Zeroth-Order Algorithms for Stochastic Distributed Nonconvex Optimization

Xinlei Yi, Shengjun Zhang, Tao Yang, and Karl H. Johansson

Abstract

In this paper, we consider a stochastic distributed nonconvex optimization problem with the cost function being distributed over n agents having access only to zeroth-order (ZO) information of the cost. This problem has various machine learning applications. As a solution, we propose two distributed ZO algorithms, in which at each iteration each agent samples the local stochastic ZO oracle at two points with an adaptive smoothing parameter. We show that the proposed algorithms achieve the linear speedup convergence rate $\mathcal{O}(\sqrt{p/(nT)})$ for smooth cost functions and $\mathcal{O}(p/(nT))$ convergence rate when the global cost function additionally satisfies the Polyak–Łojasiewicz (P–Ł) condition, where p and T are the dimension of the decision variable and the total number of iterations, respectively. To the best of our knowledge, this is the first linear speedup result for distributed ZO algorithms, which enables systematic processing performance improvements by adding more agents. We also show that the proposed algorithms converge linearly when considering deterministic centralized optimization problems under the P–Ł condition. We demonstrate through numerical experiments the efficiency of our algorithms on generating adversarial examples from deep neural networks in comparison with baseline and recently proposed centralized and distributed ZO algorithms.

Index Terms—Distributed Nonconvex Optimization, Gradient-Free, Linear Speedup, Polyak-Łojasiewicz Condition, Stochastic Optimization.

This work was supported by the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, the Swedish Research Council, and the National Natural Science Foundation of China under grants 61991403 and 61991400.

X. Yi and K. H. Johansson are with the Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44, Stockholm, Sweden. {xinleiy, kallej}@kth.se.

S. Zhang is with the Department of Electrical Engineering, University of North Texas, Denton, TX 76203 USA. ShengjunZhang@my.unt.edu.

T. Yang is with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, 110819, Shenyang, China. {yangtao,tychai}@mail.neu.edu.cn.

I. Introduction

We consider stochastic distributed nonconvex optimization with zeroth-order (ZO) information feedback. Specifically, consider a network of n agents/machines collaborating to solve the following optimization problem

$$\min_{x\in\mathbb{R}^p} f(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}_{\xi_i}[F_i(x,\xi_i)], \tag{1}$$

where x is the decision variable, ξ_i is a random variable, and $F_i(\cdot,\xi_i) : \mathbb{R}^p \to \mathbb{R}$ is a stochastic component function (not necessarily convex). Each agent i only has information about its own stochastic ZO oracle $F_i(x,\xi_i)$. In other words, for any given x and ξ_i, each agent i can sample $F_i(x,\xi_i)$ as a stochastic approximation of the true local cost function value $f_i(x) = \mathbf{E}_{\xi_i}[F_i(x,\xi_i)]$, but other information such as the first-order oracle cannot be observed. Agents can communicate with their neighbors through an underlying communication network. The network is modeled by an undirected graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$, where $\mathcal{V} = \{1,\dots,n\}$ is the agent set, $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$ the edge set, and $(i,j) \in \mathcal{E}$ if agents i and j can communicate with each other. The neighboring set of agent i is denoted $\mathcal{N}_i = \{j \in \mathcal{V} :\ (i,j) \in \mathcal{E}\}$. The ZO information feedback setting has wide usage in applications, particularly when explicit expressions of the gradients are unavailable or difficult to obtain [1]–[3]. For example, the cost functions of many big data problems that deal with complex data generating processes cannot be explicitly defined [4]. Moreover, the distributed setting is a core aspect of many important applications in view of flexibility and scalability to large-scale datasets and systems, data privacy and locality, communication reduction to the central entity, and robustness to potential failures of the central entity [5]–[7].

A. Literature Review

The study of gradient-free (derivative-free) optimization has a long history, which can be traced back at least to the 1960's [8]–[10]. It has recently gained renewed attention by the machine learning community. Classical gradient-free optimization methods can be classified into direct-search and model-based methods. For example, stochastic direct-search and model-based trust-region algorithms have been proposed in [11]–[14] and [15]–[17], respectively. In recent years, the more popular gradient-free optimization methods are ZO methods, which are gradient-free counterparts of first-order optimization methods and thus easy to implement. In ZO optimization methods, the full or stochastic gradients are approximated by directional derivatives, which are calculated through sampled function values. A commonly used method to calculate directional derivatives is simply using the function difference at two points [18]–[20].

Various ZO optimization methods have been proposed, e.g., ZO (stochastic) gradient descent algorithms [20]–[31], ZO stochastic coordinate descent algorithms [32], ZO (stochastic) variance reduction algorithms [24], [25], [29], [30], [33]–[45], ZO (stochastic) proximal algorithms [33], [41], [46], [47], ZO Frank-Wolfe algorithms [24], [43], [45], [48], ZO mirror descent algorithms [18], [39], [49], ZO adaptive momentum methods [47], [50], ZO methods of multipliers [34], [35], [51], [52], ZO stochastic path-integrated differential estimator [37], [42], [52]. Convergence properties of these algorithms have been analyzed in detail. For instance, the typical convergence result for deterministic centralized optimization problems is that first-order stationary points can be found at an $\mathcal{O}(p/T)$ convergence rate by the two-point sampling based ZO algorithms [20], [28], while under stochastic settings the convergence rate is reduced to $\mathcal{O}(\sqrt{p/T})$ [22], [32], where T is the total number of iterations.

Aforementioned ZO optimization algorithms are all centralized and thus not suitable to solve distributed optimization problems. Recently distributed ZO algorithms have being proposed, e.g., distributed ZO gradient descent algorithms [53]–[57], distributed ZO push-sum algorithm [58], distributed ZO mirror descent algorithm [59], distributed ZO gradient tracking algorithm [57], distributed ZO primal–dual algorithms [60], [61], distributed ZO sliding algorithm [62]. Among these algorithms, the algorithms in [53], [54], [57]–[59], [61] are suitable to solve the deterministic form of (1), while the algorithm in [60] can be directly applied to solve the stochastic optimization problem (1). However, the algorithm in [60] requires each agent to have an $\mathcal{O}(T)$ sampling size per iteration, which is not favorable in practice, although it was shown that first-order stationary points can be found at an $\mathcal{O}(p^2n/T)$ convergence rate.

From the discussion above, three core theoretical questions arise when considering stochastic distributed optimization problems:

(Q1) Can distributed ZO algorithms achieve similar convergence properties as centralized ZO algorithms? For instance, can distributed ZO algorithms based on two-point sampling have an $\mathcal{O}(\sqrt{p/T})$ convergence rate as their centralized counterparts in [22], [32]?

(Q2) As shown in [63], distributed stochastic gradient descent (SGD) algorithms can achieve linear speedup with respect to the number of agents compared with centralized SGD algorithms. Can distributed ZO algorithms also achieve linear speedup? In particular, can distributed ZO algorithms based on two-point sampling achieve the linear speedup convergence rate $\mathcal{O}(\sqrt{p/nT})$?

(Q3) For deterministic optimization problems, centralized and distributed ZO algorithms can achieve faster convergence rates under more stringent conditions such as strong convexity or Polyak–Łojasiewicz (P–Ł) conditions, as shown in [20], [26], [28], [42], [44], [46] and [57], [61], respectively. For stochastic optimization problems, can ZO algorithms also achieve faster convergence rates under strong convexity or P–Ł conditions?

B. Main Contributions

This paper provides positive answers to the above three questions. More specifically, the contributions of this paper are summarized as follows.

(C1) We propose two distributed ZO algorithms, one primal–dual and one primal algorithm, to solve the stochastic optimization problem (1). In both algorithms, at each iteration each agent communicates its local primal variables to its neighbors through an arbitrarily connected communication network. Moreover, each agent samples its local stochastic ZO oracle at two points with an adaptive smoothing parameter.

(C2) We show in Theorems 2 and 8 that our algorithms find a stationary point with the linear speedup convergence rate $\mathcal{O}(\sqrt{p/(nT)})$ for general nonconvex cost functions. This rate is smaller than that achieved by the centralized ZO algorithms in [22], [24], [29]–[32], [50] and the distributed ZO algorithm in [57]. To the best of our knowledge, this is the first linear speedup result for distributed ZO algorithms; thus (Q1) and (Q2) are answered.

(C3) We show in Theorems 4, 5, 10, and 11 that our proposed algorithms find a global optimum with an $\mathcal{O}(p/(nT))$ convergence rate when the global cost function satisfies the P–Ł condition. This rate is smaller than that achieved by the centralized ZO algorithms in [21], [23] and the distributed ZO algorithms in [54], [57], even though [21], [23], [54] assumed strongly convex cost functions and only considered additive sampling noise, and [57] only considered the deterministic problem. This paper presents the first performance analysis for ZO algorithms to solve stochastic optimization problems under P–Ł or strong convexity assumptions; thus (Q3) is answered.

(C4) When considering deterministic centralized optimization problems, we show in Theorems 6 and 12 that our algorithms linearly find a global optimum under the P–Ł condition. Compared with [20], [26], [28], [42], [44], [46] which also achieved linear convergence, we use weaker assumptions on the cost function and/or less samplings per iteration.

The detailed comparison between this paper and the literature is summarized in TABLE I.

TABLE I: Summary of existing works on ZO optimization, where NoSPPI denotes the number of sampled points per iteration, and the sampling complexity is the total number of function samplings to achieve $\mathbf{E}[\|\nabla f(x_T)\|^2] \le \epsilon$ for nonconvex problems or $\mathbf{E}[f(x_T) - f^*] \le \epsilon$ for (strongly) convex problems or problems satisfying the P–Ł condition.

Reference	Problem settings	NoSPPI	Convergence rate	Sampling complexity
[20]	Deterministic, centralized, unconstrained, nonconvex, smooth	Two	$\mathcal{O}(p/T)$	$\mathcal{O}(p/\epsilon)$
	Strongly convex in addition		Linear	$\mathcal{O}(p\log(1/\epsilon))$
[26]	Deterministic, centralized, strongly convex, unconstrained, smooth, Lipschitz Hessian	p	Linear	$\mathcal{O}(p\log(1/\epsilon))$
[28]	Deterministic, centralized, unconstrained, nonconvex, smooth	Two	$\mathcal{O}(p/T)$	$\mathcal{O}(p/\epsilon)$
	P–Ł condition in addition		Linear	$\mathcal{O}(p\log(1/\epsilon))$
[46]	Deterministic, centralized, restricted strongly convex, unconstrained, smooth, s-sparse gradient	$4s\log(p/s)$	Linear	$\mathcal{O}(s\log(p/s)\log(1/\epsilon))$
[21]	Deterministic, centralized, quadratic, unconstrained, additive sampling noise	One	$\mathcal{O}(p^2/T)$	$\mathcal{O}(p^2/\epsilon)$
[23]	Deterministic, centralized, strongly convex, unconstrained, smooth, additive sampling noise	Two	$\mathcal{O}(p/\sqrt{T})$	$\mathcal{O}(p^2/\epsilon^2)$
[31]	Deterministic, centralized, unconstrained, nonconvex, Lipschitz, smooth	One	$\mathcal{O}(p^2/T^{2/3})$	$\mathcal{O}(p^3/\epsilon^{3/2})$
	Stochastic, centralized, unconstrained, nonconvex, Lipschitz, smooth		$\mathcal{O}(p^{4/3}/T^{1/3})$	$\mathcal{O}(p^4/\epsilon^3)$
[22], [32]	Stochastic, centralized, unconstrained, nonconvex, smooth	Two	$\mathcal{O}(\sqrt{p/T})$	$\mathcal{O}(p/\epsilon^2)$
[24]	Stochastic, centralized, unconstrained, nonconvex, smooth, s-sparse gradient	Two	$\mathcal{O}(s\log(p)/\sqrt{T})$	$\mathcal{O}((s\log(p))^2/\epsilon^2)$
[33]	Stochastic, centralized, constrained, nonconvex, Lipschitz, smooth	$\mathcal{O}(pT)$	$\mathcal{O}(1/T)$	$\mathcal{O}(p/\epsilon^2)$
[50]	Stochastic, centralized, constrained, nonconvex, Lipschitz, smooth	Two	$\mathcal{O}(p/\sqrt{T})$	$\mathcal{O}(p^2/\epsilon^2)$
[29]	Deterministic, finite-sum, nonconvex, constrained, Lipschitz, smooth	$\mathcal{O}(\sqrt{T})$	$\mathcal{O}(p/\sqrt{T})$	$\mathcal{O}(p^3/\epsilon^3)$
[30]	Deterministic, finite-sum, nonconvex, unconstrained, Lipschitz, smooth	$\mathcal{O}(pT)$	$\mathcal{O}(\sqrt{p/T})$	$\mathcal{O}(p^3/\epsilon^4)$
[36]	Deterministic, finite-sum, nonconvex, unconstrained, smooth, the original and mixture gradients are proportional	$2b$	$\mathcal{O}(pn^\theta/(bT))$	$\mathcal{O}(pn^\theta/\epsilon)$, $\forall\theta \in (0,1)$
[37]	Deterministic, finite-sum, nonconvex, unconstrained, smooth	$\mathcal{O}(pn^{1/2})$	$\mathcal{O}(1/T)$	$\mathcal{O}(pn^{1/2}/\epsilon)$
[38]	Deterministic, finite-sum, nonconvex, unconstrained, smooth, similar f_i	$2n$	$\mathcal{O}(p/T)$	$\mathcal{O}(pn/\epsilon)$
[41]	Deterministic, finite-sum, nonconvex, unconstrained, Lipschitz, smooth	$\mathcal{O}(pn^{2/3})$	$\mathcal{O}(p/T)$	$\mathcal{O}(p^2n^{2/3}/\epsilon)$
[42]	Deterministic, finite-sum, nonconvex, unconstrained, smooth, similar f_i	$\mathcal{O}(pn^{1/2})$	$\mathcal{O}(1/T)$	$\mathcal{O}(pn^{1/2}/\epsilon)$
	P–Ł condition in addition		Linear	$\mathcal{O}(pn^{1/2}\log(1/\epsilon))$
[44]	Deterministic, finite-sum, strongly convex, unconstrained, smooth	Four	Linear	$\mathcal{O}(pn\log(p/\epsilon))$
[35]	Stochastic, finite-sum, nonconvex, constrained, Lipschitz, smooth	$\mathcal{O}(nT)$	$\mathcal{O}(p/T)$	$\mathcal{O}(p^2n/\epsilon^2)$
[40]	Stochastic, finite-sum, nonconvex, unconstrained, smooth	Four	$\mathcal{O}(p^{1/3}n^{2/3}/T)$	$\mathcal{O}(p^{1/3}n^{2/3}/\epsilon)$
[53]	Deterministic, distributed, convex, constrained, Lipschitz	$2n$	Asymptotic	—-
[58]	Deterministic, distributed, convex, unconstrained, Lipschitz	$2n$	$\mathcal{O}(p^3n^2/\sqrt{T})$	$\mathcal{O}(p^6n^5/\epsilon^2)$
[59]	Deterministic, distributed, convex, compact constrained, Lipschitz	$2n$	$\mathcal{O}(p\sqrt{n/T})$	$\mathcal{O}(p^2n^2/\epsilon^2)$
	Deterministic, distributed, strongly convex, constrained, Lipschitz		$\mathcal{O}(p^2n^2/T)$	$\mathcal{O}(p^2n^3/\epsilon)$
[54]	Deterministic, distributed, strongly convex, unconstrained, smooth, additive sampling noise	$2n$	$\mathbb{O}(pn^2/\sqrt{T})$	$\mathcal{O}(p^2n^5/\epsilon^2)$
[55]	Deterministic, distributed, convex, compact constrained, Lipschitz, additive sampling noise	$2n$	$\mathcal{O}(1/\sqrt{T})$	$\mathcal{O}(n/\epsilon^2)$
[62]	Stochastic, distributed, convex, compact constrained, Lipschitz	$\mathcal{O}(pnT)$	$\mathcal{O}(1/T)$	$\mathcal{O}(pn/\epsilon^2)$
[57]	Deterministic, distributed, nonconvex, unconstrained, Lipschitz, smooth	$2n$	$\mathcal{O}(\sqrt{p/T})$	$\mathcal{O}(pn/\epsilon^2)$
	Deterministic, distributed, nonconvex, unconstrained, smooth, P–Ł condition		$\mathcal{O}(p/T)$	$\mathcal{O}(pn/\epsilon)$
	Deterministic, distributed, nonconvex, unconstrained, smooth	$2pn$	$\mathcal{O}(1/T)$	$\mathcal{O}(pn/\epsilon)$
	Deterministic, distributed, nonconvex, unconstrained, smooth, P–Ł condition		Linear	$\mathcal{O}(pn\log(1/\epsilon))$
[61]	Deterministic, distributed, nonconvex, unconstrained, smooth	$(p+1)n$	$\mathcal{O}(1/T)$	$\mathcal{O}(pn/\epsilon)$
	P–Ł condition in addition (without using the P–Ł constant)		Linear	$\mathcal{O}(pn\log(1/\epsilon))$
[60]	Stochastic, distributed, nonconvex, unconstrained, Lipschitz, smooth	$\mathcal{O}(nT)$	$\mathcal{O}(p^2n/T)$	$\mathcal{O}(p^4n^3/\epsilon^2)$
This paper	Stochastic, distributed, nonconvex, unconstrained, smooth, similar f_i	$2n$	$\mathcal{O}(\sqrt{p/(nT)})$	$\mathcal{O}(p/\epsilon^2)$
	Stochastic, distributed, nonconvex, unconstrained, smooth, P–Ł condition		$\mathcal{O}(p/(nT))$	$\mathcal{O}(p/\epsilon)$
	Deterministic, centralized, nonconvex, unconstrained, smooth,	Two	$\mathcal{O}(p/T)$	$\mathcal{O}(p/\epsilon)$
	P–Ł condition in addition (without using the P–Ł constant)		Linear	$\mathcal{O}(p\log(1/\epsilon))$

C. Outline

The rest of this paper is organized as follows. Section II introduces some preliminaries. Sections III and IV provide the distributed primal–dual and primal ZO algorithms, respectively, and analyze their convergence properties. Numerical evaluations for an image classification problem from the literature are given in Section V. Finally, concluding remarks are offered in Section VI. All the proofs are given in Appendix.

Notations: $\mathbb{N}_+$ denotes the set of positive integers. $[n]$ denotes the set $\{1,\dots,n\}$ for any $n\in\mathbb{N}_+$. $\|\cdot\|$ represents the Euclidean norm for vectors or the induced 2-norm for matrices. $\mathbb{B}^p$ and $\mathbb{S}^p$ are the unit ball and sphere centered around the origin in $\mathbb{R}^p$ under Euclidean norm, respectively. Given a differentiable function f, ∇f denotes its gradient.

II. Preliminaries

In this section, we introduce the P–Ł condition, the random gradient estimator, and the assumptions used in this paper.

A. Polyak–Łojasiewicz Condition

Definition 1. *[64] A differentiable function $f(x):\ \mathbb{R}^p\mapsto\mathbb{R}$ satisfies the Polyak–Łojasiewicz (P–Ł) condition with constant $\nu>0$ if $f^*>-\infty$ and*

$$\frac{1}{2}\|\nabla f(x)\|^2\ge\nu(f(x)-f^*),\ \forall x\in\mathbb{R}^p. \tag{2}$$

It is straightforward to see that every (essentially, weakly, or restricted) strongly convex function satisfies the P–Ł condition. The P–Ł condition implies that every stationary point is a global minimizer. But unlike (essentially, weakly, or restricted) strong convexity, the P–Ł condition alone does not imply convexity of f. Moreover, it does not imply that the set of global minimizers is a singleton [64], [65]. Examples of nonconvex functions which satisfy the P–Ł condition can be found in [64], [65].

B. Gradient Approximation

Let $f(x):\ \mathbb{R}^p\mapsto\mathbb{R}$ be a function. The authors of [18] proposed the following random gradient estimator:

$$\hat{\nabla}_2 f(x,\delta,u)=\frac{p}{\delta}(f(x+\delta u)-f(x))u, \tag{3}$$

where $\delta > 0$ is the smoothing/exploration parameter and $u \in \mathbb{S}^p$ is a uniformly distributed random vector. This gradient estimator can be calculated by sampling the function f at two points (e.g., x and $x + \delta u$). The intuition of this estimator follows from directional derivatives [18]. From a practical point of view, the larger the smoothing parameter δ the better, since in this case it is easier to distinguish the two sampled function values.

C. Assumptions

The following assumptions are made.

Assumption 1. *The undirected graph $\mathcal{G}$ is connected.*

Assumption 2. *The optimal set $\mathbb{X}^*$ is nonempty and $f^* > -\infty$, where $\mathbb{X}^*$ and f^* are the optimal set and the minimum function value of the optimization problem* (1)*, respectively.*

Assumption 3. *For almost all ξ_i, the stochastic ZO oracle $F_i(\cdot, \xi_i)$ is smooth with constant $L_f > 0$.*

Assumption 4. *The stochastic gradient $\nabla_x F_i(x, \xi_i)$ has bounded variance, i.e., there exists $\sigma_1 \in \mathbb{R}$ such that $\mathbf{E}_{\xi_i}[\|\nabla_x F_i(x, \xi_i) - \nabla f_i(x)\|^2] \le \sigma_1^2, \ \forall i \in [n], \ \forall x \in \mathbb{R}^p$.*

Assumption 5. *Local cost functions are similar, i.e., there exists $\sigma_2 \in \mathbb{R}$ such that $\|\nabla f_i(x) - \nabla f(x)\|^2 \le \sigma_2^2, \ \forall i \in [n], \ \forall x \in \mathbb{R}^p$.*

Assumption 6. *The global cost function $f(x)$ satisfies the P–Ł condition with constant $\nu > 0$.*

Remark 1. *It should be highlighted that no convexity assumptions are made. Assumption 1 is common in distributed optimization, e.g., [57], [60], [62], [66]–[69]. Assumption 2 is basic. Assumptions 3 and 4 are standard in stochastic optimization with ZO information feedback, e.g., [22], [24], [32]–[35], [39], [40], [47], [48], [60]. Assumption 5 is slightly weaker than the assumption that each ∇f_i is bounded, which is normally used in the literature studying finite-sum ZO optimization, e.g., [18], [29], [30], [34]–[36], [41], [48], [51]–[53], [57], [58], [60], [62]. Bounded gradient does not hold for many unconstrained optimization problems, e.g., quadratic optimization problems. Assumption 5 is not needed when Assumption 6 holds and the constant ν is known in advance as shown in Theorems 5 and 11. Assumption 6 is weaker than the assumption that the global or each local cost function is (restricted) strongly convex.*

It plays a key role to guarantee that a global optimum can be found and to show that smaller convergence rate can be achieved.

III. Distributed ZO primal–dual Algorithm

In this section, we propose a distributed ZO primal–dual algorithm and analyze its convergence properties.

