|^2] \\
&= \mathbb{E}[\|\nabla h(\bar{x}_k) - \bar{p}_k + \bar{p}_k - \bar{g}_k + \bar{g}_k - s_k^i\|^2] \\
&\le 3\mathbb{E}[\|\nabla h(\bar{x}_k) - \bar{p}_k\|^2] + 3\mathbb{E}[\|\bar{p}_k - \bar{g}_k\|^2] \\
&\quad + 3\mathbb{E}[\|\bar{g}_k - s_k^i\|^2], \qquad (40)
\end{aligned}
$$

where the last inequality follows from (13). The first term of the RHS of (40) is rewritten as

$$
\begin{aligned}
&3\mathbb{E}[\|\nabla h(\bar{x}_k) - \bar{p}_k\|^2] \\
&= 3\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\big(\nabla H_i(\bar{x}_k) - \nabla H_i(\bar{x}_k^i)\big)\right\|^2\right] \\
&\le 3\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\|\nabla H_i(\bar{x}_k) - \nabla H_i(\bar{x}_k^i)\|\right)^2\right] \\
&\le \frac{3L^2}{N}\sum_{i=1}^{N}\|\bar{x}_k - \bar{x}_k^i\|^2 \\
&\le 3L^2C_1^2\gamma_k^2, \qquad (41)
\end{aligned}
$$

where we use the smoothness of h_i and Lemma 1. Substituting (41), (16) and (14) into (40) and taking $\beta_k = \frac{2}{k+1}$, $\gamma_k = \frac{2}{k+2}$, $\rho_k \le \frac{d+2C_1}{\sqrt{n}(k+2)}$, we have the result. □

1.7 Proof of Theorem 1

Proof It follows from the update (7) in Algorithm 1 and Assumption 3 that

$$
\begin{aligned}
h(\bar{x}_{k+1}) &\le h(\bar{x}_k) + \gamma_k\langle\nabla h(\bar{x}_k), z_k - \bar{x}_k\rangle + \frac{L\gamma_k^2}{2}\|z_k - \bar{x}_k\|^2 \\
&\le h(\bar{x}_k) + \gamma_k\langle\nabla h(\bar{x}_k), z_k - \bar{x}_k\rangle + \frac{L\gamma_k^2 d^2}{2}, \qquad (42)
\end{aligned}
$$

where we use Assumption 4 in the last inequality. Focusing on the second term of the RHS of (42) and the definition

of $\bar{z}_k$, we have

$$\begin{aligned}
&\langle \nabla h(\bar{x}_k), z_k - \bar{x}_k \rangle \\
&= \frac{1}{N}\sum_{i=1}^{N} \langle \nabla h(\bar{x}_k), z_k^i - \bar{x}_k \rangle \\
&= \frac{1}{N}\sum_{i=1}^{N} \langle \nabla h(\bar{x}_k) - s_k^i + s_k^i, z_k^i - \bar{x}_k \rangle \\
&\le \frac{1}{N}\sum_{i=1}^{N} \big[\langle \nabla h(\bar{x}_k) - s_k^i, z_k^i - \bar{x}_k \rangle + \langle s_k^i, x^* - \bar{x}_k \rangle \big],
\end{aligned}$$

where we use the optimality of z_k^i in the last inequality. Adding and subtracting the term $\frac{1}{N}\sum_{i=1}^{N}[\langle \nabla h(\bar{x}_k) - s_k^i, x^* \rangle]$ into the RHS of the above inequality, we arrive at

$$\begin{aligned}
&\langle \nabla h(\bar{x}_k), z_k - \bar{x}_k \rangle \\
&\le \frac{1}{N}\sum_{i=1}^{N} \big[\langle \nabla h(\bar{x}_k) - s_k^i, z_k^i - x^* \rangle + \langle \nabla h(\bar{x}_k), x^* - \bar{x}_k \rangle \big] \\
&\overset{(a)}{\le} \frac{1}{N}\sum_{i=1}^{N} \| \nabla h(\bar{x}_k) - s_k^i \| \| z_k^i - x^* \| + h(x^*) - h(\bar{x}_k) \\
&\overset{(b)}{\le} \frac{1}{N}\sum_{i=1}^{N} \| \nabla h(\bar{x}_k) - s_k^i \| d + h(x^*) - h(\bar{x}_k),
\end{aligned} \tag{43}$$