When gradient information is available, in [61] the following distributed first-order primal–dual algorithm was proposed to solve (1):

$$x_{i,k+1} = x_{i,k} - \eta\Big(\alpha\sum_{j=1}^{n} L_{ij}x_{j,k} + \beta v_{i,k} + \nabla f_i(x_{i,k})\Big), \tag{4a}$$

$$v_{i,k+1} = v_{i,k} + \eta\beta\sum_{j=1}^{n} L_{ij}x_{j,k},\ \forall x_{i,0}\in\mathbb{R}^p,\ \sum_{j=1}^{n} v_{j,0} = \mathbf{0}_p,\ \forall i\in[n], \tag{4b}$$

where α, β, and η are positive algorithm parameters, and $L = [L_{ij}]$ is the weighted Laplacian matrix associated with the undirected communication graph $\mathcal{G}$. It has been shown in [61] that the distributed first-order algorithm (4) can find a stationary point with an $\mathcal{O}(1/k)$ convergence rate.

Noting that we consider the scenario where only stochastic ZO oracles rather than the explicit expressions of the gradients are available, we need to estimate the gradients used in the distributed first-order algorithm (4). Inspired by (3), we introduce

$$g^e_{i,k} = \frac{p(F_i(x_{i,k}+\delta_{i,k}u_{i,k},\xi_{i,k}) - F_i(x_{i,k},\xi_{i,k}))}{\delta_{i,k}}u_{i,k}, \tag{5}$$

where $\delta_{i,k} > 0$ is an adaptive smoothing parameter and $u_{i,k}\in\mathbb{S}^p$ is a uniformly distributed random vector chosen by agent i at iteration k; $\xi_{i,k}$ is a random variable sampled by agent i at iteration k according to the distribution of ξ_i; and $F_i(x_{i,k}+\delta_{i,k}u_{i,k},\xi_{i,k})$ and $F_i(x_{i,k},\xi_{i,k})$ are the values sampled by agent i at iteration k. We replace the gradient and fixed algorithm parameters in (4) with the stochastic gradient estimator (5) and time-varying parameters, respectively. Then we get the following ZO algorithm:

$$x_{i,k+1} = x_{i,k} - \eta_k\Big(\alpha_k\sum_{j=1}^{n} L_{ij}x_{j,k} + \beta_k v_{i,k} + g^e_{i,k}\Big), \tag{6a}$$

$$v_{i,k+1} = v_{i,k} + \eta_k\beta_k\sum_{j=1}^{n} L_{ij}x_{j,k},\ \forall x_{i,0}\in\mathbb{R}^p,\ \sum_{j=1}^{n} v_{j,0} = \mathbf{0}_p,\ \forall i\in[n]. \tag{6b}$$

Algorithm 1 Distributed ZO Primal–Dual Algorithm

```
Input: positive sequences $\{\alpha_k\}$, $\{\beta_k\}$, $\{\eta_k\}$, and $\{\delta_{i,k}\}$.
Initialize: $x_{i,0} \in \mathbb{R}^p$ and $v_{i,0} = \mathbf{0}_p,\ \forall i \in [n]$.
for $k = 0, 1, \dots$ do
    for $i = 1, \dots, n$ in parallel do
        Broadcast $x_{i,k}$ to $\mathcal{N}_i$ and receive $x_{j,k}$ from $j \in \mathcal{N}_i$;
        Select vector $u_{i,k} \in \mathbb{S}^p$ independently and uniformly at random;
        Select $\xi_{i,k}$ independently;
        Sample $F_i(x_{i,k}, \xi_{i,k})$ and $F_i(x_{i,k} + \delta_{i,k} u_{i,k}, \xi_{i,k})$;
        Update $x_{i,k+1}$ by (6a);
        Update $v_{i,k+1}$ by (6b).
    end for
end for
Output: $\{\boldsymbol{x}_k\}$.
```

Here, we assume that $u_{i,k}$ and $\xi_{i,k},\ \forall i \in [n], k \geq 1$ are mutually independent, which is commonly assumed in stochastic optimization, e.g., [18], [22], [24], [32]–[35], [39], [40], [47], [48], [50], [54], [55], [60]. Let $\mathfrak{L}_k$ denote the σ-algebra generated by the independent random variables $u_{1,k}, \dots, u_{n,k}, \xi_{1,k}, \dots, \xi_{n,k}$ and let $\mathcal{L}_k = \bigcup_{t=0}^{k} \mathfrak{L}_t$. It is straightforward to see that $x_{i,k}$ and $v_{i,k+1},\ i \in [p]$ depend on $\mathcal{L}_{k-1}$ and are independent of $\mathfrak{L}_t$ for all $t \geq k$.

We write the distributed ZO algorithm (6) in pseudo-code as Algorithm 1.

Remark 2. *In Algorithm 1, each agent i maintains two local sequences, i.e., the local primal and dual variable sequences $\{x_{i,k}\}$ and $\{v_{i,k}\}$, and communicates its local primal variables to its neighbors through the network. Moreover, at each iteration each agent samples its local stochastic ZO oracle at two points to estimate the gradient of its local cost function. It should be highlighted that the agent-wise smoothing parameter $\delta_{i,k}$ is adaptive. It can in many situations be chosen larger than the fixed smoothing parameter used in existing ZO algorithms. For example, in the following we use an $\mathcal{O}(1/k^{1/4})$ smoothing parameter, which is larger than the $\mathcal{O}(1/T^{1/2})$ smoothing parameter used in [22].*

A. Find stationary points

Let us consider the case when Algorithm 1 is able to find stationary points. We first have the following convergence result.

Theorem 1. *Suppose Assumptions 1–5 hold. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1 with*

$$\alpha_k = \kappa_1\beta_k,\ \beta_k = \kappa_0(k+t_1)^{\theta},\ \eta_k = \frac{\kappa_2}{\beta_k},\ \delta_{i,k} \le \kappa_\delta\sqrt{\eta_k},\ \forall k \in \mathbb{N}_0, \tag{7}$$

where $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, $\theta \in (0.5, 1)$, $t_1 \ge (\sqrt{p}c_3(\kappa_1,\kappa_2))^{1/\theta}$, $\kappa_0 \ge c_0(\kappa_1,\kappa_2)/t_1^{\theta}$, and $\kappa_\delta > 0$ with $c_0(\kappa_1,\kappa_2)$, c_1, $c_2(\kappa_1)$, and $c_3(\kappa_1,\kappa_2)$ being given in Appendix B. Then, for any $T \in \mathbb{N}_+$,

$$\frac{\sum_{k=0}^{T-1}\eta_k\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2]}{\sum_{k=0}^{T-1}\eta_k} = \mathcal{O}(\frac{\sqrt{p}}{T^{1-\theta}}), \tag{8a}$$

$$\mathbf{E}[f(\bar{x}_T)] - f^* = \mathcal{O}(1), \tag{8b}$$

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,T} - \bar{x}_T\|^2\Big] = \mathcal{O}(\frac{1}{T^{2\theta}}), \tag{8c}$$

$$\lim_{T\to+\infty}\mathbf{E}[\|\nabla f(\bar{x}_T)\|^2] = 0, \tag{8d}$$

where $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n}x_{i,k}$.

Proof: The explicit expressions of the right-hand sides of (8a)–(8c) and the proof are given in Appendix B. ■

If the total number of iterations T and the number of agents n are known in advance, then, as shown in the following, Algorithm 1 can find a stationary point of (1) with an $\mathcal{O}(\sqrt{p/(nT)})$ convergence rate, and thus achieves linear speedup with respect to the number of agents compared to the $\mathcal{O}(\sqrt{p/T})$ convergence rate achieved by the centralized stochastic ZO algorithms in [22], [32]. The linear speedup property enables us to scale up the computing capability by adding more agents into the algorithm [70].

Theorem 2 (Linear speedup)**.** *Suppose Assumptions 1–5 hold. For any given $T > \max\{n(\tilde{c}_0(\kappa_1,\kappa_2)/\kappa_2)^2, n^3\}/p$, let $\{\boldsymbol{x}_k, k = 0,\dots,T\}$ be the output generated by Algorithm 1 with*

$$\alpha_k = \kappa_1\beta_k,\ \beta_k = \beta = \frac{\kappa_2\sqrt{pT}}{\sqrt{n}},\ \eta_k = \frac{\kappa_2}{\beta_k},\ \delta_{i,k} \le \frac{\kappa_\delta}{p^{\frac{1}{4}}n^{\frac{1}{4}}(k+1)^{\frac{1}{4}}},\ \forall k \le T, \tag{9}$$

where $\tilde{c}_0(\kappa_1,\kappa_2)$ is given in Appendix C, $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, and $\kappa_\delta > 0$ with c_1 and

$c_2(\kappa_1)$ *being given in Appendix B. Then,*

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] = \mathcal{O}(\frac{\sqrt{p}}{\sqrt{nT}}) + \mathcal{O}(\frac{n}{T}), \tag{10a}$$

$$\mathbf{E}[f(\bar{x}_T)] - f^* = \mathcal{O}(1), \tag{10b}$$

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2\Big] = \mathcal{O}(\frac{n}{T}), \tag{10c}$$

$$\lim_{T\to+\infty}\mathbf{E}[\|\nabla f(\bar{x}_T)\|^2] = 0. \tag{10d}$$

Proof: The explicit expressions of the right-hand sides of (10a)–(10b) and the proof are given in Appendix C. It should be highlighted that the omitted constants in the first term in the right-hand side of (10a) do not depend on any parameters related to the communication network. ■

Remark 3. *To the best of our knowledge, Theorem 2 is the first result to establish linear speedup for a distributed ZO algorithm to solve stochastic optimization problems. The achieved rate is smaller than that achieved by centralized ZO algorithms [22], [24], [29]–[32], [50] and distributed ZO gradient descent algorithm [57]. The rate is greater than that achieved by centralized ZO algorithms in [33], [35]–[38], [40]–[42], which is reasonable since these algorithms not only are centralized but also use variance reduction techniques. The distributed ZO gradient tracking algorithm in [57] and the distributed ZO primal–dual algorithms in [60], [61] also achieved smaller convergence rates than ours. However, in [36]–[38], [41], [42], [57], [61], the considered problems are deterministic; in [57], [61], the sampling size of each agent at each iteration is $\mathcal{O}(p)$, which results in a heavy sampling burden when p is large; in [33], [35], [60], the sampling size of each agent at each iteration is $\mathcal{O}(T)$, which is difficult to execute in practice. One of our future research directions is to establish faster convergence with reduced sampling complexity by using variance reduction techniques.*

B. Find global optimum

Let us next consider cases when Algorithm 1 finds global optimum.

Theorem 3. *Suppose Assumptions 1–6 hold. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1 with*

$$\alpha_k = \kappa_1\beta_k,\ \beta_k = \kappa_0(k+t_1)^\theta,\ \eta_k = \frac{\kappa_2}{\beta_k},\ \delta_{i,k} \le \kappa_\delta\eta_k,\ \forall k\in\mathbb{N}_0, \tag{11}$$

where $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, $\theta \in (0,1)$, $t_1 \in [(pc_3(\kappa_1,\kappa_2))^{1/\theta}, (pc_4c_3(\kappa_1,\kappa_2))^{1/\theta}]$, $\kappa_0 \geq c_0(\kappa_1,\kappa_2)/t_1^{\theta}$, $\kappa_\delta > 0$, *and* $c_4 \geq 1$ *with* $c_0(\kappa_1,\kappa_2)$, c_1, $c_2(\kappa_1)$, *and* $c_3(\kappa_1,\kappa_2)$ *being given in Appendix B. Then, for any* $T \in \mathbb{N}_+$,

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,T} - \bar{x}_T\|^2\Big] = \mathcal{O}(\frac{p}{T^{2\theta}}), \tag{12a}$$

$$\mathbf{E}[f(\bar{x}_T) - f^*] = \mathcal{O}(\frac{p}{nT^{\theta}}) + \mathcal{O}(\frac{p}{T^{2\theta}}). \tag{12b}$$

Proof: The explicit expressions of the right-hand sides of (12a) and (12b), and the proof are given in Appendix D. It should be highlighted that the omitted constants in the first term in the right-hand side of (12b) do not depend on any parameters related to the communication network. ■

From Theorem 3, we see that the convergence rate is strictly greater than $\mathcal{O}(p/(nT))$. In the following we show that the $\mathcal{O}(p/(nT))$ convergence rate can be achieved if the P–Ł constant ν is known in advance. Information about the total number of iterations T is not needed.

Theorem 4 (Linear speedup)**.** *Suppose Assumptions 1–6 hold and the P–Ł constant* ν *is known in advance. Let* $\{\boldsymbol{x}_k\}$ *be the sequence generated by Algorithm 1 with*

$$\alpha_k = \kappa_1\beta_k,\ \beta_k = \kappa_0(k + t_1),\ \eta_k = \frac{\kappa_2}{\beta_k},\ \delta_{i,k} \leq \kappa_\delta\eta_k,\ \forall k \in \mathbb{N}_0, \tag{13}$$

where $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, $\kappa_0 \in [3\hat{c}_0\nu\kappa_2/16, 3\nu\kappa_2/16)$, $t_1 > \hat{c}_3(\kappa_0,\kappa_1,\kappa_2)$, $\kappa_\delta > 0$, *and* $\hat{c}_0 \in (0,1)$ *with* c_1 *and* $c_2(\kappa_1)$ *being given in Appendix B, and* $\hat{c}_3(\kappa_0,\kappa_1,\kappa_2)$ *being given in Appendix E. Then, for any* $T \in \mathbb{N}_+$,

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,T} - \bar{x}_T\|^2\Big] = \mathcal{O}(\frac{p}{T^{2}}), \tag{14a}$$

$$\mathbf{E}[f(\bar{x}_T) - f^*] = \mathcal{O}(\frac{p}{nT}) + \mathcal{O}(\frac{p}{T^{2}}). \tag{14b}$$

Proof: The explicit expressions of the right-hand sides of (14a) and (14b), and the proof are given in Appendix E. It should be highlighted that the omitted constants in the first term in the right-hand side of (14b) do not depend on any parameters related to the communication network. ■

Although Assumption 5 is weaker than the bounded gradient assumption, it can be further relaxed by a mild assumption. Specifically, if each $f_i^* > -\infty$, where $f_i^* = \min_{x\in\mathbb{R}^p} f_i(x)$, then without Assumption 5, the convergence results stated in (14a) and (14b) still hold, as shown in the following.

Theorem 5 (Linear speedup)**.** *Suppose Assumptions 1–4 and 6 hold, and the P–Ł constant ν is known in advance, and each $f_i^* > -\infty$. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1 with*

$$\alpha_k = \kappa_1\beta_k,\ \beta_k = \kappa_0(k+t_1),\ \eta_k = \frac{\kappa_2}{\beta_k},\ \delta_{i,k} \le \kappa_\delta\eta_k,\ \forall k \in \mathbb{N}_0, \tag{15}$$

where $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, $\kappa_0 \in [3\hat{c}_0\nu\kappa_2/16, 3\nu\kappa_2/16)$, $t_1 > \check{c}_3(\kappa_0,\kappa_1,\kappa_2)$, $\kappa_\delta > 0$, and $\hat{c}_0 \in (0,1)$ with c_1 and $c_2(\kappa_1)$ being given in Appendix B, and $\check{c}_3(\kappa_0,\kappa_1,\kappa_2)$ being given in Appendix F. Then, for any $T \in \mathbb{N}_+$,

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,T} - \bar{x}_T\|^2\Big] = \mathcal{O}(\frac{p}{T^2}), \tag{16a}$$

$$\mathbf{E}[f(\bar{x}_T) - f^*] = \mathcal{O}(\frac{p}{nT}) + \mathcal{O}(\frac{p}{T^2}). \tag{16b}$$

Proof: The explicit expressions of the right-hand sides of (16a) and (16b), and the proof are given in Appendix F. It should be highlighted that the omitted constants in the first term in the right-hand side of (16b) do not depend on any parameters related to the communication network.

■

Remark 4. *To the best of our knowledge, Theorems 3–5 are the first performance analysis results for ZO algorithms to solve stochastic optimization problems under the P–Ł condition or strong convexity assumption. In [21], a centralized ZO algorithm based on one-point sampling with additive sampling noise was proposed and an $\mathcal{O}(p^2/T)$ convergence rate was achieved for deterministic optimization problems with strongly convex quadratic cost functions. In [23], a centralized ZO algorithm based on two-point sampling with additive noise was proposed and an $\mathcal{O}(p/\sqrt{T})$ convergence rate was achieved for deterministic strongly convex and smooth optimization problems. In [54], a distributed ZO gradient descent algorithm based on $2p$-point sampling with additive noise was proposed and an $\mathcal{O}(pn^2/\sqrt{T})$ convergence rate was achieved for deterministic strongly convex and smooth optimization problems. In [57], a distributed ZO gradient descent algorithm based on two-point sampling was proposed and an $\mathcal{O}(p/T)$ convergence rate was achieved for deterministic smooth optimization problems under the P–*

Ł condition. It is straightforward to see that aforementioned convergence rates achieved in [21], [23], [54], [57] are greater than that achieved by our distributed stochastic ZO primal–dual algorithm as stated in Theorem 5. Moreover, we consider stochastic optimization problems and use the P–Ł condition, which is slightly weaker than the strong convexity condition. The distributed ZO gradient tracking algorithm in [57] and the distributed ZO primal–dual algorithms in [61] achieved linear convergence under the P–Ł condition. However, both algorithms require each agent at each iteration to sample $\mathcal{O}(p)$ points, which results in a heavy sampling burden when p is large.

As shown in Theorems 3–5, in expectation, the convergence rate of Algorithm 1 with diminishing stepsizes is sublinear. The following theorem establishes that, in expectation, the output of Algorithm 1 with constant algorithm parameters linearly converges to a neighborhood of a global optimum.

Theorem 6 (Linear convergence)**.** *Suppose Assumptions 1–5 hold. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1 with*

$$\alpha_k = \alpha = \kappa_1\beta,\ \beta_k = \beta,\ \eta_k = \eta = \frac{\kappa_2}{\beta},\ \delta_{i,k} \le \kappa_\delta \hat{\varepsilon}^{\frac{k}{2}},\ \forall k \in \mathbb{N}_0, \tag{17}$$

where $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, $\beta \ge \tilde{c}_0(\kappa_1, \kappa_2)$, $\hat{\varepsilon} \in (0,1)$, and $\kappa_\delta > 0$ with $\tilde{c}_0(\kappa_1,\kappa_2)$ being given in Appendix C, and c_1 and $c_2(\kappa_1)$ being given in Appendix B. Then, for any $T \in \mathbb{N}_+$,

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2\Big] \le \frac{c_5}{T} + \eta^2(\sigma_1^2 + 3\sigma_2^2)c_6, \tag{18a}$$

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] \le \frac{pc_7}{\eta T} + \eta(\sigma_1^2 + 3\eta\sigma_2^2)c_8, \tag{18b}$$

where c_5, c_6, c_7, and c_8 are positive constants given in Appendix G. Moreover, if Assumption 6 also holds, then

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^*\Big] \le \varepsilon^k c_9 + \eta(\sigma_1^2 + 3\sigma_2^2)c_{10},\ \forall k \in \mathbb{N}_+, \tag{19}$$

where $\varepsilon \in (0,1)$, c_9, and c_{10} are positive constants given in Appendix G.

Proof: The proof is given in Appendix G. ■

Remark 5. *When considering centralized deterministic nonconvex smooth optimization, i.e.,* $\sigma_1 = \sigma_2 = 0$*, the result stated in* (18b) *shows that a stationary point can be found with a rate* $\mathcal{O}(p/T)$*. This rate is the same as that achieved by the ZO algorithms in [20], [28], [36], [38], [41]. Although the ZO variance reduced algorithms in [37], [42] and the stochastic direct-search algorithms in [11]–[13] achieved a faster rate* $\mathcal{O}(1/T)$*, these algorithms require three or more samplings at each iteration, while our proposed algorithm requires only two samplings. Moreover, the result stated in* (19) *shows that a global optimum can be found linearly. The ZO algorithms in [20], [26], [28], [42], [44], [46] and the stochastic direct-search algorithms in [11]–[14] also achieved linear convergence. However, the algorithms in [11]–[14], [26], [42], [44] require three or more samplings at each iteration; the P–Ł constant needs to be known in advance in [28], [42], which is not needed in Theorem 6; and the cost functions in [11]–[14], [20], [26], [44], [46] are (restricted) strongly convex, which is stronger than the P–Ł condition used in Theorem 6.*

To end this section, we would like to briefly explain the challenges when analyzing the performance of Algorithm 1. Algorithm 1 is simple in the sense that it is a combination of the first-order algorithm proposed in [61] with zeroth-order gradient estimators. For such a kind of combination, the standard technique to handle the bias in the ZO gradients is using smoothing function, which is also used in our proofs. However, there still is a gap between the smoothing function and the original function. This gap leads to that the proof details become complicated, especially under the distributed and stochastic setting. Moreover, to the best of our knowledge, how to show linear speedup for distributed ZO algorithms is an open problem in the literature.

IV. Distributed ZO Primal Algorithm

In this section, we propose a distributed ZO primal algorithm and analyze its convergence rate. Inspired by distributed first-order (sub)gradient descent algorithm proposed in [71], we propose the following distributed ZO primal algorithm:

$$x_{i,k+1} = x_{i,k} - \gamma \sum_{j=1}^{n} L_{ij} x_{j,k} - \eta_k g^e_{i,k}, \tag{20}$$

where γ is a positive constant, $\{\eta_k\}$ is a positive sequence to be specified later, and $g^e_{i,k}$ is the stochastic gradient estimator defined in (5).

Algorithm 2 Distributed ZO Primal Algorithm

```
Input: positive constant $\gamma$ and positive sequences $\{\eta_k\}$ and $\{\delta_{i,k}\}$.
Initialize: $x_{i,0} \in \mathbb{R}^p,\ \forall i \in [n]$.
for $k = 0, 1, \dots$ do
    for $i = 1, \dots, n$ in parallel do
        Broadcast $x_{i,k}$ to $\mathcal{N}_i$ and receive $x_{j,k}$ from $j \in \mathcal{N}_i$;
        Select vector $u_{i,k} \in \mathbb{S}^p$ independently and uniformly at random;
        Select $\xi_{i,k}$ independently;
        Sample $F_i(x_{i,k}, \xi_{i,k})$ and $F_i(x_{i,k} + \delta_{i,k}u_{i,k}, \xi_{i,k})$;
        Update $x_{i,k+1}$ by (20).
    end for
end for
Output: $\{\boldsymbol{x}_k\}$.
```

We write the distributed random ZO algorithm (20) in pseudo-code as Algorithm 2. Compared with Algorithm 1, in Algorithm 2 each agent only computes the primal variable. Similar results as stated in Theorems 1–6 also holds for Algorithm 2.

A. Find stationary points

Theorem 7. *Suppose Assumptions 1–5 hold. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 2 with*

$$\gamma \in (0, d_1),\ \eta_k = \frac{\kappa_\eta}{(k + t_1)^\theta},\ \delta_{i,k} \le \kappa_\delta \sqrt{\eta_k},\ \forall k \in \mathbb{N}_0, \tag{21}$$

where $\kappa_\delta > 0$, $\kappa_\eta \in (0, d_2(\gamma)t_1^\theta]$, $\theta \in (0.5, 1)$, and $t_1 \ge p^{1/(2\theta)}$ with d_1 and $d_2(\gamma)$ being given in Appendix H. Then, for any $T \in \mathbb{N}_+$,

$$\frac{\sum_{k=0}^{T-1} \eta_k \mathbf{E}[\|\nabla f(\bar{x}_k)\|^2]}{\sum_{k=0}^{T-1} \eta_k} = \mathcal{O}(\frac{\sqrt{p}}{T^{1-\theta}}), \tag{22a}$$

$$\mathbf{E}[f(\bar{x}_T)] - f^* = \mathcal{O}(1), \tag{22b}$$

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n} \|x_{i,T} - \bar{x}_T\|^2\Big] = \mathcal{O}(\frac{1}{T^{2\theta}}), \tag{22c}$$

$$\lim_{T \to +\infty} \mathbf{E}[\|\nabla f(\bar{x}_T)\|^2] = 0. \tag{22d}$$

Proof : The explicit expressions of the right-hand sides of (22a)–(22c) and the proof are given in Appendix H. ■

Theorem 8 (Linear speedup). *Suppose Assumptions 1–5 hold. For any given $T \geq \max\{n/d_2^2(\gamma),\ n^3\}/p$, let $\{\boldsymbol{x}_k, k = 0, \dots, T\}$ be the output generated by Algorithm 2 with*

$$\gamma \in (0, d_1),\ \eta_k = \frac{\sqrt{n}}{\sqrt{pT}},\ \delta_{i,k} \leq \frac{\kappa_\delta}{p^{\frac{1}{4}} n^{\frac{1}{4}} (k+1)^{\frac{1}{4}}},\ \forall k \leq T, \tag{23}$$

where $\kappa_\delta > 0$ and d_1 and $d_2(\gamma)$ are given in Appendix H, then

$$\frac{1}{T}\sum_{k=0}^{T-1} \mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] = \mathcal{O}(\frac{\sqrt{p}}{\sqrt{nT}}) + \mathcal{O}(\frac{n}{T}), \tag{24a}$$

$$\mathbf{E}[f(\bar{x}_T)] - f^* = \mathcal{O}(1), \tag{24b}$$

$$\frac{1}{T}\sum_{k=0}^{T-1} \mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2\Big] = \mathcal{O}(\frac{n}{T}), \tag{24c}$$

$$\lim_{T \to +\infty} \mathbf{E}[\|\nabla f(\bar{x}_T)\|^2] = 0. \tag{24d}$$

Proof: The explicit expressions of the right-hand sides of (24a)–(24c) and the proof are given in Appendix I. It should be highlighted that the omitted constants in the first term in the right-hand side of (24a) do not depend on any parameters related to the communication network. ■

B. Find global optimum

Theorem 9. *Suppose Assumptions 1–6 hold. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 2 with*

$$\gamma \in (0, d_1),\ \eta_k = \frac{\kappa_\eta}{(k + t_1)^\theta},\ \delta_{i,k} \leq \kappa_\delta \eta_k,\ \forall k \in \mathbb{N}_0, \tag{25}$$

where $\kappa_\delta > 0$, $\kappa_\eta \in (0, d_2(\gamma) t_1^\theta]$, $\theta \in (0, 1)$, and $t_1 \geq p^{1/\theta}$ with d_1 and $d_2(\gamma)$ being given in Appendix H. Then, for any $T \in \mathbb{N}_+$,

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n} \|x_{i,T} - \bar{x}_T\|^2\Big] = \mathcal{O}(\frac{p}{T^{2\theta}}), \tag{26a}$$

$$\mathbf{E}[f(\bar{x}_T) - f^*] = \mathcal{O}(\frac{p}{nT^\theta}) + \mathcal{O}(\frac{p}{T^{2\theta}}). \tag{26b}$$

Proof: The explicit expressions of the right-hand sides of (26a) and (26b), and the proof are given in Appendix J. It should be highlighted that the omitted constants in the first term in the right-hand side of (26b) do not depend on any parameters related to the communication network. ■

Theorem 10 (Linear speedup)**.** *Suppose Assumptions 1–6 hold and the P–Ł constant ν is known in advance. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 2 with*

$$\gamma \in (0, d_1),\ \eta_k = \frac{\kappa_\eta}{k + t_1},\ \delta_{i,k} \le \kappa_\delta \eta_k,\ \forall k \in \mathbb{N}_0, \tag{27}$$

where $\kappa_\delta > 0$, $\kappa_\eta > 4/\nu$, and $t_1 > \hat{d}_2(\gamma)$ with d_1 and $\hat{d}_2(\gamma)$ being given in Appendices H and K, respectively. Then, for any $T \in \mathbb{N}_+$,

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,T} - \bar{x}_T\|^2\Big] = \mathcal{O}(\frac{p}{T^2}), \tag{28a}$$

$$\mathbf{E}[f(\bar{x}_T) - f^*] = \mathcal{O}(\frac{p}{nT}) + \mathcal{O}(\frac{p}{T^2}). \tag{28b}$$

Proof : The explicit expressions of the right-hand sides of (28a) and (28b), and the proof are given in Appendix K. It should be highlighted that the omitted constants in the first term in the right-hand side of (28b) do not depend on any parameters related to the communication network.

■

Theorem 11 (Linear speedup)**.** *Suppose Assumptions 1–4 and 6 hold, and the P–Ł constant ν is known in advance, and each $f_i^* > -\infty$. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 2 with*

$$\gamma \in (0, d_1),\ \eta_k = \frac{\kappa_\eta}{k + t_1},\ \delta_{i,k} \le \kappa_\delta \eta_k,\ \forall k \in \mathbb{N}_0, \tag{29}$$

where $\kappa_\delta > 0$, $\kappa_\eta > 4/\nu$, and $t_1 > \check{d}_2(\gamma)$ with d_1 and $\check{d}_2(\gamma)$ being given in Appendices H and L, respectively. Then, for any $T \in \mathbb{N}_+$,

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,T} - \bar{x}_T\|^2\Big] = \mathcal{O}(\frac{p}{T^2}), \tag{30a}$$

$$\mathbf{E}[f(\bar{x}_T) - f^*] = \mathcal{O}(\frac{p}{nT}) + \mathcal{O}(\frac{p}{T^2}). \tag{30b}$$

Proof : The explicit expressions of the right-hand sides of (30a) and (30b), and the proof are given in Appendix L. It should be highlighted that the omitted constants in the first term in the right-hand side of (30b) do not depend on any parameters related to the communication network.

■

Theorem 12. *Suppose Assumptions 1–5 hold. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 2* *with*

$$\gamma \in (0, d_1),\ \eta_k = \eta,\ \delta_{i,k} \leq \hat{\epsilon}^{\frac{k}{2}},\ \forall k \in \mathbb{N}_0, \tag{31}$$

where $\eta \in (0, d_2(\gamma)$ *and* $\hat{\epsilon} \in (0, 1)$ *with* d_1 *and* $d_2(\gamma)$ *being given in Appendix H. Then, for any* $T \in \mathbb{N}_+$,

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2\Big] \leq \frac{d_3}{T} + \eta^2(\sigma_1^2 + 3\sigma_2^2)d_4, \tag{32a}$$

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] \leq \frac{pd_5}{\eta T} + \eta(\sigma_1^2 + 3\sigma_2^2)d_6, \tag{32b}$$

where d_3*,* d_4*,* d_5*, and* d_6 *are positive constants given in Appendix M. Moreover, if Assumption 6 also holds, then*

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^*\Big] \leq \epsilon^k d_7 + \eta(\sigma_1^2 + 3\sigma_2^2)d_8,\ \forall k \in \mathbb{N}_+, \tag{33}$$

where $\epsilon \in (0, 1)$*,* d_7*, and* d_8 *are positive constants given in Appendix M.*

Proof: The proof is given in Appendix M. ■

V. Simulations

In this section, we verify the theoretical results through numerical simulations. Specifically, we evaluate the performance of Algorithms 1 and 2 in generating adversarial examples from black-box deep neural networks (DNNs).

In image classification tasks, DNNs are vulnerable to adversarial examples [72] even under small perturbations, which leads misclassifications. Considering the setting of ZO attacks in [38], [73], the model is hidden and no gradient information is available. We treat this task of generating adversarial examples as a ZO optimization problem. The black-box attack loss function [38], [73] is given as

$$\begin{aligned} f_i(x) = \max\Big\{F_{y_i}\Big(\frac{1}{2}\tanh(\tanh^{-1}2a_i + x)\Big) - \max_{j\neq y_i}\Big\{F_j\Big(\frac{1}{2}\tanh(\tanh^{-1}2a_i + x)\Big)\Big\},\ 0\Big\} \\ + c\Big\|\frac{1}{2}\tanh(\tanh^{-1}2a_i + x) - a_i\Big\|_2^2, \end{aligned}$$

where c is a constant, (a_i, y_i) denotes the pair of the ith natural image a_i and its original class label y_i. The output of function $F(z) = \mathrm{col}(F_1(z), \dots, F_m(z))$ is the well-trained model prediction of the input z in all m image classes.

The well-trained DNN model[1] on the MNIST handwritten dataset has 99.4% test accuracy on natural examples [38]. We compare the proposed distributed primal–dual ZO algorithm (Algorithm 1) and distributed primal ZO algorithm (Algorithm 2) with state-of-the-art centralized and distributed ZO algorithms: RSGF [22], SZO-SPIDER [37], ZO-SVRG [38], SZVR-G [40], and ZO-SPIDER-Coord [42], ZO-GDA [57], and ZONE-M [60].

We consider $n = 10$ agents and assume the communication network is generated randomly following the Erdős–Rényi model with probability of 0.4. All the hyper-parameters used in the experiment are given in TABLE II.

TABLE II: Parameters in each algorithm.

Algorithm	Distributed	Parameters
Algorithm 1	✔	$\eta = 0.5/k^{10^{-5}}$, $\alpha = 0.5k^{10^{-5}}$, $\beta = 0.1k^{10^{-5}}$
Algorithm 2	✔	$\gamma = 0.01$, $\eta = 0.08/k^{10^{-5}}$
ZO-GDA	✔	$\eta = 0.08/k^{10^{-5}}$
ZONE-M	✔	$\rho = 0.1\sqrt{k}$
RSGF	✗	$\mu = 0.01$
SZO-SPIDER	✗	$\mu = 0.01$
ZO-SVRG	✗	$\mu = 0.01$
SZVR-G	✗	$\mu = 0.01$
ZO-SPIDER-Coord	✗	$\mu = 0.01$

Fig. 1 and Fig. 2 show the evolutions of the black-box attack loss achieved by each ZO algorithm with respect to the number of iterations and function value queries, respectively. From these two figures, we can see that our proposed distributed ZO algorithms are as efficient as ZO-GDA [57] in terms of both convergence rate and sampling complexity, and more efficient than the other algorithms. The least ℓ_2 distortions of the successful adversarial perturbations are listed in TABLE III. We can see that the adversarial examples generated by the distributed algorithms in general have slightly larger ℓ_2 distortions than those generated by the centralized

[1] https://github.com/carlini/nn_robust_attacks

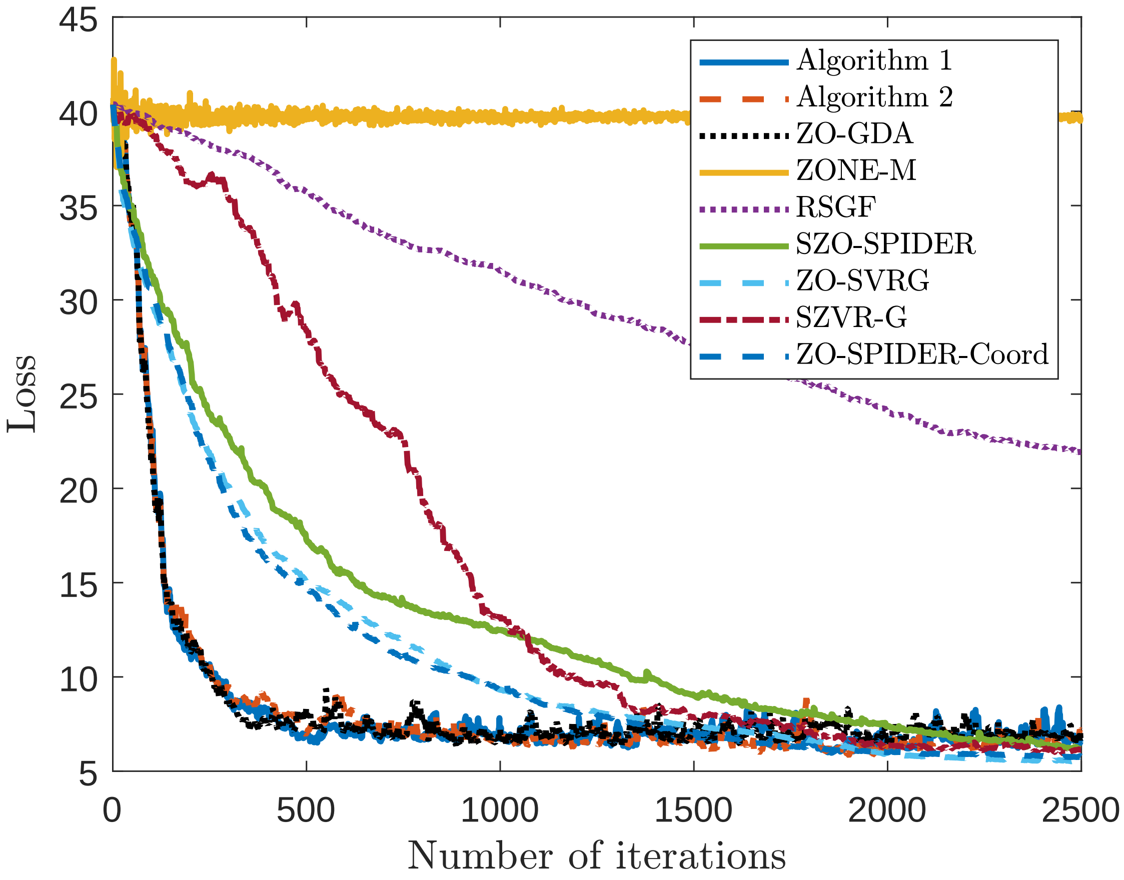


Fig. 1: Evolutions of the black-box attack loss with respect to the number of iterations.

algorithms. TABLE IV provides a comparison of generated adversarial examples from the DNN on the MNIST dataset: digit class “4”.

In order to verify the result that linear speedup convergence is achieved with respect to the number of agents, we also consider $n = 100$ agents. Fig. 3 compares the evolutions of the black-box attack loss achieved by the proposed distributed ZO algorithms with respect to the number of iterations when using different numbers of agents, which matches the theoretical result.

VI. Conclusions

In this paper, we studied stochastic distributed nonconvex optimization with ZO information feedback. We proposed two distributed ZO algorithms and analyzed their convergence properties. More specifically, linear speedup convergence rate $\mathcal{O}(\sqrt{p/(nT)})$ was established for smooth nonconvex cost functions under arbitrarily connected communication networks. The convergence

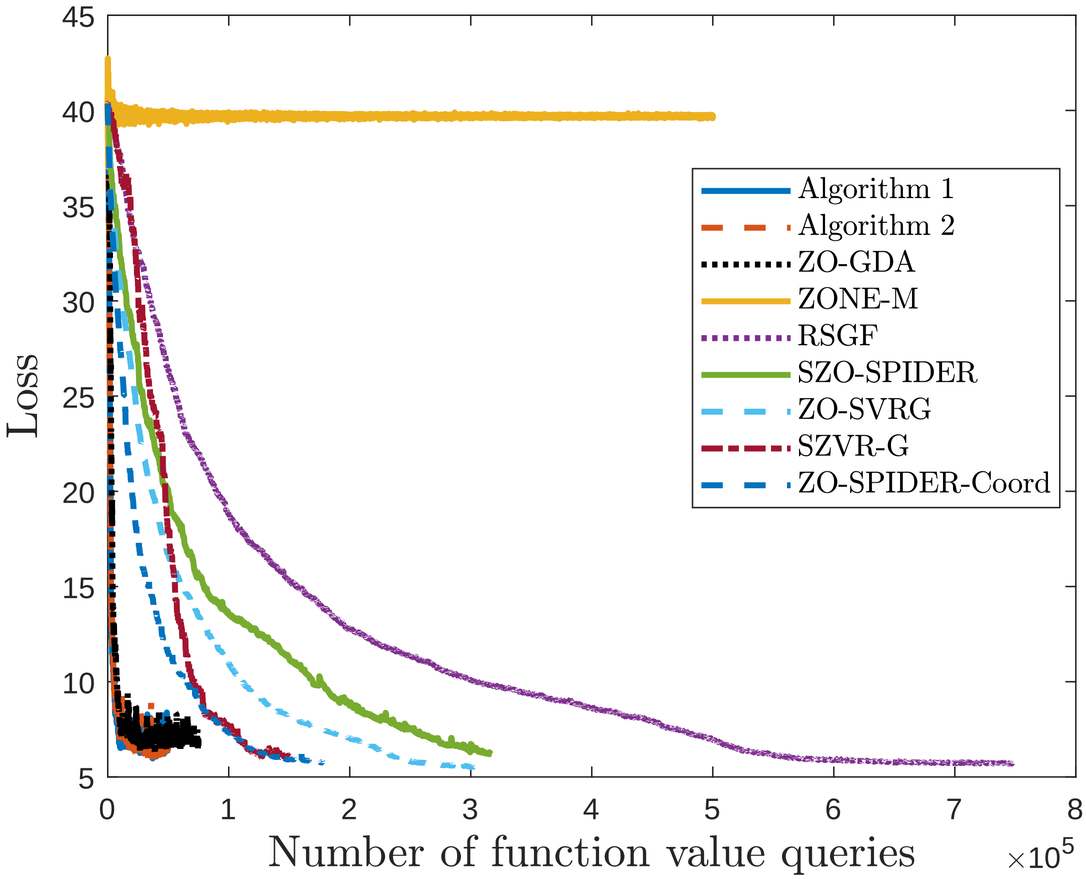
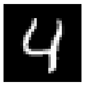
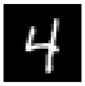
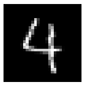
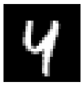
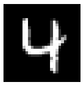
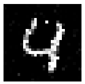
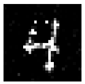
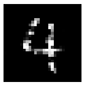
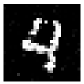
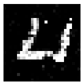
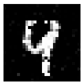
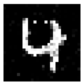
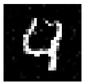
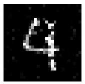
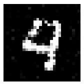
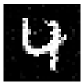
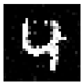
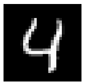
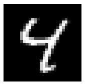
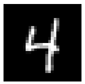
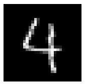
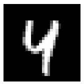
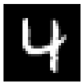
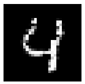
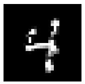
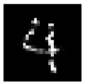
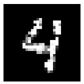
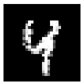
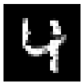
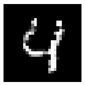
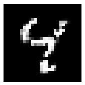
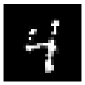
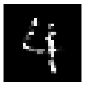
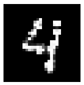
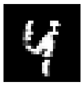
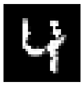
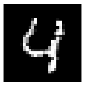
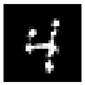
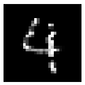
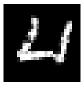
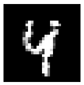
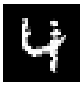
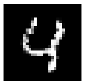
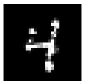
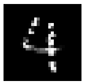
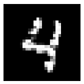
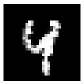
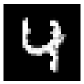
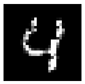
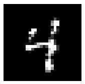
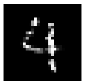
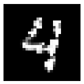
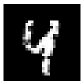
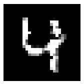


Fig. 2: Evolutions of the black-box attack loss with respect to the number of function value queries.

TABLE III: Distortion

Algorithm	ℓ_2 distortion
Algorithm 1	6.44
Algorithm 2	5.77
ZO-GDA	7.23
RSGF	5.69
SZO-SPIDER	6.19
ZO-SVRG	4.76
SZVR-G	5.16
ZO-SPIDER-Coord	5.76

TABLE IV: Comparison of generated adversarial examples from a black-box DNN on MNIST: digit class "4".

Image ID	4	6	19	24	27	33	42	48	49	56
Original										
Algorithm 1										
Classified as	9	8	2	7	2	2	9	9	9	9
Algorithm 2										
Classified as	9	9	7	9	9	2	9	9	9	9
ZO-GDA										
Classified as	9	9	2	2	2	2	9	9	9	3
ZONE-M										
Classified as	4	4	4	4	4	4	4	4	4	4
RSGF										
Classified as	9	9	2	9	9	2	9	9	9	9
SZO-SPIDER										
Classified as	9	9	7	9	9	2	9	9	9	9
ZO-SVRG										
Classified as	9	8	2	9	9	2	9	9	9	9
SZVR-G										
Classified as	9	8	2	2	2	2	9	9	9	9
ZO-SPIDER-Coord										
Classified as	9	9	2	9	9	2	9	9	9	9

rate was improved to $\mathcal{O}(p/(nT))$ when the global cost function satisfies the P–Ł condition. It was also shown that the output of the proposed algorithms linearly converges to a neighborhood of a global optimum. Interesting directions for future work include establishing faster convergence with reduced sampling complexity by using variance reduction techniques, and considering communication reduction with asynchronous, periodic, or compressed communication.