where (a) holds by the convexity of function $h(x)$; (b) follows from Assumption 4. Substituting (43) into (42), rearranging, and subtracting $h(x^*)$ from the RHS and LHS of the result, we arrive at

$$\begin{aligned}
h(\bar{x}_{k+1}) - h(x^*) \le (1-\gamma_k)\big(h(\bar{x}_k) - h(x^*)\big) + \frac{L\gamma_k^2 d^2}{2} \\
+ \frac{d\gamma_k}{N}\sum_{i=1}^{N} \| \nabla h(\bar{x}_k) - s_k^i \|.
\end{aligned} \tag{44}$$

Taking the expectation on the RHS and LHS of (44), and then using the Jensen's inequality in the last term of RHS of (44), we have

$$\begin{aligned}
&\mathbb{E}\big[h(\bar{x}_{k+1})\big] - h(x^*) \\
&\le (1-\gamma_k)\big(\mathbb{E}\big[h(\bar{x}_k)\big] - h(x^*)\big) + \frac{L\gamma_k^2 d^2}{2} \\
&\quad + \frac{d\gamma_k}{N}\sum_{i=1}^{N} \sqrt{\mathbb{E}\big[\| \nabla h(\bar{x}_k) - s_k^i \|^2\big]}.
\end{aligned} \tag{45}$$

It follows from Lemma 4, $\gamma_k = \frac{2}{k+2}$, and (45) that

$$\begin{aligned}
&\mathbb{E}\big[h(\bar{x}_{k+1})\big] - h(x^*) \\
&\le \left(1 - \frac{2}{k+2}\right)\big(\mathbb{E}\big[h(\bar{x}_k)\big] - h(x^*)\big) \\
&\quad + \frac{2d\sqrt{18L^2(d+2C_1)^2 + 12C_2 + 6C_3}}{(k+2)^{\frac{3}{2}}} \\
&\quad + \frac{2Ld^2}{(k+2)^{\frac{3}{2}}}.
\end{aligned} \tag{46}$$

Using Lemma 5 with

$$\begin{aligned}
&A_1 = 2, \\
&A_2 = 2Ld^2 + 2d\sqrt{18L^2(d+2C_1)^2 + 12C_2 + 6C_3}, \\
&t_0 = 2, \qquad a_1 = 1, \quad \text{and} \quad a_2 = 3/2,
\end{aligned}$$

we prove the result. □

1.8 Proof of Theorem 2

Proof Define $v_k \in \operatorname{argmin}_{v\in\mathcal{X}}\langle \nabla h(\bar{x}_k), v\rangle$. We have $p_k = \langle \nabla h(\bar{x}_k), \bar{x}_k - v_k\rangle$ by (18) and the definition of v_k. We also obtain from the smoothness property of $h(\cdot)$ that

$$\begin{aligned}
h(\bar{x}_{k+1}) &\le \frac{L}{2}\|\bar{x}_{k+1} - \bar{x}_k\|^2 + h(\bar{x}_k) + \langle \nabla h(\bar{x}_k), \bar{x}_{k+1} - \bar{x}_k \rangle \\
&\overset{(a)}{=} \frac{L\gamma_k^2}{2}\|\bar{z}_k - \bar{x}_k\| + h(\bar{x}_k) \\
&\quad + \frac{\gamma_k}{N}\sum_{i=1}^{N} \langle \nabla h(\bar{x}_k) + s_k^i - s_k^i, z_k^i - \bar{x}_k \rangle \\
&\overset{(b)}{\le} \frac{L\gamma_k^2}{2}\|\bar{z}_k - \bar{x}_k\|^2 + h(\bar{x}_k) + \frac{\gamma_k}{N}\sum_{i=1}^{N} \langle s_k^i, v_k - \bar{x}_k \rangle \\
&\quad + \frac{\gamma_k}{N}\sum_{i=1}^{N} \langle \nabla h(\bar{x}_k) - s_k^i, z_k^i - \bar{x}_k \rangle,
\end{aligned}$$