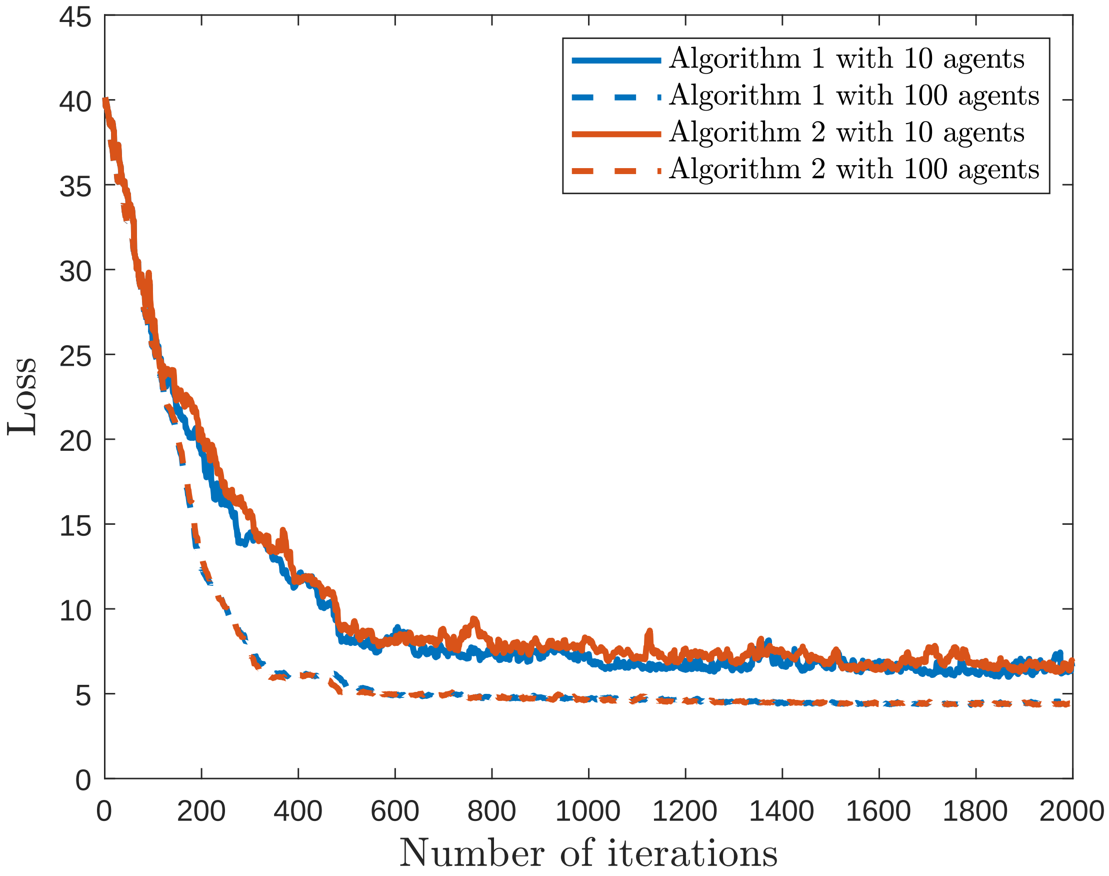


Fig. 3: Evolutions of the black-box attack loss with respect to the number of iterations when using different numbers of agents.

References

[1] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*. MPS-SIAM Series on Optimization. SIAM Philadelphia, 2009.

[2] C. Audet and W. Hare, *Derivative-Free and Blackbox Optimization*. Springer, 2017.

[3] J. Larson, M. Menickelly, and S. M. Wild, “Derivative-free optimization methods,” *Acta Numerica*, vol. 28, pp. 287–404, 2019.

[4] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, “ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models,” in *ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[5] A. Nedić and J. Liu, “Distributed optimization for control,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 77–103, 2018.

[6] A. Koloskova, S. Stich, and M. Jaggi, “Decentralized stochastic optimization and gossip algorithms with compressed communication,” in *International Conference on Machine Learning*, 2019, pp. 3478–3487.

[7] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, “A survey of distributed optimization,” *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[8] R. Hooke and T. A. Jeeves, ““Direct search” solution of numerical and statistical problems,” *Journal of the ACM*, vol. 8, no. 2, pp. 212–229, 1961.

[9] J. Matyas, “Random optimization,” *Automation and Remote Control*, vol. 26, no. 2, pp. 246–253, 1965.

[10] J. A. Nelder and R. Mead, “A simplex method for function minimization,” *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[11] E. H. Bergou, E. Gorbunov, and P. Richtarik, “Stochastic three points method for unconstrained smooth minimization,” *arXiv:1902.03591*, 2019.

[12] A. Bibi, E. H. Bergou, O. Sener, B. Ghanem, and P. Richtarik, “A stochastic derivative-free optimization method with importance sampling: Theory and learning to control,” *arXiv:1902.01272*, 2019.

[13] E. Gorbunov, A. Bibi, O. Sener, E. H. Bergou, and P. Richtárik, “A stochastic derivative free optimization method with momentum,” in *International Conference on Learning Representations*, 2020.

[14] D. Golovin, J. Karro, G. Kochanski, C. Lee, X. Song *et al.*, “Gradientless descent: High-dimensional zeroth-order optimization,” *arXiv:1911.06317*, 2019.

[15] M. Marazzi and J. Nocedal, “Wedge trust region methods for derivative free optimization,” *Mathematical Programming*, vol. 91, no. 2, pp. 289–305, 2002.

[16] A. R. Conn, K. Scheinberg, and L. N. Vicente, “Global convergence of general derivative-free trust-region algorithms to first- and second-order critical points,” *SIAM Journal on Optimization*, vol. 20, no. 1, pp. 387–415, 2009.

[17] K. Scheinberg and P. L. Toint, “Self-correcting geometry in model-based algorithms for derivative-free unconstrained optimization,” *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 3512–3532, 2010.

[18] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, “Optimal rates for zero-order convex optimization: The power of two function evaluations,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[19] O. Shamir, “An optimal algorithm for bandit and zero-order convex optimization with two-point feedback,” *Journal of Machine Learning Research*, vol. 18, no. 52, pp. 1–11, 2017.

[20] Y. Nesterov and V. Spokoiny, “Random gradient-free minimization of convex functions,” *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[21] O. Shamir, “On the complexity of bandit and derivative-free stochastic convex optimization,” in *Conference on Learning Theory*, 2013, pp. 3–24.

[22] S. Ghadimi and G. Lan, “Stochastic first- and zeroth-order methods for nonconvex stochastic programming,” *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[23] F. Bach and V. Perchet, “Highly-smooth zero-th order online optimization,” in *Conference on Learning Theory*, 2016, pp. 257–283.

[24] K. Balasubramanian and S. Ghadimi, “Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates,” in *Advances in Neural Information Processing Systems*, 2018, pp. 3455–3464.

[25] C. Jin, L. T. Liu, R. Ge, and M. I. Jordan, “On the local minima of the empirical risk,” in *Advances in Neural Information Processing Systems*, 2018, pp. 4896–4905.

[26] H. Ye, Z. Huang, C. Fang, C. J. Li, and T. Zhang, “Hessian-aware zeroth-order optimization for black-box adversarial attack,” *arXiv:1812.11377*, 2018.

[27] E.-V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras, “Efficiently avoiding saddle points with zero order methods: No gradients required,” in *Advances in Neural Information Processing Systems*, 2019, pp. 10 066–10 077.

[28] D. Kozak, S. Becker, A. Doostan, and L. Tenorio, “A stochastic subspace approach to gradient-free optimization in high dimensions,” *arXiv:2003.02684*, 2020.

[29] S. Liu, X. Li, P.-Y. Chen, J. Haupt, and L. Amini, “Zeroth-order stochastic projected gradient descent for nonconvex optimization,” in *IEEE Global Conference on Signal and Information Processing*, 2018, pp. 1179–1183.

[30] S. Liu, P.-Y. Chen, X. Chen, and M. Hong, “signSGD via zeroth-order oracle,” in *International Conference on Learning Representations*, 2019.

[31] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, “Improving the convergence rate of one-point zeroth-order optimization using residual feedback,” *arXiv:2006.10820*, 2020.

[32] X. Lian, H. Zhang, C.-J. Hsieh, Y. Huang, and J. Liu, “A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3054–3062.

[33] S. Ghadimi, G. Lan, and H. Zhang, “Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization,” *Mathematical Programming*, vol. 155, no. 1-2, pp. 267–305, 2016.

[34] X. Gao, B. Jiang, and S. Zhang, “On the information-adaptive variants of the ADMM: An iteration complexity perspective,” *Journal of Scientific Computing*, vol. 76, no. 1, pp. 327–363, 2018.

[35] E. Kazemi and L. Wang, “A proximal zeroth-order algorithm for nonconvex nonsmooth problems,” in *Annual Allerton Conference on Communication, Control, and Computing*, 2018, pp. 64–71.

[36] B. Gu, Z. Huo, C. Deng, and H. Huang, “Faster derivative-free stochastic algorithm for shared memory machines,” in *International Conference on Machine Learning*, 2018, pp. 1812–1821.

[37] C. Fang, C. J. Li, Z. Lin, and T. Zhang, “Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator,” in *Advances in Neural Information Processing Systems*, 2018, pp. 689–699.

[38] S. Liu, B. Kailkhura, P.-Y. Chen, P. Ting, S. Chang, and L. Amini, “Zeroth-order stochastic variance reduction for nonconvex optimization,” in *Advances in Neural Information Processing Systems*, 2018, pp. 3727–3737.

[39] E. Gorbunov, P. Dvurechensky, and A. Gasnikov, “An accelerated method for derivative-free smooth stochastic convex optimization,” *arXiv:1802.09022*, 2018.

[40] L. Liu, M. Cheng, C.-J. Hsieh, and D. Tao, “Stochastic zeroth-order optimization via variance reduction method,” *arXiv:1805.11811*, 2018.

[41] F. Huang, B. Gu, Z. Huo, S. Chen, and H. Huang, “Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization,” in *AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1503–1510.

[42] K. Ji, Z. Wang, Y. Zhou, and Y. Liang, “Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization,” in *International Conference on Machine Learning*, 2019, pp. 3100–3109.

[43] F. Huang, L. Tao, and S. Chen, “Accelerated stochastic gradient-free and projection-free methods,” in *International Conference on Machine Learning*, 2020.

[44] Y. Chen, A. Orvieto, and A. Lucchi, “An accelerated DFO algorithm for finite-sum convex functions,” in *International Conference on Machine Learning*, 2020.

[45] H. Gao and H. Huang, “Can stochastic zeroth-order Frank–Wolfe method converge faster for non-convex problems?” in *International Conference on Machine Learning*, 2020.

[46] H. Cai, D. Mckenzie, W. Yin, and Z. Zhang, “Zeroth-order regularized optimization (ZORO): Approximately sparse gradients and adaptive sampling,” *arXiv:2003.13001*, 2020.

[47] P. Nazari, D. A. Tarzanagh, and G. Michailidis, “Adaptive first- and zeroth-order methods for weakly convex stochastic optimization problems,” *arXiv:2005.09261*, 2020.

[48] A. K. Sahu, M. Zaheer, and S. Kar, "Towards gradient free and projection free stochastic optimization," in *International Conference on Artificial Intelligence and Statistics*, 2019, pp. 3468–3477.

[49] Y. Wang, S. Du, S. Balakrishnan, and A. Singh, "Stochastic zeroth-order optimization in high dimensions," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1356–1365.

[50] X. Chen, S. Liu, K. Xu, X. Li, X. Lin, M. Hong, and D. Cox, "ZO-AdaMM: Zeroth-order adaptive momentum method for black-box optimization," in *Advances in Neural Information Processing Systems*, 2019, pp. 7204–7215.

[51] F. Huang, S. Gao, S. Chen, and H. Huang, "Zeroth-order stochastic alternating direction method of multipliers for nonconvex nonsmooth optimization," in *International Conference on Artificial Intelligence and Statistics*, 2019, pp. 2549–2555.

[52] F. Huang, S. Gao, J. Pei, and H. Huang, "Nonconvex zeroth-order stochastic ADMM methods with lower function query complexity," *arXiv:1907.13463*, 2019.

[53] D. Yuan and D. W. Ho, "Randomized gradient-free method for multiagent optimization over time-varying networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1342–1347, 2014.

[54] A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar, "Distributed zeroth order optimization over random networks: A Kiefer–Wolfowitz stochastic approximation approach," in *IEEE Conference on Decision and Control*, 2018, pp. 4951–4958.

[55] Y. Wang, W. Zhao, Y. Hong, and M. Zamani, "Distributed subgradient-free stochastic optimization algorithm for nonsmooth convex functions over time-varying networks," *SIAM Journal on Control and Optimization*, vol. 57, no. 4, pp. 2821–2842, 2019.

[56] Y. Pang and G. Hu, "Randomized gradient-free distributed optimization methods for a multi-agent system with unknown cost function," *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 333–340, 2020.

[57] Y. Tang, J. Zhang, and N. Li, "Distributed zero-order algorithms for nonconvex multi-agent optimization," *arXiv:1908.11444v3*, 2020.

[58] D. Yuan, S. Xu, and J. Lu, "Gradient-free method for distributed multi-agent optimization via push-sum algorithms," *International Journal of Robust and Nonlinear Control*, vol. 25, no. 10, pp. 1569–1580, 2015.

[59] Z. Yu, D. W. Ho, and D. Yuan, "Distributed randomized gradient-free mirror descent algorithm for constrained optimization," *arXiv:1903.04157*, 2019.

[60] D. Hajinezhad, M. Hong, and A. Garcia, "ZONE: Zeroth-order nonconvex multiagent optimization over networks," *IEEE Transactions on Automatic Control*, vol. 64, no. 10, pp. 3995–4010, 2019.

[61] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Linear convergence of first- and zeroth-order algorithms for distributed nonconvex optimization under the Polyak–Łojasiewicz condition," *arXiv:1912.12110*, 2019.

[62] A. Beznosikov, E. Gorbunov, and A. Gasnikov, "Derivative-free method for composite optimization with applications to decentralized distributed optimization," *arXiv:1911.10645v4*, 2020.

[63] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 5330–5340.

[64] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016, pp. 795–811.

[65] H. Zhang and L. Cheng, "Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization," *Optimization Letters*, vol. 9, no. 5, pp. 961–979, 2015.

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

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

[68] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2018.

[69] ——, "Accelerated distributed Nesterov gradient descent," *IEEE Transactions on Automatic Control*, vol. 65, no. 6, pp. 2566–2581, 2020.

[70] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," in *International Conference on Machine Learning*, 2019, pp. 7184–7193.

[71] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[72] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[73] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.

[74] X. Yi, L. Yao, T. Yang, J. George, and K. H. Johansson, "Distributed optimization for second-order multi-agent systems with dynamic event-triggered communication," in *IEEE Conference on Decision and Control*, 2018, pp. 3397–3402.

[75] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer International Publishing, 2018.

[76] X. Yi, X. Li, T. Yang, L. Xie, T. Chai, and K. H. Johansson, "Distributed bandit online convex optimization with time-varying coupled inequality constraints," *arXiv:1912.03719*, 2019.

[77] S. Kar, J. M. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3575–3605, 2012.

Appendix

A. Notations, Definitions, and Useful Lemmas

1) Notations: $\mathbf{1}_n$ ($\mathbf{0}_n$) denotes the column one (zero) vector of dimension n. $\mathrm{col}(z_1,\dots,z_k)$ is the concatenated column vector of vectors $z_i \in \mathbb{R}^{p_i},\ i \in [k]$. $\boldsymbol{I}_n$ is the n-dimensional identity matrix. Given a vector $[x_1,\dots,x_n]^\top \in \mathbb{R}^n$, $\mathrm{diag}([x_1,\dots,x_n])$ is a diagonal matrix with the i-th diagonal element being x_i. The notation $A \otimes B$ denotes the Kronecker product of matrices A and B. $\mathrm{null}(A)$ is the null space of matrix A. Given two symmetric matrices M, N, $M \geq N$ means that $M - N$ is positive semi-definite. $\rho(\cdot)$ stands for the spectral radius for matrices and $\rho_2(\cdot)$ indicates the minimum positive eigenvalue for matrices having positive eigenvalues. For any square matrix A, $\|x\|_A^2$ denotes $x^\top A x$. $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the ceiling and floor functions, respectively. For any $x \in \mathbb{R}$, $[x]_+$ is the positive part of x. $\mathbf{1}_{(\cdot)}$ is the indicator function.

2) Graph Theory: For an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, let $\mathcal{A} = (a_{ij})$ be the associated weighted adjacency matrix with $a_{ij} > 0$ if $(i,j) \in \mathcal{E}$ if $a_{ij} > 0$ and zero otherwise. It is assumed that $a_{ii} = 0$ for all $i \in [n]$. Let $\mathrm{deg}_i = \sum\limits_{j=1}^{n} a_{ij}$ denotes the weighted degree of vertex

i. The degree matrix of graph $\mathcal{G}$ is $\mathrm{Deg} = \mathrm{diag}([\mathrm{deg}_1, \cdots, \mathrm{deg}_n])$. The Laplacian matrix is $L = (L_{ij}) = \mathrm{Deg} - \mathcal{A}$. A path of length k between vertices i and j is a subgraph with distinct vertices $i_0 = i, \dots, i_k = j \in [n]$ and edges $(i_j, i_{j+1}) \in \mathcal{E},\ j = 0, \dots, k-1$. An undirected graph is connected if there exists at least one path between any two distinct vertices.

For a connected undirected graph, we have the following results.

Lemma 1. *(Lemmas 1 and 2 in [74]) Let L be the Laplacian matrix of the connected graph $\mathcal{G}$ and $K_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. Then L and K_n are positive semi-definite,* $\mathrm{null}(L) = \mathrm{null}(K_n) = \{\mathbf{1}_n\}$, *$L \le \rho(L)\boldsymbol{I}_n$, $\rho(K_n) = 1$,*

$$K_n L = L K_n = L, \tag{34a}$$

$$0 \le \rho_2(L) K_n \le L \le \rho(L) K_n. \tag{34b}$$

Moreover, there exists an orthogonal matrix $[r\ R] \in \mathbb{R}^{n\times n}$ with $r = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $R \in \mathbb{R}^{n\times(n-1)}$ such that

$$R\Lambda_1^{-1}R^\top L = LR\Lambda_1^{-1}R^\top = K_n, \tag{35a}$$

$$\frac{1}{\rho(L)}K_n \le R\Lambda_1^{-1}R^\top \le \frac{1}{\rho_2(L)}K_n, \tag{35b}$$

where $\Lambda_1 = \mathrm{diag}([\lambda_2, \dots, \lambda_n])$ with $0 < \lambda_2 \le \cdots \le \lambda_n$ being the eigenvalues of the Laplacian matrix L.

3) Smooth Functions:

Definition 2. *[75] A function $f(x)$: $\mathbb{R}^p \mapsto \mathbb{R}$ is smooth with constant $L_f > 0$ if it is differentiable and*

$$\|\nabla f(x) - \nabla f(y)\| \le L_f\|x - y\|,\ \forall x, y \in \mathbb{R}^p. \tag{36}$$

From Lemma 1.2.3 in [75], we know that (36) implies

$$|f(y) - f(x) - (y - x)^\top \nabla f(x)| \le \frac{L_f}{2}\|y - x\|^2,\ \forall x, y \in \mathbb{R}^p, \tag{37}$$

which further implies

$$\|\nabla f(x)\|^2 \le 2L_f(f(x) - f^*),\ \forall x, y \in \mathbb{R}^p, \tag{38}$$

where $f^* = \min_{x\in\mathbb{R}^p} f(x)$.

4) Properties of Gradient Approximation: The random gradient estimator $\hat{\nabla}_2 f$ defined in (3) is an unbiased gradient estimator of f^s, where f^s is the uniformly smoothed version of f defined as

$$f^s(x,\delta) = \mathbf{E}_{u\in\mathbb{B}^p}[f(x+\delta u)], \tag{39}$$

with the expectation is taken with respect to uniform distribution.

From Lemma 2 in [76], Lemma 5 in [57], and Proposition 7.6 in [34], we have the following properties of f^s and $\hat{\nabla}_2 f$.

Lemma 2. *(i) The uniform smoothing $f^s(x,\delta)$ is differentiable with respect to x, and*

$$\nabla f^s(x,\delta) = \mathbf{E}_{u\in\mathbb{S}^p}[\hat{\nabla}_2 f(x,\delta,u)]. \tag{40}$$

(ii) If f is smooth with constant $L_f > 0$, then

$$\|\nabla f^s(x,\delta) - \nabla f(x)\| \le \delta L_f, \tag{41a}$$

$$\mathbf{E}_{u\in\mathbb{S}^p}[\|\hat{\nabla}_2 f(x,\delta,u)\|^2] \le 2p\|\nabla f(x)\|^2 + \frac{1}{2}p^2\delta^2 L_f^2. \tag{41b}$$

5) Useful Lemmas on Series:

Lemma 3. *Let $a, b \in (0,1)$ be two constants, then*

$$\sum_{\tau=0}^{k} a^\tau b^{k-\tau} \le \begin{cases} \frac{a^{k+1}}{a-b}, & \textit{if } a > b \\ \frac{b^{k+1}}{b-a}, & \textit{if } a < b \\ \frac{c^{k+1}}{c-b}, & \textit{if } a = b, \end{cases} \tag{42}$$

where c is any constant in $(a,1)$.

Proof: If $a > b$, then

$$\sum_{\tau=0}^{k} a^\tau b^{k-\tau} = a^k \sum_{\tau=0}^{k} (\frac{b}{a})^{k-\tau} \le \frac{a^{k+1}}{a-b}.$$

Similarly, when $a < b$, we have

$$\sum_{\tau=0}^{k} a^{\tau} b^{k-\tau} = b^{k} \sum_{\tau=0}^{k} (\frac{a}{b})^{\tau} \leq \frac{b^{k+1}}{b-a}.$$

If $a = b$, then for any $c \in (a, 1)$, we have

$$\sum_{\tau=0}^{k} a^{\tau} b^{k-\tau} \leq \sum_{\tau=0}^{k} c^{\tau} b^{k-\tau} = c^{k} \sum_{\tau=0}^{k} (\frac{b}{c})^{k-\tau} \leq \frac{c^{k+1}}{c-b}.$$

Hence, this lemma holds. ■

Lemma 4. *Let k and τ be two integers and δ be a constant. Suppose $k \geq \tau \geq 1$, then*

$$\sum_{l=\tau}^{k} l^{\delta} \leq \begin{cases} \frac{(k+1)^{\delta+1}}{\delta+1}, & \textit{if } \delta > -1 \\ \ln(k), & \textit{if } \delta = -1 \\ \frac{-(\tau-1)^{\delta+1}}{\delta+1}, & \textit{if } \delta < -1 \textit{ and } \tau \geq 2. \end{cases} \tag{43}$$

Proof: If $\delta \geq 0$, then $h(t) = t^{\delta}$ is an increasing function in the interval $[1, +\infty)$. Hence,

$$\sum_{l=\tau}^{k} l^{\delta} \leq \int_{\tau}^{k+1} t^{\delta} dt = \frac{(k+1)^{\delta+1} - \tau^{\delta+1}}{\delta+1} \leq \frac{(k+1)^{\delta+1}}{\delta+1}. \tag{44}$$

If $\delta < 0$, then $h(t) = t^{\delta}$ is a decreasing function in the interval $[1, +\infty)$. Hence,

$$\sum_{l=\tau}^{k} l^{\delta} \leq \int_{\tau-1}^{k} t^{\delta} dt = \begin{cases} \ln(\frac{k}{\tau-1}), & \text{if } \delta = -1, \\ \frac{k^{\delta+1}-(\tau-1)^{\delta+1}}{\delta+1}, & \text{if } -1 < \delta < 0, \\ \frac{k^{\delta+1}-(\tau-1)^{\delta+1}}{\delta+1}, & \text{if } \delta < -1 \text{ and } \tau \geq 2, \end{cases}$$

$$\leq \begin{cases} \ln(k), & \text{if } \delta = -1, \\ \frac{(k+1)^{\delta+1}}{\delta+1}, & \text{if } -1 < \delta < 0, \\ \frac{-(\tau-1)^{\delta+1}}{\delta+1}, & \text{if } \delta < -1 \text{ and } \tau \geq 2. \end{cases} \tag{45}$$

Finally, (44) and (45) yield (43). ■

Lemma 5. *Let $\{z_k\}$, $\{r_{1,k}\}$, and $\{r_{2,k}\}$ be sequences. Suppose there exists $t_1 \in \mathbb{N}_+$ such that*

$$z_k \geq 0, \tag{46a}$$

$$z_{k+1} \leq (1 - r_{1,k}) z_k + r_{2,k}, \tag{46b}$$

$$1 > r_{1,k} \geq \frac{a_1}{(k+t_1)^{\delta_1}}, \tag{46c}$$

$$r_{2,k} \leq \frac{a_2}{(k+t_1)^{\delta_2}},\ \forall k \in \mathbb{N}_0, \tag{46d}$$

where $a_1 > 0$*,* $a_2 > 0$*,* $\delta_1 \in [0,1]$*, and* $\delta_2 > \delta_1$ *are constants.*

(i) If $\delta_1 \in (0,1)$*, then*

$$z_k \leq \phi_1(k, t_1, a_1, a_2, \delta_1, \delta_2, z_0),\ \forall k \in \mathbb{N}_+, \tag{47}$$

where

$$\begin{aligned}\phi_1(k, t_1, a_1, a_2, \delta_1, \delta_2, z_0) = {} & \frac{1}{s_1(k+t_1)}\Big(s_1(t_1)z_0 + \frac{[t_2 - 1 - t_1]_+ s_1(t_1+1)a_2}{t_1^{\delta_2}}\Big) \\ & + \frac{a_2}{(k+t_1-1)^{\delta_2}} + \frac{\mathbf{1}_{(k+t_1-1 \geq t_2)}(\frac{t_1+1}{t_1})^{\delta_2} a_2 \delta_2}{a_1 \delta_1 (k+t_1)^{\delta_2 - \delta_1}},\end{aligned} \tag{48}$$

$s_1(k) = e^{\frac{a_1}{1-\delta_1}k^{1-\delta_1}}$ *and* $t_2 = \lceil (\frac{\delta_2}{a_1})^{\frac{1}{1-\delta_1}} \rceil$*.*

(ii) If $\delta_1 = 1$*, then*

$$z_k \leq \phi_2(k, t_1, a_1, a_2, \delta_2, z_0),\ \forall k \in \mathbb{N}_+, \tag{49}$$

where

$$\phi_2(k, t_1, a_1, a_2, \delta_2, z_0) = \frac{t_1^{a_1} z_0}{(k+t_1)^{a_1}} + \frac{a_2}{(k+t_1-1)^{\delta_2}} + \Big(\frac{t_1+1}{t_1}\Big)^{\delta_2} a_2 s_2(k+t_1), \tag{50}$$

and

$$s_2(k) = \begin{cases} \frac{1}{(a_1-\delta_2+1)k^{\delta_2-1}}, & \text{if } a_1 - \delta_2 > -1, \\ \frac{\ln(k-1)}{k^{a_1}}, & \text{if } a_1 - \delta_2 = -1, \\ \frac{-t_1^{a_1-\delta_2+1}}{(a_1-\delta_2+1)k^{a_1}}, & \text{if } a_1 - \delta_2 < -1. \end{cases}$$

(iii) If $\delta_1 = 0$*, then*

$$z_k \leq \phi_3(k, t_1, a_1, a_2, \delta_2, z_0),\ \forall k \in \mathbb{N}_+, \tag{51}$$

where

$$\begin{aligned}\phi_3(k, t_1, a_1, a_2, \delta_2, z_0) = {} & (1-a_1)^k z_0 + a_2(1-a_1)^{k+t_1-1}\Big([t_3 - t_1]_+ s_3(t_1) \\ & + ([t_4 - t_1]_+ - [t_3 - t_1]_+)s_3(t_4)\Big)\end{aligned}$$

$$+\frac{\mathbf{1}_{(k+t_1-1\geq t_4)}2a_2}{-\ln(1-a_1)(k+t_1)^{\delta_2}(1-a_1)}, \tag{52}$$

$s_3(k)=\frac{1}{k^{\delta_2}(1-a_1)^k}$, $t_3=\lceil\frac{-\delta_2}{\ln(1-a_1)}\rceil$, *and* $t_4=\lceil\frac{-2\delta_2}{\ln(1-a_1)}\rceil$.

Proof: This proof is inspired by the proof of Lemma 25 in [77].

From (46a)–(46c), for any $k\in\mathbb{N}_+$, it holds that

$$z_k\leq\prod_{\tau=0}^{k-1}(1-r_{1,\tau})z_0+r_{2,k-1}+\sum_{l=0}^{k-2}\prod_{\tau=l+1}^{k-1}(1-r_{1,\tau})r_{2,l}. \tag{53}$$

For any $t\in[0,1]$, it holds that $1-t\leq e^{-t}$ since $s_4(t)=1-t-e^{-t}$ is a non-increasing function in the interval $[0,1]$ and $s_4(0)=0$. Thus, for any $k>l\geq0$, it holds that

$$\prod_{\tau=l}^{k-1}(1-r_{1,\tau})\leq e^{-\sum_{\tau=l}^{k-1}r_{1,\tau}}. \tag{54}$$

We also have

$$\begin{aligned}\sum_{\tau=l}^{k-1}r_{1,\tau}&\geq\sum_{\tau=l}^{k-1}\frac{a_1}{(\tau+t_1)^{\delta_1}}=\sum_{\tau=l+t_1}^{k-1+t_1}\frac{a_1}{\tau^{\delta_1}}\geq\int_{t=l+t_1}^{k+t_1}\frac{a_1}{t^{\delta_1}}dt\\&=\begin{cases}\frac{a_1}{1-\delta_1}((k+t_1)^{1-\delta_1}-(l+t_1)^{1-\delta_1}), & \text{if } \delta_1\in(0,1),\\ a_1\ln(\frac{k+t_1}{l+t_1}), & \text{if } \delta_1=1,\end{cases}\end{aligned} \tag{55}$$

where the first inequality holds due to (46c) and the second inequality holds since $s_5(t)=a_1/t^{\delta_1}$ is a decreasing function in the interval $[1,+\infty)$.

Hence, (54) and (55) yield

$$\prod_{\tau=l}^{k-1}(1-r_{1,\tau})\leq e^{-\sum_{\tau=l}^{k-1}r_{1,\tau}}\leq\begin{cases}\frac{s_1(l+t_1)}{s_1(k+t_1)}, & \text{if } \delta_1\in(0,1),\\ \frac{(l+t_1)^{a_1}}{(k+t_1)^{a_1}}, & \text{if } \delta_1=1.\end{cases} \tag{56}$$

(i) When $\delta_1\in(0,1)$, from (56) and (46d), we have

$$\begin{aligned}\sum_{l=0}^{k-2}\prod_{\tau=l+1}^{k-1}(1-r_{1,\tau})r_{2,l}&\leq\sum_{l=0}^{k-2}\frac{s_1(l+t_1+1)}{s_1(k+t_1)}\frac{a_2}{(l+t_1)^{\delta_2}}\\&=\frac{a_2}{s_1(k+t_1)}\sum_{l=0}^{k-2}\frac{s_1(l+t_1+1)}{(l+t_1)^{\delta_2}}\\&\leq\frac{a_2}{s_1(k+t_1)}\sum_{l=0}^{k-2}\frac{s_1(l+t_1+1)}{(\frac{t_1}{t_1+1}l+t_1)^{\delta_2}}\end{aligned}$$

$$\begin{aligned}
&= \frac{(\frac{t_1+1}{t_1})^{\delta_2} a_2}{s_1(k+t_1)} \sum_{l=0}^{k-2} \frac{s_1(l+t_1+1)}{(l+t_1+1)^{\delta_2}} \\
&= \frac{(\frac{t_1+1}{t_1})^{\delta_2} a_2}{s_1(k+t_1)} \sum_{l=t_1+1}^{k+t_1-1} \frac{s_1(l)}{l^{\delta_2}} \\
&= \frac{(\frac{t_1+1}{t_1})^{\delta_2} a_2}{s_1(k+t_1)} \Big(\sum_{l=t_1+1}^{t_2-1} \frac{s_1(l)}{l^{\delta_2}} + \sum_{l=t_2}^{k+t_1-1} \frac{s_1(l)}{l^{\delta_2}} \Big). \qquad (57)
\end{aligned}$$

We know that $s_6(t) = s_1(t)/t^{\delta_2}$ is a decreasing function in the interval $[1, t_2 - 1]$ since

$$\frac{ds_6(t)}{dt} = \Big(a_1 - \frac{\delta_2}{t^{1-\delta_1}}\Big) \frac{s_6(t)}{t^{\delta_1}} \le 0, \ \forall t \in \Big(0, \Big(\frac{\delta_2}{a_1}\Big)^{\frac{1}{1-\delta_1}}\Big].$$

Thus, for any $k \in [1, t_2 - 1]$, we have

$$\sum_{l=k}^{t_2-1} \frac{s_1(l)}{l^{\delta_2}} \le (t_2 - k) \frac{s_1(k)}{k^{\delta_2}}. \qquad (58)$$

Noting that $s_6(t) = s_1(t)/t^{\delta_2}$ is an increasing function in the interval $[t_2, +\infty)$, for any $k \ge t_2$, we have

$$\sum_{l=t_2}^{k} \frac{s_1(l)}{l^{\delta_2}} \le \int_{t_2}^{k+1} \frac{s_1(t)}{t^{\delta_2}} dt. \qquad (59)$$

We have

$$\begin{aligned}
\int_{t_2}^{k+1} \frac{s_1(t)}{t^{\delta_2}} dt &= \int_{t_2}^{k+1} \frac{1}{a_1 t^{\delta_2-\delta_1}} ds_1(t) \\
&= \frac{s_1(k+1)}{a_1(k+1)^{\delta_2-\delta_1}} - \frac{s_1(t_2)}{a_1 t_2^{\delta_2-\delta_1}} + \int_{t_2}^{k+1} \frac{(\delta_2-\delta_1)s_1(t)}{a_1 t^{\delta_2-\delta_1+1}} dt \\
&\le \frac{s_1(k+1)}{a_1(k+1)^{\delta_2-\delta_1}} + \int_{t_2}^{k+1} \frac{(\delta_2-\delta_1)}{a_1 t^{1-\delta_1}} \frac{s_1(t)}{t^{\delta_2}} dt \\
&\le \frac{s_1(k+1)}{a_1(k+1)^{\delta_2-\delta_1}} + \frac{\delta_2-\delta_1}{a_1 t_2^{1-\delta_1}} \int_{t_2}^{k+1} \frac{s_1(t)}{t^{\delta_2}} dt \\
&\le \frac{s_1(k+1)}{a_1(k+1)^{\delta_2-\delta_1}} + \frac{\delta_2-\delta_1}{\delta_2} \int_{t_2}^{k+1} \frac{s_1(t)}{t^{\delta_2}} dt, \qquad (60)
\end{aligned}$$

where the second inequality holds since $s_7(t) = 1/t^{1-\delta_1}$ is a decreasing function in the interval $[1, +\infty)$; and the last inequality holds due to $t_2^{1-\delta_1} \ge \frac{\delta_2}{a_1}$.

From (59) and (60), for any $k \geq t_2$, we have

$$\sum_{l=t_2}^{k} \frac{s_1(l)}{l^{\delta_2}} \leq \int_{t_2}^{k+1} \frac{s_1(t)}{t^{\delta_2}} dt \leq \frac{\delta_2 s_1(k+1)}{a_1\delta_1(k+1)^{\delta_2-\delta_1}}. \tag{61}$$

From (57), (58), and (61), we have

$$\begin{aligned}&\sum_{l=0}^{k-2} \prod_{\tau=l+1}^{k-1} (1-r_{1,\tau})r_{2,l} \\ &\leq \frac{(\frac{t_1+1}{t_1})^{\delta_2} a_2}{s_1(k+t_1)} \Big(\frac{[t_2-1-t_1]_+ s_1(t_1+1)}{(t_1+1)^{\delta_2}} + \mathbf{1}_{(k+t_1-1\geq t_2)} \frac{\delta_2 s_1(k+t_1)}{a_1\delta_1(k+t_1)^{\delta_2-\delta_1}} \Big).\end{aligned} \tag{62}$$

Then, (53), (56), and (62) yield (47).

(ii) When $\delta_1 = 1$, from (56) and (46d), we have

$$\begin{aligned}\sum_{l=0}^{k-2} \prod_{\tau=l+1}^{k-1} (1-r_{1,\tau})r_{2,l} &\leq \sum_{l=0}^{k-2} \frac{(l+t_1+1)^{a_1}}{(k+t_1)^{a_1}} \frac{a_2}{(l+t_1)^{\delta_2}} \\ &\leq \sum_{l=0}^{k-2} \frac{(l+t_1+1)^{a_1}}{(k+t_1)^{a_1}} \frac{a_2}{(\frac{t_1}{t_1+1}l+t_1)^{\delta_2}} \\ &= \frac{(\frac{t_1+1}{t_1})^{\delta_2} a_2}{(k+t_1)^{a_1}} \sum_{l=0}^{k-2} \frac{(l+t_1+1)^{a_1}}{(l+t_1+1)^{\delta_2}} \\ &= \frac{(\frac{t_1+1}{t_1})^{\delta_2} a_2}{(k+t_1)^{a_1}} \sum_{l=t_1+1}^{k+t_1-1} l^{a_1-\delta_2},\end{aligned} \tag{63}$$

where the first inequality holds due to (56) and (46d).

From (53), (56), (63), and (43), we have (49).

(iii) Denote $a = 1 - a_1$. From (46c) and $\delta_1 = 0$, we know that $a_1 \in (0,1)$. Thus, $a \in (0,1)$.

From (46a)–(46d) and $\delta_1 = 0$, for any $k \in \mathbb{N}_+$, it holds that

$$\begin{aligned}z_k &\leq (1-a_1)^k z_0 + \sum_{\tau=0}^{k-1} (1-a_1)^{k-1-\tau} r_{2,\tau} \\ &\leq a^k z_0 + a_2 a^{k+t_1-1} \sum_{\tau=0}^{k-1} \frac{1}{(\tau+t_1)^{\delta_2} a^{\tau+t_1}}.\end{aligned} \tag{64}$$

We have

$$\sum_{\tau=0}^{k-1} \frac{1}{(\tau+t_1)^{\delta_2} a^{\tau+t_1}} = \sum_{\tau=t_1}^{k+t_1-1} \frac{1}{\tau^{\delta_2} a^{\tau}} = \sum_{\tau=t_1}^{t_3-1} s_3(\tau) + \sum_{\tau=t_3}^{t_4-1} s_3(\tau) + \sum_{\tau=t_4}^{k+t_1-1} s_3(\tau). \tag{65}$$

We know that $s_3(t) = 1/(t^{\delta_2} a^t)$ is decreasing and increasing in the intervals $[1, t_3 - 1]$ and $[t_3, +\infty)$, respectively, since

$$\frac{ds_3(t)}{dt} = -s_3(t)\Big(\frac{\delta_2}{t} + \ln(a)\Big) \le 0,\ \forall t \in \Big(0, \frac{-\delta_2}{\ln(a)}\Big],$$
$$\frac{ds_3(t)}{dt} = -s_3(t)\Big(\frac{\delta_2}{t} + \ln(a)\Big) \ge 0,\ \forall t \in \Big[\frac{-\delta_2}{\ln(a)}, +\infty\Big).$$

Thus, we have

$$\sum_{\tau=k_1}^{t_3-1} s_3(\tau) \le (t_3 - k_1) s_3(k_1),\ \forall k_1 \in [1, t_3 - 1], \tag{66a}$$
$$\sum_{\tau=k_2}^{t_4-1} s_3(\tau) \le (t_4 - k_2) s_3(t_4),\ \forall k_2 \in [t_3, t_4 - 1], \tag{66b}$$
$$\sum_{\tau=t_4}^{k_3} s_3(\tau) \le \int_{t_4}^{k_3+1} s_3(t) dt,\ \forall k_3 \ge t_4. \tag{66c}$$

Denote $b = 1/a$. We have

$$\begin{aligned}\int_{t_4}^{k_3+1} s_3(t) dt &= \int_{t_4}^{k_3+1} \frac{b^t}{t^{\delta_2}} dt = \int_{t_4}^{k_3+1} \frac{1}{\ln(b) t^{\delta_2}} db^t \\ &= \frac{b^{k_3+1}}{\ln(b)(k_3+1)^{\delta_2}} - \frac{b^{t_4}}{\ln(b) t_4^{\delta_2}} + \int_{t_4}^{k_3+1} \frac{\delta_2 b^t}{\ln(b) t^{\delta_2+1}} dt \\ &\le \frac{b^{k_3+1}}{\ln(b)(k_3+1)^{\delta_2}} + \int_{t_4}^{k_3+1} \frac{\delta_2}{\ln(b) t} s_3(t) dt \\ &\le \frac{b^{k_3+1}}{\ln(b)(k_3+1)^{\delta_2}} + \frac{\delta_2}{\ln(b) t_4} \int_{t_4}^{k_3+1} s_3(t) dt \\ &\le \frac{b^{k_3+1}}{\ln(b)(k_3+1)^{\delta_2}} + \frac{1}{2} \int_{t_4}^{k_3+1} s_3(t) dt, \end{aligned} \tag{67}$$

where the last inequality holds due to $t_4 = \lceil -2\delta_2 / \ln(1 - a_1) \rceil \ge -2\delta_2 / \ln(1 - a_1) = 2\delta_2 / \ln(b)$.

From (66c) and (67), we have

$$\sum_{\tau=t_4}^{k_3} s_3(\tau) \le \frac{2}{-\ln(a)(k_3+1)^{\delta_2} a^{k_3+1}},\ \forall k_3 \ge t_4. \tag{68}$$

From (64), (65), (66a), (66b), and (68), we get (51). ■

B. Proof of Theorem 1

Denote $\boldsymbol{L} = L \otimes \boldsymbol{I}_p$, $\boldsymbol{K} = K_n \otimes \boldsymbol{I}_p$, $\boldsymbol{H} = \frac{1}{n}(\mathbf{1}_n \mathbf{1}_n^\top \otimes \boldsymbol{I}_p)$, $\boldsymbol{Q} = R\Lambda_1^{-1}R^\top \otimes \boldsymbol{I}_p$, $\delta_k = \max_{i\in[n]}\{\delta_{i,k}\}$, $\boldsymbol{x} = \mathrm{col}(x_1,\dots,x_n)$, $\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^n f_i(x_i)$, $\bar{x}_k = \frac{1}{n}(\mathbf{1}_n^\top \otimes \boldsymbol{I}_p)\boldsymbol{x}_k$, $\bar{\boldsymbol{x}}_k = \mathbf{1}_n \otimes \bar{x}_k$, $\boldsymbol{g}_k = \nabla\tilde{f}(\boldsymbol{x}_k)$, $\bar{\boldsymbol{g}}_k = \boldsymbol{H}\boldsymbol{g}_k$, $\boldsymbol{g}_k^0 = \nabla\tilde{f}(\bar{\boldsymbol{x}}_k)$, $\bar{\boldsymbol{g}}_k^0 = \boldsymbol{H}\boldsymbol{g}_k^0 = \mathbf{1}_n \otimes \nabla f(\bar{x}_k)$, $\boldsymbol{g}_k^e = \mathrm{col}(g_{1,k}^e,\dots,g_{n,k}^e)$, $\bar{g}_k^e = \frac{1}{n}(\mathbf{1}_n^\top \otimes \boldsymbol{I}_p)\boldsymbol{g}_k^e$, $\bar{\boldsymbol{g}}_k^e = \mathbf{1}_n \otimes \bar{g}_k^e = \boldsymbol{H}\boldsymbol{g}_k^e$, $f_i^s(x,\delta_{i,k}) = \mathbf{E}_{u\in\mathbb{B}^p}[f_i(x+\delta_{i,k}u)]$, $g_{i,k}^s = \nabla f_i^s(x_{i,k},\delta_{i,k})$, $\boldsymbol{g}_k^s = \mathrm{col}(g_{1,k}^s,\dots,g_{n,k}^s)$, and $\bar{\boldsymbol{g}}_k^s = \boldsymbol{H}\boldsymbol{g}_k^s$.