where (a) holds by Lemma 6(b) and introducing s_k^i; (b) is due to the optimality of z_k^i in the update (6). It follows from the definition of p_k and Assumption 4 that

$$\begin{aligned}
h(\bar{x}_{k+1}) &= \frac{L\gamma_k^2}{2}\|\bar{z}_k - \bar{x}_k\|^2 + h(\bar{x}_k) \\
&\quad + \frac{\gamma_k}{N}\sum_{i=1}^{N} \langle s_k^i - \nabla h(\bar{x}_k), v_k - \bar{x}_k \rangle \\
&\quad + \frac{\gamma_k}{N}\sum_{i=1}^{N} \langle \nabla h(\bar{x}_k), v_k - \bar{x}_k \rangle \\
&\quad + \frac{\gamma_k}{N}\sum_{i=1}^{N} \langle \nabla h(\bar{x}_k) - s_k^i, z_k^i - \bar{x}_k \rangle
\end{aligned}$$

$$\overset{\text{(c-)}}{\leq} \frac{Ld^2\gamma_k^2}{2} - \gamma_k p_k + h(\bar{x}_k) + \frac{2d\gamma_k}{N}\sum_{i=1}^{N}\left\|\nabla h(\bar{x}_k) - s_k^i\right\|. \tag{47}$$

Taking the full expectation on the RHS and LHS of (47) and using Jensen's inequality, we yield that

$$\begin{aligned}\mathbb{E}\left[h(\bar{x}_{k+1})\right] &\leq \frac{2d\gamma_k}{N}\sum_{i=1}^{N}\sqrt{\mathbb{E}\left[\left\|\nabla h(\bar{x}_k) - s_k^i\right\|^2\right]} + \mathbb{E}\left[h(\bar{x}_k)\right] + \frac{Ld^2\gamma_k^2}{2} - \gamma_k\mathbb{E}[p_k] \\ &\leq \mathbb{E}\left[h(\bar{x}_k)\right] + \frac{4d\sqrt{18L^2(d+2C_1)^2 + 12C_2 + 6C_3}}{(k+2)^{\frac{3}{2}}} - \frac{2\mathbb{E}[p_k]}{k+2} + \frac{2Ld^2}{(k+2)^2},\end{aligned} \tag{48}$$

where we use (17) in the last inequality, and substitute $\gamma_k = \frac{2}{k+2}$ into the last inequality. Summing the RHS and LHS of (48) from $k = 1$ to $k = K$ and rearranging, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^{K}\frac{2p_k}{k+2}\right] &\leq h(\bar{x}_1) - h(\bar{x}_{K+1}) + \sqrt{18L^2(d+2C_1)^2 + 12C_2 + 6C_3} \times \sum_{k=1}^{K}\frac{4d}{(k+2)^{\frac{3}{2}}} + Ld^2\sum_{k=1}^{K}\frac{2}{(k+2)^2}.\end{aligned} \tag{49}$$

Define m such that $2^m = K$, i.e., $m = \log_2(K)$. According to the property of p-series, we yield that $m - 1 \leq \sum_{k=1}^{2^m}\frac{2}{k+2}$, $\sum_{k=1}^{2^m}\frac{2}{(k+2)^2} \leq 4$, and there are some constant c such that $\sum_{k=1}^{2^m}\frac{4d}{(k+2)^{\frac{3}{2}}} \leq c$. Hence, we rewrite (49) as

$$(m-1)\mathbb{E}\left[\min_{k\in[1,K]} p_k\right] \leq h(\bar{x}_1) - h(\bar{x}_{K+1}) + 4Ld^2 + \sqrt{18L^2(d+2C_1)^2 + 12C_2 + 6C_3}.$$

By rearranging and substituting $m = \log_2(K)$, we obtain the result. □

Acknowledgements
The authors would like to thank the anonymous reviewers and potential users for their valuable comments and suggestions.