We also denote the following notations.

$$
\begin{aligned}
&c_0(\kappa_1,\kappa_2) = \max\Big\{\varepsilon_1,\ \frac{2\varepsilon_5}{\varepsilon_4},\ \Big(\frac{2p\varepsilon_7}{\varepsilon_4}\Big)^{\frac{1}{2}},\ \frac{\varepsilon_8}{2\varepsilon_6},\ \frac{24\kappa_4}{\kappa_2},\ 96p\kappa_2\varepsilon_{10}\Big\},\\
&c_1 = \frac{1}{\rho_2(L)} + 1,\\
&c_2(\kappa_1) = \min\Big\{\frac{\varepsilon_2}{\varepsilon_3},\ \frac{1}{5}\Big\},\\
&c_3(\kappa_1,\kappa_2) = \frac{24\kappa_3}{\kappa_2},\\
&\kappa_3 = \frac{1}{\rho_2(L)} + \kappa_1 + 1,\\
&\kappa_4 = \frac{1}{\rho_2(L)} + \kappa_1,\\
&\kappa_5 = \frac{1}{\rho_2(L)} + \kappa_1 + \frac{3}{2},\\
&\kappa_6 = \frac{\kappa_1+1}{2} + \frac{1}{2\rho_2(L)},\\
&\kappa_7 = \min\Big\{\frac{1}{2\rho(L)},\ \frac{\kappa_1-1}{2\kappa_1}\Big\},\\
&\varepsilon_1 = \max\Big\{1+3L_f^2,\ (8+12p(3+0.5L_f))^{\frac{1}{2}}L_f,\ p\kappa_3\Big\},\\
&\varepsilon_2 = (\kappa_1-1)\rho_2(L) - 1,\\
&\varepsilon_3 = \rho(L) + (2\kappa_1^2+1)\rho(L^2) + 1,\\
&\varepsilon_4 = 0.5(\varepsilon_2\kappa_2 - \varepsilon_3\kappa_2^2),\\
&\varepsilon_5 = 0.5 - \kappa_1\kappa_2\rho_2(L) + \kappa_2^2\rho(L) + 0.5(1+3\kappa_1\kappa_2+2\kappa_2)\kappa_1\kappa_2\rho(L^2),\\
&\varepsilon_6 = 0.25(\kappa_2 - 5\kappa_2^2),\\
&\varepsilon_7 = 6(1+6\kappa_2+2\kappa_4+10\kappa_2\kappa_4)\kappa_2L_f^4 + \frac{1}{2p}(1+2L_f^2)\kappa_2 + \Big(\frac{5}{p}+24\Big)L_f^2\kappa_2^2,\\
&\varepsilon_8 = \kappa_4 + \kappa_1\kappa_2 + 3\kappa_2^2 + \kappa_2\kappa_4,
\end{aligned}
$$

$$\begin{aligned}
&\varepsilon_9 = \frac{3\kappa_0}{2\kappa_2^2}(2\kappa_4+1),\\
&\varepsilon_{10} = 10 + L_f + \frac{1}{\kappa_2}(2\kappa_4+1)L_f^2 + (10\kappa_4+6)L_f^2,\\
&\varepsilon_{11} = L_f^2\Big(\frac{1}{384} + \frac{1}{p}(13\kappa_2+4)\Big),\\
&\varepsilon_{12} = 2\varepsilon_{10}\sigma_1^2 + \frac{1}{p}\varepsilon_9\sigma_2^2 + 6\varepsilon_{10}\sigma_2^2,\\
&\varepsilon_{13} = \frac{1}{p}\varepsilon_9 + 6\varepsilon_{10},\\
&\varepsilon_{14} = \frac{W_0}{n} + \frac{2\theta p(\varepsilon_{11}\kappa_\delta^2 + \varepsilon_{12})\kappa_2^2}{(2\theta-1)\kappa_0^2},\\
&a_1 = \frac{1}{\kappa_6}\min\{\varepsilon_4,\ \varepsilon_6\},\\
&a_2 = pn(\varepsilon_{11}\kappa_\delta^2 + \varepsilon_{12} + 2L_f\varepsilon_{13}\varepsilon_{14})\frac{\kappa_2^2}{\kappa_0^2}.
\end{aligned}$$

To prove Theorem 1, the following three lemmas are used.

Lemma 6. *Suppose Assumption 3 holds. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1, then*

$$\boldsymbol{g}_k^s = \mathbf{E}_{\mathfrak{L}_k}[\boldsymbol{g}_k^e], \tag{69a}$$

$$\|\boldsymbol{g}_k^0 - \boldsymbol{g}_k^s\|^2 \le 2L_f^2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + 2nL_f^2\delta_k^2, \tag{69b}$$

$$\|\bar{\boldsymbol{g}}_k^0 - \bar{\boldsymbol{g}}_k^s\|^2 \le 2L_f^2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + 2nL_f^2\delta_k^2, \tag{69c}$$

$$\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2] \le \frac{1}{n}\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2] + \|\bar{\boldsymbol{g}}_k^s\|^2, \tag{69d}$$

$$\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^0 - \boldsymbol{g}_k^e\|^2] \le 4L_f^2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + 4nL_f^2\delta_k^2 + 2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2], \tag{69e}$$

$$\|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0\|^2 \le \eta_k^2L_f^2\|\bar{\boldsymbol{g}}_k^e\|^2 \le \eta_k^2L_f^2\|\boldsymbol{g}_k^e\|^2, \tag{69f}$$

$$\|\bar{\boldsymbol{g}}_k^0\|^2 \le 2nL_f(f(\bar{x}_k) - f^*). \tag{69g}$$

If Assumptions 4 and 5 also hold, then

$$\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2] \le 12p\|\bar{\boldsymbol{g}}_k^0\|^2 + 12pL_f^2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + 4np\sigma_1^2 + 12np\sigma_2^2 + 0.5np^2L_f^2\delta_k^2, \tag{70a}$$

$$\|\boldsymbol{g}_{k+1}^0\|^2 \le 3(\eta_k^2L_f^2\|\boldsymbol{g}_k^e\|^2 + n\sigma_2^2 + \|\bar{\boldsymbol{g}}_k^0\|^2). \tag{70b}$$

Proof: (i) From $u_{i,k}$ and $\xi_{i,k}$ are mutually independent, $x_{i,k}$ is independent of $u_{i,k}$ and $\xi_{i,k}$, and

(40), we have

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[g^e_{i,k}] &= \mathbf{E}_{u_{i,k}}\big[\mathbf{E}_{\xi_{i,k}}\big[\frac{p}{\delta_{i,k}}(F_i(x_{i,k}+\delta_{i,k}u_{i,k},\xi_{i,k})-F_i(x_{i,k},\xi_{i,k}))u_{i,k}\big]\big]\\ &= \mathbf{E}_{u_{i,k}}\big[\frac{p}{\delta_{i,k}}(f_i(x_{i,k}+\delta_{i,k}u_{i,k})-f_i(x_{i,k}))u_{i,k}\big]\\ &= \mathbf{E}_{u_{i,k}}[\hat{\nabla}_2 f_i(x_{i,k},\delta_{i,k},u_{i,k})] = \nabla f^s_i(x_{i,k},\delta_{i,k}) = g^s_{i,k},\end{aligned}$$

which gives (69a).

(ii) From Assumption 3, we know that each $f_i(x) = \mathbf{E}_{\xi_i}[F_i(x,\xi_i)]$ is smooth with constant L_f since

$$\begin{aligned}\|\nabla f_i(x)-\nabla f_i(y)\| &= \|\mathbf{E}_{\xi_i}[\nabla_x F_i(x,\xi_i)-\nabla_x F_i(y,\xi_i)]\|\\ &\le \mathbf{E}_{\xi_i}[\|\nabla_x F_i(x,\xi_i)-\nabla_x F_i(y,\xi_i)\|]\\ &\le \mathbf{E}_{\xi_i}[L_f\|x-y\|] = L_f\|x-y\|,\ \forall x,y\in\mathbb{R}^p.\end{aligned} \tag{71}$$

From (71), we have

$$\begin{aligned}\|\boldsymbol{g}^0_k-\boldsymbol{g}_k\|^2 &= \|\nabla\tilde{f}(\bar{\boldsymbol{x}}_k)-\nabla\tilde{f}(\boldsymbol{x}_k)\|^2\\ &= \sum_{i=1}^n\|\nabla f_i(\bar{x}_k)-\nabla f_i(x_{i,k})\|^2 \le \sum_{i=1}^n L_f^2\|\bar{x}_k-x_{i,k}\|^2\\ &= L_f^2\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}_k\|^2 = L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}.\end{aligned} \tag{72}$$

From (71) and (41a), we have

$$\|g^s_{i,k}-g_{i,k}\| \le L_f\delta_{i,k}.$$

Thus,

$$\|\boldsymbol{g}^s_k-\boldsymbol{g}_k\|^2 = \sum_{i=1}^n\|g^s_{i,k}-g_{i,k}\|^2 \le nL_f^2\delta_k^2. \tag{73}$$

Noting $\|\boldsymbol{g}^0_k-\boldsymbol{g}^s_k\|^2 \le 2\|\boldsymbol{g}^0_k-\boldsymbol{g}_k\|^2+2\|\boldsymbol{g}_k-\boldsymbol{g}^s_k\|^2$, from (72) and (73), we know (69b) holds.

(iii) Noting $\|\bar{\boldsymbol{g}}^0_k-\bar{\boldsymbol{g}}^s_k\|^2 = \|\boldsymbol{H}(\boldsymbol{g}^0_k-\boldsymbol{g}^s_k)\|^2$, from $\rho(\boldsymbol{H})=1$ and (69b), we know (69c) hold.

(iv) We have

$$\mathbf{E}_{\mathfrak{L}_k}[\|\bar{g}^e_k\|^2] = \mathbf{E}_{\mathfrak{L}_k}\Big[\Big\|\sum_{i=1}^n\frac{1}{n}g^e_{i,k}\Big\|^2\Big]$$

$$\begin{aligned}
&= \frac{1}{n^2}\mathbf{E}_{\mathfrak{L}_k}\Big[\sum_{i=1}^{n}\|g^e_{i,k}\|^2 + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\langle g^e_{i,k}, g^e_{j,k}\rangle\Big]\\
&= \frac{1}{n^2}\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^e_k\|^2] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\langle \mathbf{E}_{\mathfrak{L}_k}[g^e_{i,k}], \mathbf{E}_{\mathfrak{L}_k}[g^e_{j,k}]\rangle\\
&= \frac{1}{n^2}\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^e_k\|^2] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\langle g^s_{i,k}, g^s_{j,k}\rangle\\
&= \frac{1}{n^2}\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^e_k\|^2] + \|\bar{g}^s_k\|^2 - \frac{1}{n^2}\|\boldsymbol{g}^s_k\|^2, \qquad (74)
\end{aligned}$$

where the third equality holds since $u_{i,k}$ and $\xi_{i,k},\ \forall i \in [n], k \geq 1$ are mutually independent; and the fourth equality holds due to (69a).

From (74), $\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}^e_k\|^2] = n\mathbf{E}_{\mathfrak{L}_k}[\|\bar{g}^e_k\|^2]$ and $\|\bar{\boldsymbol{g}}^s_k\|^2 = n\|\bar{g}^s_k\|^2$, we know that (69d) holds.

(v) We have

$$\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^0_k - \boldsymbol{g}^e_k\|^2] &\leq 2\|\boldsymbol{g}^0_k - \boldsymbol{g}^s_k\|^2 + 2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^s_k - \boldsymbol{g}^e_k\|^2]\\
&= 2\|\boldsymbol{g}^0_k - \boldsymbol{g}^s_k\|^2 + 2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^e_k\|^2] - 2\|\boldsymbol{g}^s_k\|^2, \qquad (75)
\end{aligned}$$

where the inequality holds due to the Cauchy–Schwarz inequality; and the equality holds since (69a) and $\boldsymbol{x}_k$ is independent of $\mathfrak{L}_k$.

From (75) and (69b), we know (69e) holds.

(vi) The distributed ZO algorithm (6) can be rewritten as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k(\alpha_k\boldsymbol{L}\boldsymbol{x}_k + \beta_k\boldsymbol{v}_k + \boldsymbol{g}^e_k), \qquad (76a)$$

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \eta_k\beta_k\boldsymbol{L}\boldsymbol{x}_k,\ \forall \boldsymbol{x}_0 \in \mathbb{R}^{np},\ \sum_{i=1}^{n} v_{i,0} = \mathbf{0}_p. \qquad (76b)$$

From (76b), we know that

$$\bar{v}_{k+1} = \bar{v}_k. \qquad (77)$$

Then, from (77), $\sum_{i=1}^{n} v_{i,0} = \mathbf{0}_p$, and (76a), we know that $\bar{v}_k = \mathbf{0}_p$ and

$$\bar{\boldsymbol{x}}_{k+1} = \bar{\boldsymbol{x}}_k - \eta_k\bar{\boldsymbol{g}}^e_k. \qquad (78)$$

Then, we have

$$\|\boldsymbol{g}^0_{k+1} - \boldsymbol{g}^0_k\|^2 = \|\nabla\tilde{f}(\bar{\boldsymbol{x}}_{k+1}) - \nabla\tilde{f}(\bar{\boldsymbol{x}}_k)\|^2$$

$$\leq L_f^2\|\bar{\boldsymbol{x}}_{k+1}-\bar{\boldsymbol{x}}_k\|^2=\eta_k^2L_f^2\|\bar{\boldsymbol{g}}_k^e\|^2\leq\eta_k^2L_f^2\|\boldsymbol{g}_k^e\|^2,$$

where the first inequality holds due to (71); the last equality holds due to (78); and the last equality holds due to $\bar{\boldsymbol{g}}_k^e=\boldsymbol{H}\boldsymbol{g}_k^e$ and $\rho(\boldsymbol{H})=1$. Thus, (69f) holds.

(vii) From (38), we have

$$\|\bar{\boldsymbol{g}}_k^0\|^2=n\|\nabla f(\bar{x}_k)\|^2\leq 2nL_f(f(\bar{x}_k)-f^*),\tag{79}$$

which yields (69g).

(viii) From Assumption 3, $x_{i,k}$ and $\xi_{i,k}$ are independent of $u_{i,k}$, and (41b), we know that for almost every $\xi_{i,k}$ it holds that

$$\mathbf{E}_{u_{i,k}}[\|g_{i,k}^e\|^2]\leq 2p\|\nabla_xF_i(x_{i,k},\xi_{i,k})\|^2+0.5p^2L_f^2\delta_{i,k}^2.\tag{80}$$

Then,

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[\|g_{i,k}^e\|^2]&\leq 2p\mathbf{E}_{\xi_{i,k}}[\|\nabla_xF_i(x_{i,k},\xi_{i,k})\|^2]+0.5p^2L_f^2\delta_{i,k}^2\\&=2p\mathbf{E}_{\xi_{i,k}}[\|\nabla_xF_i(x_{i,k},\xi_{i,k})-\nabla f_i(x_{i,k})+\nabla f_i(x_{i,k})\|^2]+0.5p^2L_f^2\delta_{i,k}^2\\&\leq 4p\mathbf{E}_{\xi_{i,k}}[\|\nabla_xF_i(x_{i,k},\xi_{i,k})-\nabla f_i(x_{i,k})\|^2+\|\nabla f_i(x_{i,k})\|^2]+0.5p^2L_f^2\delta_{i,k}^2\\&\leq 4p\|\nabla f_i(x_{i,k})\|^2+4p\sigma_1^2+0.5p^2L_f^2\delta_{i,k}^2,\end{aligned}\tag{81}$$

where the first inequality holds due to (80); the second inequality holds due to the Cauchy–Schwarz inequality; and the last inequality holds since Assumption 4 and $x_{i,k}$ is independent of $\xi_{i,k}$.

From (71), we have

$$\begin{aligned}\|\nabla f(x)-\nabla f(y)\|^2&=\Big\|\frac{1}{n}\sum_{i=1}^n(\nabla f_i(x)-\nabla f_i(y))\Big\|^2\\&\leq\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x)-\nabla f_i(y)\|^2\leq L_f^2\|x-y\|^2,\ \forall x,y\in\mathbb{R}^p.\end{aligned}\tag{82}$$

Then, we have

$$\begin{aligned}\|\nabla f_i(x_{i,k})\|^2&=\|\nabla f_i(x_{i,k})-\nabla f(x_{i,k})+\nabla f(x_{i,k})-\nabla f(\bar{x}_k)+\nabla f(\bar{x}_k)\|^2\\&\leq 3(\|\nabla f_i(x_{i,k})-\nabla f(x_{i,k})\|^2+\|\nabla f(x_{i,k})-\nabla f(\bar{x}_k)\|^2+\|\nabla f(\bar{x}_k)\|^2)\end{aligned}$$

$$\leq 3(\sigma_2^2 + L_f^2\|x_{i,k} - \bar{x}_k\|^2 + \|\nabla f(\bar{x}_k)\|^2), \tag{83}$$

where the first inequality holds due to the Cauchy–Schwarz inequality; and the last inequality holds due to Assumption 5 and (82).

From (81) and (83), we know (70a) holds.

(ix) From the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\|\boldsymbol{g}_{k+1}^0\|^2 &= \|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0 + \boldsymbol{g}_k^0 - \bar{\boldsymbol{g}}_k^0 + \bar{\boldsymbol{g}}_k^0\|^2 \\ &\leq 3(\|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0\|^2 + \|\boldsymbol{g}_k^0 - \bar{\boldsymbol{g}}_k^0\|^2 + \|\bar{\boldsymbol{g}}_k^0\|^2). \end{aligned}\tag{84}$$

From Assumption 5, we have

$$\|\boldsymbol{g}_k^0 - \bar{\boldsymbol{g}}_k^0\|^2 = \sum_{i=1}^n \|f_i(\bar{x}_k) - f(\bar{x}_k)\|^2 \leq n\sigma_2^2. \tag{85}$$

From (84), (85), and (69f), we know (70b) holds. ■

Lemma 7. *Suppose Assumptions 1–5 hold. Suppose $\{\beta_k\}$ is non-decreasing, $\alpha_k = \kappa_1\beta_k$, and $\eta_k = \frac{\kappa_2}{\beta_k}$, where $\kappa_1 > 1$ and $\kappa_2 > 0$ are constants. Moreover, suppose $\beta_k \geq \varepsilon_1$. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1, then*

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}] \leq{}& W_k - \|\boldsymbol{x}_k\|^2_{(2\varepsilon_4 - \varepsilon_5\omega_k - b_{1,k})\boldsymbol{K}} - \Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{b_{2,k}\boldsymbol{K}} \\ &- \eta_k(0.25 - (b_{3,k} + 6pb_{4,k})\eta_k)\|\bar{\boldsymbol{g}}_k^0\|^2 + 2pn\sigma_1^2 b_{4,k}\eta_k^2 \\ &+ n\sigma_2^2(b_{3,k} + 6pb_{4,k})\eta_k^2 + b_{5,k}\eta_k\delta_k^2, \end{aligned}\tag{86a}$$

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[\breve{W}_{k+1}] \leq{}& \breve{W}_k - \|\boldsymbol{x}_k\|^2_{(2\varepsilon_4 - \varepsilon_5\omega_k - b_{1,k})\boldsymbol{K}} - \Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{b_{2,k}\boldsymbol{K}} \\ &+ (b_{3,k} + 6pb_{4,k})\eta_k^2\|\bar{\boldsymbol{g}}_k^0\|^2 + 2pn\sigma_1^2 b_{4,k}\eta_k^2 \\ &+ n\sigma_2^2(b_{3,k} + 6pb_{4,k})\eta_k^2 + b_{5,k}\eta_k\delta_k^2, \end{aligned}\tag{86b}$$

where $W_k = \sum_{i=1}^4 W_{i,k}$, $\breve{W}_k = \sum_{i=1}^3 W_{i,k}$, and

$$\begin{aligned} W_{1,k} &= \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}, \\ W_{2,k} &= \frac{1}{2}\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}, \\ W_{3,k} &= \boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big), \end{aligned}$$

$$
\begin{aligned}
W_{4,k} &= n(f(\bar{x}_k) - f^*) = \tilde{f}(\bar{\boldsymbol{x}}_k) - f^*,\\
b_{1,k} &= 6p\kappa_3 L_f^4 \frac{\eta_k}{\beta_k^2} + 12p\kappa_5 L_f^4 \frac{\eta_k^2}{\beta_k^2} + (0.5 + L_f^2)\eta_k\omega_k\\
&\quad + 6p\kappa_4 L_f^4 \frac{\eta_k\omega_k}{\beta_k^2} + (5 + 24p + 18p\kappa_3 L_f^2)L_f^2\eta_k^2\omega_k\\
&\quad + 12p\kappa_4 L_f^4 \frac{\eta_k^2\omega_k}{\beta_k^2} + 18p\kappa_4 L_f^4 \eta_k^2\omega_k^2,\\
b_{2,k} &= 2\varepsilon_6 - 0.5\omega_k(\kappa_1 + \kappa_4 + \kappa_1\kappa_2 + 3\kappa_2^2) - 0.5\omega_k\eta_k\kappa_4,\\
b_{3,k} &= \frac{3}{2}\kappa_3\frac{\omega_k}{\eta_k^2} + \frac{3}{2}\kappa_4\frac{\omega_k^2}{\eta_k^2},\\
b_{4,k} &= 6 + L_f + \frac{\kappa_3}{\kappa_2}L_f^2\frac{1}{\beta_k} + (4 + 3\kappa_3 L_f^2)\omega_k + 3\kappa_4 L_f^2\omega_k^2\\
&\quad + 2\kappa_5 L_f^2\frac{1}{\beta_k^2} + \frac{\kappa_4}{\kappa_2}L_f^2\frac{\omega_k}{\beta_k} + 2\kappa_4 L_f^2\frac{\omega_k}{\beta_k^2},\\
b_{5,k} &= nL_f^2(0.25p^2 b_{4,k}\eta_k + 3 + \omega_k + 8\eta_k + 5\eta_k\omega_k).
\end{aligned}
$$

Proof: Note that $W_{4,k}$ is well defined due to $f^* > -\infty$ as assumed in Assumption 2. Thus, W_k is well defined.

(i) We have

$$
\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[W_{1,k+1}] &= \mathbf{E}_{\mathfrak{L}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_{k+1}\|^2_{\boldsymbol{K}}\Big]\\
&= \mathbf{E}_{\mathfrak{L}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_k - \eta_k(\alpha_k\boldsymbol{L}\boldsymbol{x}_k + \beta_k\boldsymbol{v}_k + \boldsymbol{g}_k^e)\|^2_{\boldsymbol{K}}\Big]\\
&= \mathbf{E}_{\mathfrak{L}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \eta_k\alpha_k\|\boldsymbol{x}_k\|^2_{\boldsymbol{L}} + \frac{1}{2}\eta_k^2\alpha_k^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{L}^2}\\
&\quad - \eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{I}_{np} - \eta_k\alpha_k\boldsymbol{L})\boldsymbol{K}\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^e\Big) + \frac{1}{2}\eta_k^2\beta_k^2\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^e\Big\|^2_{\boldsymbol{K}}\Big]\\
&= W_{1,k} - \|\boldsymbol{x}_k\|^2_{\eta_k\alpha_k\boldsymbol{L} - \frac{1}{2}\eta_k^2\alpha_k^2\boldsymbol{L}^2} - \eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{I}_{np} - \eta_k\alpha_k\boldsymbol{L})\boldsymbol{K}\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^s\Big)\\
&\quad + \frac{1}{2}\eta_k^2\beta_k^2\mathbf{E}_{\mathfrak{L}_k}\Big[\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0 + \frac{1}{\beta_k}\boldsymbol{g}_k^e - \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}\Big]\\
&= W_{1,k} - \|\boldsymbol{x}_k\|^2_{\eta_k\alpha_k\boldsymbol{L} - \frac{1}{2}\eta_k^2\alpha_k^2\boldsymbol{L}^2} - \eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{I}_{np}\\
&\quad - \eta_k\alpha_k\boldsymbol{L})\boldsymbol{K}\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0 + \frac{1}{\beta_k}\boldsymbol{g}_k^s - \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&\quad + \frac{1}{2}\eta_k^2\beta_k^2\mathbf{E}_{\mathfrak{L}_k}\Big[\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0 + \frac{1}{\beta_k}\boldsymbol{g}_k^e - \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}\Big]\\
&\le W_{1,k} - \|\boldsymbol{x}_k\|^2_{\eta_k\alpha_k\boldsymbol{L} - \frac{1}{2}\eta_k^2\alpha_k^2\boldsymbol{L}^2} - \eta_k\beta_k\boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{2}\eta_k\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+\frac{1}{2}\eta_k\|\boldsymbol{g}^s_k-\boldsymbol{g}^0_k\|^2\\
&+\frac{1}{2}\eta_k^2\alpha_k^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{L}^2}+\frac{1}{2}\eta_k^2\beta_k^2\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}\\
&+\frac{1}{2}\eta_k^2\alpha_k^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{L}^2}+\frac{1}{2}\eta_k^2\|\boldsymbol{g}^s_k-\boldsymbol{g}^0_k\|^2\\
&+\eta_k^2\beta_k^2\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}+\eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^e_k-\boldsymbol{g}^0_k\|^2]\\
={}&W_{1,k}-\|\boldsymbol{x}_k\|^2_{\eta_k\alpha_k\boldsymbol{L}-\frac{1}{2}\eta_k\boldsymbol{K}-\frac{3}{2}\eta_k^2\alpha_k^2\boldsymbol{L}^2}\\
&+\frac{1}{2}\eta_k(1+\eta_k)\|\boldsymbol{g}^s_k-\boldsymbol{g}^0_k\|^2+\eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^e_k-\boldsymbol{g}^0_k\|^2]\\
&-\eta_k\beta_k\boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big)+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\frac{3}{2}\eta_k^2\beta_k^2\boldsymbol{K}}\\
\le{}&W_{1,k}-\|\boldsymbol{x}_k\|^2_{\eta_k\alpha_k\boldsymbol{L}-\frac{1}{2}\eta_k\boldsymbol{K}-\frac{3}{2}\eta_k^2\alpha_k^2\boldsymbol{L}^2-\eta_k(1+5\eta_k)L_f^2\boldsymbol{K}}\\
&-\eta_k\beta_k\boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big)+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\frac{3}{2}\eta_k^2\beta_k^2\boldsymbol{K}}\\
&+nL_f^2\eta_k(1+5\eta_k)\delta_k^2+2\eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}^e_k\|^2],
\end{aligned}\tag{87}
$$

where the second equality holds due to (76a); the third equality holds due to (34a) in Lemma 1; the fourth equality holds since (69a) and that $x_{i,k}$ and $v_{i,k}$ are independent of $\mathfrak{L}_k$; the first inequality holds due to the Cauchy–Schwarz inequality and $\rho(\boldsymbol{K})=1$; and the last inequality holds due to (69b) and (69e).