Author contributions
All authors contributed to the design and implementation of the research. Material preparation and analysis were completed by Jie Hou, Xianlin Zeng and Chen Chen. Jie Hou and Xianlin Zeng contributed to the problem formulation, discussion of ideas, mathematical derivation and proof of results. Chen Chen contributed to the problem formulation and discussion of ideas. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China under Grant Nos. 62222303, 62073035, and 62088101.

Data availability
Not applicable.

Declarations

Competing interests
The authors declare no competing interests.

Received: 27 November 2023 Revised: 1 April 2024 Accepted: 8 April 2024
Published online: 01 May 2024

References

1. S. Aeron, V. Saligrama, D.A. Castanon, Efficient sensor management policies for distributed target tracking in multihop sensor networks. IEEE Trans. Signal Process. **56**(6), 2562–2574 (2008)
2. Z. Akhtar, K. Rajawat, Momentum based projection free stochastic optimization under affine constraints, in *American Control Conf.* (2021), pp. 2619–2624
3. Z. Akhtar, K. Rajawat, Zeroth and first order stochastic Frank-Wolfe algorithms for constrained optimization. IEEE Trans. Signal Process. **70**, 2119–2135 (2022)
4. K. Balasubramanian, S. Ghadimi, Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates, in *Proc. Int. Conf. Neural Inf. Process. Syst.* (2018), pp. 3459–3468
5. A. Bellet, Y. Liang, A.B. Garakani et al., A distributed Frank-Wolfe algorithm for communication-efficient sparse learning, in *Proc. SIAM Int. Conf. Data Mining* (2015), pp. 478–486. https://doi.org/10.1137/1.97816119740 10.54
6. G. Chen, P. Yi, Y. Hong et al., Distributed optimization with projection-free dynamics: a Frank-Wolfe perspective. IEEE Trans. Cybern. **54**(1), 599–610 (2024). https://doi.org/10.1109/TCYB.2023.3284822
7. J. Chen, J. Sun, G. Wang, From unmanned systems to autonomous intelligent systems. Engineering **12**(5), 16–19 (2022)
8. P. Chen, H. Zhang, Y. Sharma et al., ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in *Proc. ACM. Work. Artif. Intell. Sec.* (2017), pp. 15–26
9. A. Cutkosky, F. Orabona, Momentum-based variance reduction in non-convex SGD, in *Proc. Adv. Neural Inf. Process. Syst.* (2019), pp. 15210–15219
10. M. Frank, P. Wolfe, An algorithm for quadratic programming. Nav. Res. Logist. **3**(1–2), 95–110 (1956)
11. K. Fu, H. Chen, W. Zhao, Distributed dynamic stochastic approximation algorithm over time-varying networks. Auton. Intell. Syst. **1**(5) (2021). https://doi.org/10.1007/s43684-021-00003-1
12. E. Hazan, H. Luo, Variance-reduced and projection-free stochastic optimization, in *Proc. Int. Conf. Mach. Learn* (2016)
13. J. Hou, X. Zeng, G. Wang et al., Distributed momentum-based Frank-Wolfe algorithm for stochastic optimization. IEEE/CAA J. Autom. Sin. **10**(3), 676–690 (2023)
14. F. Huang, S. Chen, Accelerated stochastic gradient-free and projection-free methods, in *Proc. Int. Conf. Mach. Learn.* (2020), pp. 4519–4530
15. M. Jaggi, Revisiting Frank-Wolfe: projection-free sparse convex optimization, in *Proc. Int. Conf*, Mach. Learn., Atlanta, GA, USA (2013), pp. 427–435
16. Y. Kuriki, T. Namerikawa, Consensus-based cooperative formation control with collision avoidance for a multi-UAV system, in *American Control Conf.* (2014), pp. 2077–2082
17. D. Li, N. Li, L. Lewis, Projection-free distributed optimization with nonconvex local objective functions and resource allocation constraint. IEEE Trans. Control Netw. Syst. **8**(1), 413–422 (2021)

18. A. Mokhtari, H. Hassani, A. Karbasi, Stochastic conditional gradient methods: from convex minimization to submodular maximization. J. Mach. Learn. Res. **21**(105), 1–49 (2020)
19. S. Pu, A. Olshevsky, I.C. Paschalidis, Asymptotic network independence in distributed stochastic optimization for machine learning: examining distributed and centralized stochastic gradient descent. IEEE Signal Process. Mag. **37**(3), 114–122 (2020)
20. R. Rubinstein, D. Kroese, *Simulation and the Monte Carlo Method*, vol. 10 (Wiley, New York, 2016)
21. A. Sahu, D. Jakovetic, D. Bajovic et al., Distributed zeroth order optimization over random networks: a Kiefer-Wolfowitz stochastic approximation approach, in *IEEE Conf. Decision Contr.* (2018), pp. 4951–4958. https://doi.org/10.1109/CDC.2018.8619044
22. A. Sahu, S. Kar, Decentralized zeroth-order constrained stochastic optimization algorithms: Frank–Wolfe and variants with applications to black-box adversarial attacks. Proc. IEEE **108**(11), 1890–1905 (2020)
23. A. Sahu, M. Zaheer, S. Kar, Towards gradient free and projection free stochastic optimization, in *Proc. Int. Conf. Artif. Intell. Statis.* (2019), pp. 3468–3477
24. T. Salimans, J. Ho, X. Chen et al., Evolution strategies as a scalable alternative to reinforcement learning (2017). arXiv preprint https://doi.org/10.48550/arXiv.1703.03864
25. P. Sun, Z. Guo, G. Wang et al., MARVEL: enabling controller load balancing in software-defined networks with multi-agent reinforcement learning. Comput. Netw. **177**, 107230 (2020)
26. H. Wai, J. Lafond, A. Scaglione et al., Decentralized Frank-Wolfe algorithm for convex and nonconvex problems. IEEE Trans. Autom. Control **62**(11), 5522–5537 (2017)
27. D. Wang, Z. Wang, Z. Wu, Distributed convex optimization for nonlinear multi-agent systems disturbed by a second-order stationary process over a digraph. Sci. China Inf. Sci. **65**, 132201 (2022). https://doi.org/10.1007/s11432-020-3111-4
28. G. Wang, S. Lu, G.B. Giannakis et al., Decentralized TD tracking with linear function approximation and its finite-time analysis, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, vol. 1154 (2020), pp. 13762–13772
29. Z. Wang, J. Zhang, T. Chang et al., Distributed stochastic consensus optimization with momentum for nonconvex nonsmooth problems. IEEE Trans. Signal Process. **69**, 4486–4501 (2021)
30. Y. Xu, H. Deng, W. Zhu, Synchronous distributed admm for consensus convex optimization problems with self-loops. Inf. Sci. **614**, 185–205 (2022)
31. R. Yang, L. Liu, G. Feng, An overview of recent advances in distributed coordination of multi-agent systems. Unmanned Syst. **10**(03), 307–325 (2022)
32. X. Yi, S. Zhang, T. Yang et al., Linear convergence of first- and zeroth-order primal–dual algorithms for distributed nonconvex optimization. IEEE Trans. Autom. Control **67**(8), 4194–4201 (2022)
33. X. Yi, S. Zhang, T. Yang et al., Zeroth-order algorithms for stochastic distributed nonconvex optimization. Automatica **142**, 110353 (2022)
34. Z. Yu, D.W. Ho, D. Yuan, Distributed randomized gradient-free mirror descent algorithm for constrained optimization. IEEE Trans. Autom. Control **67**(2), 957–964 (2022)
35. D. Yuan, B. Zhang, D.W. Ho et al., Distributed online bandit optimization under random quantization. Automatica **146**, 110590 (2022)
36. S. Zhang, C.P. Bailey, Accelerated zeroth-order algorithm for stochastic distributed non-convex optimization, in *American Contr. Conf.* (2022), pp. 4274–4279. https://doi.org/10.23919/ACC53348.2022.9867306

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com