(ii) We know that $\omega_k\ge0$ since $\{\beta_k\}$ is non-decreasing. We have

$$
\begin{aligned}
W_{2,k+1}&=\frac{1}{2}\Big\|\boldsymbol{v}_{k+1}+\frac{1}{\beta_{k+1}}\boldsymbol{g}^0_{k+1}\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&=\frac{1}{2}\Big\|\boldsymbol{v}_{k+1}+\frac{1}{\beta_k}\boldsymbol{g}^0_{k+1}+\Big(\frac{1}{\beta_{k+1}}-\frac{1}{\beta_k}\Big)\boldsymbol{g}^0_{k+1}\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&\le\frac{1}{2}(1+\omega_k)\Big\|\boldsymbol{v}_{k+1}+\frac{1}{\beta_k}\boldsymbol{g}^0_{k+1}\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}+\frac{1}{2}(\omega_k+\omega_k^2)\|\boldsymbol{g}^0_{k+1}\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}},
\end{aligned}\tag{88}
$$

where the inequality holds due to the Cauchy–Schwarz inequality.

For the first term in the right-hand side of (88), we have

$$
\begin{aligned}
\frac{1}{2}\Big\|\boldsymbol{v}_{k+1}+\frac{1}{\beta_k}\boldsymbol{g}^0_{k+1}\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}&=\frac{1}{2}\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k+\eta_k\beta_k\boldsymbol{L}\boldsymbol{x}_k+\frac{1}{\beta_k}(\boldsymbol{g}^0_{k+1}-\boldsymbol{g}^0_k)\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&=W_{2,k}+\eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{K}+\kappa_1\boldsymbol{L})\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big)\\
&\quad+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)}+\frac{1}{2\beta_k^2}\Big\|\boldsymbol{g}^0_{k+1}-\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{\beta_k}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0+\eta_k\beta_k\boldsymbol{L}\boldsymbol{x}_k\Big)^\top(\boldsymbol{Q}+\kappa_1\boldsymbol{K})(\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0)\\
\le{}& W_{2,k}+\eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{K}+\kappa_1\boldsymbol{L})\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)}+\frac{1}{2\beta_k^2}\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&+\frac{\eta_k}{2}\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}+\frac{1}{2\eta_k\beta_k^2}\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&+\frac{1}{2}\eta_k^2\beta_k^2\|\boldsymbol{L}\boldsymbol{x}_k\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}+\frac{1}{2\beta_k^2}\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
={}& W_{2,k}+\eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{K}+\kappa_1\boldsymbol{L})\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&+\|\boldsymbol{x}_k\|^2_{\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)}+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\frac{1}{2}\eta_k(\boldsymbol{Q}+\kappa_1\boldsymbol{K})}\\
&+\frac{1}{\beta_k^2}\Big(1+\frac{1}{2\eta_k}\Big)\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
\le{}& W_{2,k}+\eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{K}+\kappa_1\boldsymbol{L})\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&+\|\boldsymbol{x}_k\|^2_{\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)}+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\frac{1}{2}\eta_k(\boldsymbol{Q}+\kappa_1\boldsymbol{K})}\\
&+\frac{1}{\beta_k^2}\Big(1+\frac{1}{2\eta_k}\Big)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2\\
\le{}& W_{2,k}+\eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{K}+\kappa_1\boldsymbol{L})\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&+\|\boldsymbol{x}_k\|^2_{\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)}+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\frac{1}{2}\eta_k(\boldsymbol{Q}+\kappa_1\boldsymbol{K})}\\
&+\frac{\eta_k}{\beta_k^2}\Big(\eta_k+\frac{1}{2}\Big)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)L_f^2\|\bar{\boldsymbol{g}}_k^e\|^2,
\end{aligned}
\tag{89}
$$

where the first equality holds due to (76b); the second equality holds due to (34a) and (35a) in Lemma 1; the first inequality holds due to the Cauchy–Schwarz inequality; the last equality holds due to (34a) and (35a); the second inequality holds due to $\rho(\boldsymbol{Q}+\kappa_1\boldsymbol{K})\le\rho(\boldsymbol{Q})+\kappa_1\rho(\boldsymbol{K})$, (35b), $\rho(\boldsymbol{K})=1$; and the last inequality holds due to (69f).

For the second term in the right-hand side of (88), we have

$$
\|\boldsymbol{g}_{k+1}^0\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\le\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\|\boldsymbol{g}_{k+1}^0\|^2. \tag{90}
$$

Also note that

$$
\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\le\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}. \tag{91}
$$

Then, from (88)–(91), we have

$$\begin{aligned}
W_{2,k+1} \le{}& W_{2,k} + (1+\omega_k)\eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{K}+\kappa_1\boldsymbol{L})\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&+ \frac{1}{2}(\eta_k+\omega_k+\eta_k\omega_k)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}\\
&+ \|\boldsymbol{x}_k\|^2_{(1+\omega_k)\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)} + \frac{\eta_k}{\beta_k^2}\Big(\eta_k+\frac{1}{2}\Big)(1+\omega_k)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)L_f^2\|\bar{\boldsymbol{g}}_k^e\|^2\\
&+ \frac{1}{2}\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)(\omega_k+\omega_k^2)\|\boldsymbol{g}_{k+1}^0\|^2. \qquad (92)
\end{aligned}$$

(iii) We have

$$\begin{aligned}
W_{3,k+1} &= \boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1} + \frac{1}{\beta_{k+1}}\boldsymbol{g}_{k+1}^0\Big)\\
&= \boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1} + \frac{1}{\beta_k}\boldsymbol{g}_{k+1}^0 + \Big(\frac{1}{\beta_{k+1}} - \frac{1}{\beta_k}\Big)\boldsymbol{g}_{k+1}^0\Big)\\
&= \boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1} + \frac{1}{\beta_k}\boldsymbol{g}_{k+1}^0\Big) - \omega_k\boldsymbol{x}_{k+1}^\top\boldsymbol{K}\boldsymbol{g}_{k+1}^0\\
&\le \boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1} + \frac{1}{\beta_k}\boldsymbol{g}_{k+1}^0\Big) + \frac{1}{2}\omega_k(\|\boldsymbol{x}_{k+1}\|^2_{\boldsymbol{K}} + \|\boldsymbol{g}_{k+1}^0\|^2). \qquad (93)
\end{aligned}$$

For the first term in the right-hand side of (93), we have

$$\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}\Big[\boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1} + \frac{1}{\beta_k}\boldsymbol{g}_{k+1}^0\Big)\Big] &= \mathbf{E}_{\mathfrak{L}_k}\Big[(\boldsymbol{x}_k - \eta_k(\alpha_k\boldsymbol{L}\boldsymbol{x}_k + \beta_k\boldsymbol{v}_k + \boldsymbol{g}_k^0 + \boldsymbol{g}_k^e - \boldsymbol{g}_k^0))^\top\\
&\quad\times\boldsymbol{K}\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0 + \eta_k\beta_k\boldsymbol{L}\boldsymbol{x}_k + \frac{1}{\beta_k}(\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0)\Big)\Big]\\
&= \boldsymbol{x}_k^\top(\boldsymbol{K} - \eta_k(\alpha_k + \eta_k\beta_k^2)\boldsymbol{L})\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&\quad+ \|\boldsymbol{x}_k\|^2_{\eta_k\beta_k(\boldsymbol{L}-\eta_k\alpha_k\boldsymbol{L}^2)}\\
&\quad+ \frac{1}{\beta_k}\boldsymbol{x}_k^\top(\boldsymbol{K} - \eta_k\alpha_k\boldsymbol{L})\mathbf{E}_{\mathfrak{L}_k}[\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0]\\
&\quad- \eta_k\beta_k\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}\\
&\quad- \eta_k\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)^\top\boldsymbol{K}\mathbf{E}_{\mathfrak{L}_k}[\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0]\\
&\quad- \eta_k(\boldsymbol{g}_k^s - \boldsymbol{g}_k^0)^\top\boldsymbol{K}\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0 + \eta_k\beta_k\boldsymbol{L}\boldsymbol{x}_k\Big)\\
&\quad- \frac{1}{\beta_k}\mathbf{E}_{\mathfrak{L}_k}[\eta_k(\boldsymbol{g}_k^e - \boldsymbol{g}_k^0)^\top\boldsymbol{K}(\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0)]\\
&\le \boldsymbol{x}_k^\top(\boldsymbol{K} - \eta_k\alpha_k\boldsymbol{L})\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big) + \frac{1}{2}\eta_k^2\beta_k^2\|\boldsymbol{L}\boldsymbol{x}_k\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\eta_k^2\beta_k^2\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2+\|\boldsymbol{x}_k\|^2_{\eta_k\beta_k(\boldsymbol{L}-\eta_k\alpha_k\boldsymbol{L}^2)}\\
&+\frac{1}{2}\eta_k\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+\frac{1}{2\eta_k\beta_k^2}\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2\\
&+\frac{1}{2}\eta_k^2\alpha_k^2\|\boldsymbol{L}\boldsymbol{x}_k\|^2+\frac{1}{2\beta_k^2}\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2]\\
&-\eta_k\beta_k\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2\\
&+\frac{1}{2}\eta_k^2\beta_k^2\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2+\frac{1}{2\beta_k^2}\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2]\\
&+\frac{1}{2}\eta_k\|\boldsymbol{g}_k^s-\boldsymbol{g}_k^0\|^2+\frac{1}{2}\eta_k\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2\\
&+\frac{1}{2}\eta_k^2\|\boldsymbol{g}_k^s-\boldsymbol{g}_k^0\|^2+\frac{1}{2}\eta_k^2\beta_k^2\|\boldsymbol{L}\boldsymbol{x}_k\|^2\\
&+\frac{1}{2}\eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e-\boldsymbol{g}_k^0\|^2]+\frac{1}{2\beta_k^2}\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2]\\
=\;&\boldsymbol{x}_k^\top(\boldsymbol{K}-\eta_k\alpha_k\boldsymbol{L})\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&+\frac{1}{2}(\eta_k+\eta_k^2)\|\boldsymbol{g}_k^s-\boldsymbol{g}_k^0\|^2+\frac{1}{2}\eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e-\boldsymbol{g}_k^0\|^2]\\
&+\|\boldsymbol{x}_k\|^2_{\eta_k(\beta_k\boldsymbol{L}+\frac{1}{2}\boldsymbol{K})+\eta_k^2(\frac{1}{2}\alpha_k^2-\alpha_k\beta_k+\beta_k^2)\boldsymbol{L}^2}\\
&+\Big(\frac{1}{2\eta_k\beta_k^2}+\frac{3}{2\beta_k^2}\Big)\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2]\\
&-\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\eta_k(\beta_k-\frac{1}{2}-\eta_k\beta_k^2)\boldsymbol{K}}\\
\le\;&\boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)-(1+\omega_k)\eta_k\alpha_k\boldsymbol{x}_k^\top\boldsymbol{L}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&+\omega_k\eta_k\alpha_k\boldsymbol{x}_k^\top\boldsymbol{L}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&+\|\boldsymbol{x}_k\|^2_{\eta_k(\beta_k\boldsymbol{L}+\frac{1}{2}\boldsymbol{K})+\eta_k^2(\frac{1}{2}\alpha_k^2-\alpha_k\beta_k+\beta_k^2)\boldsymbol{L}^2+\eta_k(1+3\eta_k)L_f^2\boldsymbol{K}}\\
&+\frac{\eta_k}{2\beta_k^2}(1+3\eta_k)L_f^2\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]+nL_f^2\eta_k(1+3\eta_k)\delta_k^2\\
&+\eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2]-\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\eta_k(\beta_k-\frac{1}{2}-\eta_k\beta_k^2)\boldsymbol{K}},
\end{aligned}
\tag{94}
$$

where the first equality holds due to (76); the second equality holds since (34a), (69a), and that $x_{i,k}$ and $v_{i,k}$ are independent of $\mathfrak{L}_k$; the first inequality holds due to the Cauchy–Schwarz inequality, the Jensen's inequality, (34a), and $\rho(\boldsymbol{K})=1$; and the last inequality holds due to (69b), (69e), and (69f).

For the third term in the right-hand side of (94), we have

$$\begin{aligned}\omega_k\eta_k\alpha_k\boldsymbol{x}_k^\top\boldsymbol{L}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)&=\omega_k\eta_k\alpha_k\boldsymbol{x}_k^\top\boldsymbol{L}\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\&\le\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\omega_k\eta_k\alpha_k\boldsymbol{L}^2}+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\frac{1}{2}\omega_k\eta_k\alpha_k\boldsymbol{K}}.\end{aligned}\tag{95}$$

Then, from (93)–(95), we have

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[W_{3,k+1}]\le{}&W_{3,k}-(1+\omega_k)\eta_k\alpha_k\boldsymbol{x}_k^\top\boldsymbol{L}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\&+\|\boldsymbol{x}_k\|^2_{\eta_k(\beta_k\boldsymbol{L}+\frac{1}{2}\boldsymbol{K})+\eta_k^2(\frac{1}{2}\alpha_k^2-\alpha_k\beta_k+\beta_k^2)\boldsymbol{L}^2}+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\omega_k\eta_k\alpha_k\boldsymbol{L}^2+\eta_k(1+3\eta_k)L_f^2\boldsymbol{K}}\\&+\frac{\eta_k}{2\beta_k^2}(1+3\eta_k)L_f^2\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]+nL_f^2\eta_k(1+3\eta_k)\delta_k^2+\eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2]\\&-\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\eta_k(\beta_k-\frac{1}{2}-\eta_k\beta_k^2-\frac{1}{2}\omega_k\alpha_k)\boldsymbol{K}}+\frac{1}{2}\omega_k\mathbf{E}_{\mathfrak{L}_k}[2W_{1,k+1}+\|\boldsymbol{g}_{k+1}^0\|^2].\end{aligned}\tag{96}$$

(iv) We have

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[W_{4,k+1}]&=\mathbf{E}_{\mathfrak{L}_k}[\tilde{f}(\bar{\boldsymbol{x}}_{k+1})-nf^*]\\&=\mathbf{E}_{\mathfrak{L}_k}[\tilde{f}(\bar{\boldsymbol{x}}_k)-nf^*+\tilde{f}(\bar{\boldsymbol{x}}_{k+1})-\tilde{f}(\bar{\boldsymbol{x}}_k)]\\&\le\mathbf{E}_{\mathfrak{L}_k}[\tilde{f}(\bar{\boldsymbol{x}}_k)-nf^*-\eta_k(\bar{\boldsymbol{g}}_k^e)^\top\boldsymbol{g}_k^0+0.5\eta_k^2L_f\|\bar{\boldsymbol{g}}_k^e\|^2]\\&=W_{4,k}-\eta_k(\bar{\boldsymbol{g}}_k^s)^\top\boldsymbol{g}_k^0+0.5\eta_k^2L_f\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]\\&=W_{4,k}-\eta_k(\bar{\boldsymbol{g}}_k^s)^\top\bar{\boldsymbol{g}}_k^0+0.5\eta_k^2L_f\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]\\&=W_{4,k}-0.5\eta_k(\bar{\boldsymbol{g}}_k^s)^\top(\bar{\boldsymbol{g}}_k^s+\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k^s)\\&\quad-0.5\eta_k(\bar{\boldsymbol{g}}_k^s-\bar{\boldsymbol{g}}_k^0+\bar{\boldsymbol{g}}_k^0)^\top\bar{\boldsymbol{g}}_k^0+0.5\eta_k^2L_f\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]\\&\le W_{4,k}-0.25\eta_k(\|\bar{\boldsymbol{g}}_k^s\|^2-\|\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k^s\|^2+\|\bar{\boldsymbol{g}}_k^0\|^2\\&\quad-\|\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k^s\|^2)+0.5\eta_k^2L_f\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]\\&=W_{4,k}-0.25\eta_k\|\bar{\boldsymbol{g}}_k^s\|^2+0.5\eta_k\|\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k^s\|^2\\&\quad-0.25\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2+0.5\eta_k^2L_f\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]\\&\le W_{4,k}-0.25\eta_k\|\bar{\boldsymbol{g}}_k^s\|^2+\|\boldsymbol{x}_k\|^2_{\eta_kL_f^2\boldsymbol{K}}\\&\quad+nL_f^2\eta_k\delta_k^2-0.25\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2+0.5\eta_k^2L_f\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2],\end{aligned}\tag{97}$$

where the first inequality holds since that $\tilde{f}$ is smooth as shown in (71), (37) and (78); the third equality holds since (69a) and that $x_{i,k}$ and $v_{i,k}$ are independent of $\mathfrak{L}_k$; the fourth equality holds due to $(\bar{\boldsymbol{g}}_k^s)^\top \boldsymbol{g}_k^0 = (\boldsymbol{g}_k^s)^\top \boldsymbol{H}\boldsymbol{g}_k^0 = (\boldsymbol{g}_k^s)^\top \boldsymbol{H}\boldsymbol{H}\boldsymbol{g}_k^0 = (\bar{\boldsymbol{g}}_k^s)^\top \bar{\boldsymbol{g}}_k^0$; the second inequality holds due to the Cauchy–Schwarz inequality; and the last inequality holds due to (69c).

(v) We have

$$\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}] \le{}& W_k + \frac{1}{2}\omega_k\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - (1+\omega_k)\|\boldsymbol{x}_k\|^2_{\eta_k\alpha_k\boldsymbol{L}-\frac{1}{2}\eta_k\boldsymbol{K}-\frac{3}{2}\eta_k^2\alpha_k^2\boldsymbol{L}^2-\eta_k(1+5\eta_k)L_f^2\boldsymbol{K}}\\
&+ (1+\omega_k)\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\frac{3}{2}\eta_k^2\beta_k^2\boldsymbol{K}} + (1+\omega_k)nL_f^2\eta_k(1+5\eta_k)\delta_k^2\\
&+ 2(1+\omega_k)\eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2] + \frac{1}{2}(\eta_k+\omega_k+\eta_k\omega_k)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}\\
&+ \|\boldsymbol{x}_k\|^2_{(1+\omega_k)\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)} + \frac{\eta_k}{\beta_k^2}\Big(\eta_k+\frac{1}{2}\Big)(1+\omega_k)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)L_f^2\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]\\
&+ \frac{1}{2}\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)(\omega_k+\omega_k^2)\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_{k+1}^0\|^2] + \|\boldsymbol{x}_k\|^2_{\eta_k(\beta_k\boldsymbol{L}+\frac{1}{2}\boldsymbol{K})+\eta_k^2(\frac{1}{2}\alpha_k^2-\alpha_k\beta_k+\beta_k^2)\boldsymbol{L}^2}\\
&+ \|\boldsymbol{x}_k\|^2_{\frac{1}{2}\omega_k\eta_k\alpha_k\boldsymbol{L}^2+\eta_k(1+3\eta_k)L_f^2\boldsymbol{K}} + \frac{\eta_k}{2\beta_k^2}(1+3\eta_k)L_f^2\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]\\
&+ nL_f^2\eta_k(1+3\eta_k)\delta_k^2 + \eta_k^2\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2] - \Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\eta_k(\beta_k-\frac{1}{2}-\eta_k\beta_k^2-\frac{1}{2}\omega_k\alpha_k)\boldsymbol{K}}\\
&+ \frac{1}{2}\omega_k\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_{k+1}^0\|^2] - \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^s\|^2 + \|\boldsymbol{x}_k\|^2_{\eta_kL_f^2\boldsymbol{K}}\\
&+ nL_f^2\eta_k\delta_k^2 - \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2 + \frac{1}{2}\eta_k^2L_f\mathbf{E}_{\mathfrak{L}_k}[\|\bar{\boldsymbol{g}}_k^e\|^2]\\
\le{}& W_k - \|\boldsymbol{x}_k\|^2_{\eta_k\boldsymbol{M}_{1,k}-\eta_k^2\boldsymbol{M}_{2,k}-\omega_k\boldsymbol{M}_3-b_{1,k}\boldsymbol{K}} - \Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{b_{2,k}^0\boldsymbol{K}}\\
&- \eta_k\Big(\frac{1}{4} - (b_{3,k}+6pb_{4,k})\eta_k\Big)\|\bar{\boldsymbol{g}}_k^0\|^2\\
&+ 2pn\sigma_1^2b_{4,k}\eta_k^2 + n\sigma_2^2(b_{3,k}+6pb_{4,k})\eta_k^2 + b_{5,k}\eta_k\delta_k^2, \qquad (98)
\end{aligned}$$

where the first inequality holds due to (87), (92), (96), (97), and $\alpha_k = \kappa_1\beta_k$; the last inequality holds due to (70a), (70b), $\alpha_k = \kappa_1\beta_k$, $\eta_k = \frac{\kappa_2}{\beta_k}$, and

$$\begin{aligned}
\boldsymbol{M}_{1,k} &= (\alpha_k-\beta_k)\boldsymbol{L} - (1+3L_f^2)\boldsymbol{K},\\
\boldsymbol{M}_{2,k} &= \beta_k^2\boldsymbol{L} + (2\alpha_k^2+\beta_k^2)\boldsymbol{L}^2 + 8L_f^2\boldsymbol{K} + 12p(3+0.5L_f)L_f^2\boldsymbol{K},\\
\boldsymbol{M}_3 &= 0.5\boldsymbol{K} - \kappa_1\kappa_2\boldsymbol{L} + 0.5\kappa_1\kappa_2\boldsymbol{L}^2 + 1.5\kappa_1^2\kappa_2^2\boldsymbol{L}^2 + \kappa_2^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2),\\
b_{2,k}^0 &= 0.5\eta_k(2\beta_k-\kappa_3) - 2.5\kappa_2^2 - 0.5\omega_k(\kappa_1\kappa_2+3\kappa_2^2+\kappa_4) - 0.5\omega_k\eta_k\kappa_4.
\end{aligned}$$

From (34b), $\alpha_k = \kappa_1\beta_k$, $\kappa_1 > 1$, $\beta_k \geq \varepsilon_1 \geq 1 + 3L_f^2$, and $\eta_k = \frac{\kappa_2}{\beta_k}$, we have

$$\eta_k \boldsymbol{M}_{1,k} \geq \varepsilon_2\kappa_2\boldsymbol{K}. \tag{99}$$

From (34b), $\alpha_k = \kappa_1\beta_k$, $\beta_k \geq \varepsilon_1 \geq (8 + 12p(3 + 0.5L_f))^{1/2}L_f$, and $\eta_k = \frac{\kappa_2}{\beta_k}$, we have

$$\eta_k^2 \boldsymbol{M}_{2,k} \leq \varepsilon_3\kappa_2^2\boldsymbol{K}. \tag{100}$$

From (34b), $\alpha_k = \kappa_1\beta_k$, and $\eta_k = \frac{\kappa_2}{\beta_k}$, we have

$$\boldsymbol{M}_3 \leq \varepsilon_5\boldsymbol{K}. \tag{101}$$

From $\beta_k \geq \varepsilon_1 \geq p\kappa_3 \geq \kappa_3$ and $\eta_k = \frac{\kappa_2}{\beta_k}$, we have

$$b_{2,k}^0 \geq b_{2,k}. \tag{102}$$

From (98)–(102), we know that (86a) holds.

Similar to the way to get (86a), we have (86b). ■

Lemma 8. *Suppose Assumptions 1–5 hold. Suppose $\alpha_k = \kappa_1\beta_k$, $\beta_k = \kappa_0(k+t_1)^\theta$, and $\eta_k = \frac{\kappa_2}{\beta_k}$, where $\theta \in [0,1]$, $\kappa_0 \geq c_0(\kappa_1,\kappa_2)/t_1^\theta$, $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, and $t_1 \geq (c_3(\kappa_1,\kappa_2))^{1/\theta}$. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1, then*

$$\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}] \leq W_k - \varepsilon_4\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 - \varepsilon_6\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2 - \frac{1}{16}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2 + pn\varepsilon_{12}\eta_k^2 + pn\varepsilon_{11}\eta_k\delta_k^2, \tag{103a}$$

$$\mathbf{E}_{\mathfrak{L}_k}[\breve{W}_{k+1}] \leq \breve{W}_k - \varepsilon_4\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 - \varepsilon_6\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2 + p\varepsilon_{13}\eta_k^2\|\bar{\boldsymbol{g}}_k^0\|^2 + pn\varepsilon_{12}\eta_k^2 + pn\varepsilon_{11}\eta_k\delta_k^2, \tag{103b}$$

$$\mathbf{E}_{\mathfrak{L}_k}[W_{4,k+1}] \leq W_{4,k} + \|\boldsymbol{x}_k\|_{2\eta_k L_f^2\boldsymbol{K}}^2 - \frac{3}{16}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2 + 2p\eta_k^2L_f(\sigma_1^2 + 3\sigma_2^2) + (n+p)L_f^2\eta_k\delta_k^2. \tag{103c}$$

Proof: (i) Noting that $\kappa_1 > c_1 > 1$ and $\beta_k = \kappa_0(k+t_1)^\theta \geq \kappa_0 t_1^\theta \geq c_0(\kappa_1,\kappa_2,t_1,\theta) \geq \varepsilon_1 \geq 1$, we know that all conditions needed in Lemma 7 are satisfied, so (86a) and (86b) hold.

From $\kappa_1 > c_1 = \frac{1}{\rho_2(L)} + 1$, we have

$$\varepsilon_2 > 0. \tag{104}$$

From (104) and $\kappa_2 \in (0, \min\{\frac{\varepsilon_2}{\varepsilon_3},\ \frac{1}{5}\})$, we have

$$\varepsilon_4 > 0 \text{ and } \varepsilon_6 > 0. \tag{105}$$

From $t_1 \geq (c_3(\kappa_1, \kappa_2))^{1/\theta}$ and $c_3(\kappa_1, \kappa_2) = \frac{24\kappa_3}{\kappa_2}$, we have

$$\frac{3\kappa_3}{2\kappa_2 t_1^\theta} \leq \frac{1}{16}. \tag{106}$$

From $\kappa_0 \geq \frac{24\kappa_4}{\kappa_2 t_1^\theta} \geq \frac{24\kappa_4}{\kappa_2 t_1^{3\theta}}$, we have

$$\frac{3\kappa_4}{2\kappa_2\kappa_0 t_1^{3\theta}} \leq \frac{1}{16}. \tag{107}$$

From $\beta_k = \kappa_0(k + t_1)^\theta$, we have

$$\begin{aligned}\omega_k &= \frac{1}{\beta_k} - \frac{1}{\beta_{k+1}} = \frac{1}{\kappa_0}\Big(\frac{1}{(k+t_1)^\theta} - \frac{1}{(k+t_1+1)^\theta}\Big)\\ &\leq \frac{1}{\kappa_0(k+t_1)^\theta(k+t_1+1)^\theta} \leq \frac{\kappa_0}{\beta_k^2} \leq 1.\end{aligned} \tag{108}$$

From (108), $\eta_k = \frac{\kappa_2}{\beta_k}$, $\beta_k \geq 1$, $\omega_k \leq 1$, and $\kappa_0 \geq (\frac{2p\varepsilon_7}{\varepsilon_4 t_1^{2\theta}})^{\frac{1}{2}}$, we have

$$b_{1,k} \leq \frac{p\varepsilon_7}{\kappa_0^2 t_1^{2\theta}} \leq \frac{\varepsilon_4}{2}. \tag{109}$$

From (108), (109), $\kappa_0 \geq \frac{2\varepsilon_5}{\varepsilon_4 t_1^\theta}$, and (105), we have

$$2\varepsilon_4 - \varepsilon_5\omega_k - b_{1,k} \geq 2\varepsilon_4 - \frac{\varepsilon_5}{\kappa_0 t_1^\theta} - \frac{\varepsilon_4}{2} \geq \varepsilon_4 > 0. \tag{110}$$

From (108), $\eta_k = \frac{\kappa_2}{\beta_k}$, $\kappa_0 \geq \frac{\varepsilon_8}{2\varepsilon_6 t_1^\theta} \geq \frac{\varepsilon_8}{2\varepsilon_6 t_1^{2\theta}}$, and (105), we have

$$b_{2,k} \geq 2\varepsilon_6 - \frac{\varepsilon_8}{2\kappa_0 t_1^{2\theta}} \geq \varepsilon_6 > 0. \tag{111}$$

From (106)–(108) and $\eta_k = \frac{\kappa_2}{\beta_k}$, we have

$$b_{3,k}\eta_k \leq \frac{3\kappa_3}{2\kappa_2\kappa_0 t_1^{3\theta}} + \frac{3\kappa_4}{2\kappa_2\kappa_0 t_1^{3\theta}} \leq \frac{1}{8}. \tag{112}$$

From $\beta_k \geq 1$ and $\omega_k \leq 1$, we have

$$b_{3,k} \leq \varepsilon_9, \tag{113a}$$

$$b_{4,k} \leq \varepsilon_{10}. \tag{113b}$$

From (112), (113b), and $\kappa_0 \geq \frac{96p\kappa_2\varepsilon_{10}}{t_1^\theta}$, we have

$$\frac{1}{4} - (b_{3,k} + 6pb_{4,k})\eta_k \geq \frac{1}{8} - 6pb_{4,k}\eta_k \geq \frac{1}{8} - \frac{6p\kappa_2\varepsilon_{10}}{\kappa_0 t_1^\theta} \geq \frac{1}{16}. \tag{114}$$

From (114), $\eta_k = \frac{\kappa_2}{\beta_k}$, $\beta_k \geq 1$, and $\omega_k \leq 1$, we have

$$b_{5,k} \leq pn\varepsilon_{11}. \tag{115}$$

From (86a), (110), (111), and (113a)–(115), we know that (103a) holds.

(ii) From (86b), (110), (111), (113a), (113b), and (115), we have (103b).

(iii) From (97), (69d), and (70a), we have

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[W_{4,k+1}] \leq W_{4,k} &- \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^s\|^2 + \|\boldsymbol{x}_k\|^2_{\eta_k L_f^2 \boldsymbol{K}} + nL_f^2\eta_k\delta_k^2 \\ &- \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2 + \frac{1}{2}\eta_k^2 L_f\Big(\frac{12p}{n}\|\bar{\boldsymbol{g}}_k^0\|^2 + \frac{12p}{n}L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} \\ &+ 4p\sigma_1^2 + 12p\sigma_2^2 + \frac{1}{2}p^2L_f^2\delta_k^2 + \|\bar{\boldsymbol{g}}_k^s\|^2\Big).\end{aligned} \tag{116}$$

From $\kappa_0 t_1^\theta \geq c_0(\kappa_1,\kappa_2) \geq 96p\kappa_2\varepsilon_{10} > 96p\kappa_2 L_f$, we have

$$\frac{6p}{n}\eta_k^2 L_f \leq \frac{6}{\kappa_0 t_1^\theta}p\eta_k L_f\kappa_2 < \frac{1}{16}\eta_k, \tag{117a}$$

$$\frac{6p}{n}\eta_k^2 L_f^3 < \frac{1}{16}\eta_k L_f^2, \tag{117b}$$

$$\frac{1}{2}\eta_k^2 L_f < \frac{1}{16}\eta_k, \tag{117c}$$

$$\frac{1}{4}p^2\eta_k^2 L_f^3 < pL_f^2\eta_k. \tag{117d}$$

From (116)–(117d), we have (103c). ■

Now it is ready to prove Theorem 1.

Denote

$$\hat{V}_k = \|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}} + n(f(\bar{x}_k) - f^*).$$

We have

$$\begin{aligned}W_k &= \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{2}\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}} + \boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big) + n(f(\bar{x}_k) - f^*) \\ &\geq \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{2}\Big(\frac{1}{\rho(L)} + \kappa_1\Big)\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}} - \frac{1}{2\kappa_1}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \frac{1}{2}\kappa_1\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}} + n(f(\bar{x}_k) - f^*)\end{aligned}$$

$$\geq \kappa_7\Big(\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}\Big) + n(f(\bar{x}_k) - f^*) \tag{118}$$

$$\geq \kappa_7\hat{V}_k \geq 0, \tag{119}$$

where the first inequality holds due to (35b) and the Cauchy–Schwarz inequality; and the last inequality holds due to $0 < \kappa_7 < \frac{1}{2}$. Similarly, we have

$$W_k \leq \kappa_6\hat{V}_k. \tag{120}$$

From (103a) and (105), we have

$$\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}] \leq W_k - \varepsilon_4\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \frac{1}{16}\eta_k\|\bar{\boldsymbol{g}}^0_k\|^2 + pn\varepsilon_{12}\eta_k^2 + pn\varepsilon_{11}\eta_k\delta_k^2. \tag{121}$$

Then, taking expectation in $\mathcal{L}_T$, summing (121) over $k \in [0, T]$, and using (43) and $\eta_k = \frac{\kappa_2}{\kappa_0(k+t_1)^\theta}$ and $\delta_k \leq \kappa_\delta\sqrt{\eta_k}$ as stated in (7), yield

$$\mathbf{E}[W_{T+1}] + \sum_{k=0}^{T}\mathbf{E}\Big[\varepsilon_4\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{16}\eta_k\|\bar{\boldsymbol{g}}^0_k\|^2\Big] \leq W_0 + \frac{pn(\varepsilon_{11}\kappa_\delta^2 + \varepsilon_{12})\kappa_2^2}{\kappa_0^2}\sum_{k=0}^{T}\frac{1}{(k+t_1)^{2\theta}} \leq n\varepsilon_{14}. \tag{122}$$

Noting that $t_1^\theta = \mathcal{O}(\sqrt{p})$, we have

$$\kappa_0 = \mathcal{O}(\frac{p}{t_1^\theta}) = \mathcal{O}(\sqrt{p}). \tag{123}$$

From $W_0 = \mathcal{O}(n)$ and (123), we have

$$\varepsilon_{14} = \frac{W_0}{n} + \frac{2\theta p(\varepsilon_{11}\kappa_\delta^2 + \varepsilon_{12})\kappa_2^2}{(2\theta - 1)\kappa_0^2} = \mathcal{O}(1). \tag{124}$$

From (122), (119), (105), and $\sum_{k=0}^T \eta_k = \sum_{k=0}^T \frac{\kappa_2}{\kappa_0(k+t_1)^\theta} \geq \frac{\kappa_2(T+t_1)^{1-\theta}}{\kappa_0(1-\theta)}$, we have

$$\frac{\sum_{k=0}^T \eta_k\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2]}{\sum_{k=0}^T \eta_k} = \frac{\sum_{k=0}^T \eta_k\mathbf{E}[\|\bar{\boldsymbol{g}}^0_k\|^2]}{n\sum_{k=0}^T \eta_k} \leq \frac{16\kappa_0(1-\theta)\varepsilon_{14}}{\kappa_2(T+t_1)^{1-\theta}}. \tag{125}$$

From (125), (124), and (123), we have (8a).

From (122), (118), and (105), we have

$$\mathbf{E}[f(\bar{x}_{T+1})] - f^* = \frac{1}{n}W_{T+1} \leq \varepsilon_{14},\ \forall T \in \mathbb{N}_0, \tag{126}$$

which gives (8b).

From (122), (119), and (105), we have

$$\sum_{k=0}^{T} \mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}] \leq \frac{n\varepsilon_{14}}{\varepsilon_4}, \ \forall T \in \mathbb{N}_0. \tag{127}$$

From (69g) and (126), we have

$$\|\bar{\boldsymbol{g}}^0_k\|^2 \leq 2nL_f(f(\bar{x}_k) - f^*) \leq 2nL_f\varepsilon_{14}. \tag{128}$$

From (70a), (127), and (128), we know that $\mathbf{E}[\|\boldsymbol{g}^e_k\|^2]$ is bounded. Then, same as the proof of the first part of Theorem 1 in [57], we have (8d).

From (118) and (120), we have

$$0 \leq 2\kappa_7(W_{1,k} + W_{2,k}) \leq \breve{W}_k \leq 2\kappa_6(W_{1,k} + W_{2,k}). \tag{129}$$

Denote $\breve{z}_k = \mathbf{E}[\breve{W}_k]$. From (103b), (128), (129), and (7), we have

$$\breve{z}_{k+1} \leq (1 - a_1)\breve{z}_k + \frac{a_2}{(t + t_1)^{2\theta}}. \tag{130}$$

From $\kappa_1 > 1$, we have $\kappa_6 > 1$. From $0 < \kappa_2 < \frac{1}{5}$, we have $\varepsilon_6 = \frac{1}{4}(\kappa_2 - 5\kappa_2^2) \leq \frac{1}{80}$. Thus,

$$a_1 \leq \frac{\varepsilon_6}{\kappa_6} \leq \frac{1}{80}. \tag{131}$$

From (105), we know that

$$a_1 > 0 \text{ and } a_2 > 0. \tag{132}$$

From (130)–(132) and (51), we have

$$\breve{z}_k \leq \phi_3(k, t_1, a_1, a_2, 2\theta, \breve{z}_0), \ \forall k \in \mathbb{N}_+, \tag{133}$$

where the function ϕ_3 is defined in (52).

Noting that $\phi_3(k, t_1, a_1, a_2, 2\theta, \breve{z}_0) = \mathcal{O}(n/k^{2\theta})$, from (133) and (129), we have (8c).

C. Proof of Theorem 2

In addition to the notations defined in Appendix B, we also denote the following notations.

$$\tilde{c}_0(\kappa_1, \kappa_2) = \max\Big\{\varepsilon_1, \ \Big(\frac{p\tilde{\varepsilon}_7}{\varepsilon_4}\Big)^{\frac{1}{3}}, \ 48p\kappa_2\tilde{\varepsilon}_{10}\Big\},$$

$$\begin{aligned}
&\tilde{\varepsilon}_7 = 6(1+3\kappa_2+\kappa_4+2\kappa_2\kappa_4)\kappa_2 L_f^4,\\
&\tilde{\varepsilon}_{10} = 6 + L_f + \frac{1}{\kappa_2}(\kappa_4+1)L_f^2 + (3\kappa_4+3)L_f^2,\\
&\tilde{\varepsilon}_{11} = L_f^2\Big(\frac{1}{192} + \frac{1}{p}(8\kappa_2+3)\Big),\\
&\tilde{\varepsilon}_{12} = 2(\sigma_1^2+3\sigma_2^2)\tilde{\varepsilon}_{10},\\
&\varepsilon_{15} = 2(\sigma_1^2+3\sigma_2^2)L_f,\\
&\varepsilon_{16} = 2L_f^2\kappa_\delta^2.
\end{aligned}$$

To prove Theorem 2, the following lemma is used.

Lemma 9. *Suppose Assumptions 1–5 hold. Suppose* $\alpha_k = \alpha = \kappa_1\beta$*,* $\beta_k = \beta$*, and* $\eta_k = \eta = \frac{\kappa_2}{\beta}$*, where* $\beta \ge \tilde{c}_0(\kappa_1,\kappa_2)$*,* $\kappa_1 > c_1$*, and* $\kappa_2 \in (0, c_2(\kappa_2))$ *are constants. Let* $\{\boldsymbol{x}_k\}$ *be the sequence generated by Algorithm 1, then*

$$\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}] \le W_k - \varepsilon_4\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - 2\varepsilon_6\Big\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}} - \frac{1}{8}\eta\|\bar{\boldsymbol{g}}_k^0\|^2 + pn\tilde{\varepsilon}_{12}\eta^2 + pn\tilde{\varepsilon}_{11}\eta\delta_k^2, \tag{134a}$$

$$\mathbf{E}_{\mathfrak{L}_k}[W_{4,k+1}] \le W_{4,k} + \|\boldsymbol{x}_k\|^2_{2\eta L_f^2\boldsymbol{K}} - \frac{1}{8}\eta\|\bar{\boldsymbol{g}}_k^0\|^2 + p\varepsilon_{15}\eta^2 + (n+p)L_f^2\eta\delta_k^2. \tag{134b}$$

Proof: (i) Substituting $\alpha_k = \alpha = \kappa_1\beta$, $\beta_k = \beta$, $\eta_k = \eta = \frac{\kappa_2}{\beta}$, and $\omega_k = 0$ into (87), (92), (96), and (97), similar to the way to get (98), we have

$$\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}] \le{}& W_k - \|\boldsymbol{x}_k\|^2_{\eta\tilde{\boldsymbol{M}}_1 - \eta^2\tilde{\boldsymbol{M}}_2 - \tilde{b}_1\boldsymbol{K}} - \Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\tilde{b}_2^0\boldsymbol{K}}\\
&- \eta\Big(\frac{1}{4} - 6p\tilde{b}_4\eta\Big)\|\bar{\boldsymbol{g}}_k^0\|^2 + 2pn(\sigma_1^2+3\sigma_2^2)\tilde{b}_4\eta^2 + \tilde{b}_5\eta\delta_k^2,
\end{aligned} \tag{135}$$

where

$$\begin{aligned}
&\tilde{\boldsymbol{M}}_1 = (\alpha-\beta)\boldsymbol{L} - (1+3L_f^2)\boldsymbol{K},\\
&\tilde{\boldsymbol{M}}_2 = \beta^2\boldsymbol{L} + (2\alpha^2+\beta^2)\boldsymbol{L}^2 + 8L_f^2\boldsymbol{K} + (3+0.5L_f)\frac{12p}{n}L_f^2\boldsymbol{K},\\
&\tilde{b}_1 = \frac{6p}{n}\kappa_3 L_f^4\frac{\eta}{\beta^2} + \frac{12p}{n}\kappa_5 L_f^4\frac{\eta^2}{\beta^2},\\
&\tilde{b}_2^0 = \frac{1}{2}\eta(2\beta-\kappa_3) - \frac{5}{2}\kappa_2^2,\\
&\tilde{b}_4 = 6 + L_f + \frac{\kappa_3}{\kappa_2}L_f^2\frac{1}{\beta} + 2\kappa_5 L_f^2\frac{1}{\beta^2},
\end{aligned}$$

$$\tilde{b}_5 = nL_f^2\Big(\frac{1}{4}p^2\tilde{b}_4\eta + 3 + 8\eta\Big).$$

From (135), similar to the way to get (86a), we have

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}] \le{}& W_k - \|\boldsymbol{x}_k\|^2_{(2\varepsilon_4-\tilde{b}_1)\boldsymbol{K}} - \|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\|^2_{2\varepsilon_6\boldsymbol{K}}\\ &- \eta\Big(\frac{1}{4} - 6p\tilde{b}_4\eta\Big)\|\bar{\boldsymbol{g}}_k^0\|^2 + 2pn(\sigma_1^2 + 3\sigma_2^2)\tilde{b}_4\eta^2 + \tilde{b}_5\eta\delta_k^2.\end{aligned} \tag{136}$$

From (136), similar to the way to get (103a), we have (134a).

(ii) Noting $\eta_k = \eta$, $\beta \ge 48p\kappa_2\tilde{\varepsilon}_{10} \ge 48p\kappa_2 L_f$, and $\eta = \kappa_2/\beta$, similar to the way to get (103c), we have (134b). ■

We are now ready to prove Theorem 2.

From $\beta_k = \beta = \kappa_2\sqrt{pT}/\sqrt{n}$ and $T > n(\tilde{c}_0(\kappa_1,\kappa_2)/\kappa_2)^2/p$, we have $\beta \ge \tilde{c}_0(\kappa_1,\kappa_2)$. Thus, all conditions needed in Lemma 9 are satisfied. So (134a) and (134b) hold.

From (134a) and (9), similar to the way to get (127), we have

$$\frac{1}{T+1}\sum_{k=0}^{T}\mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2] \le \frac{1}{\varepsilon_4}\Big(\frac{W_0}{n(T+1)} + \frac{n\tilde{\varepsilon}_{12}}{T} + \frac{2n\tilde{\varepsilon}_{11}\kappa_\delta^2}{\sqrt{T(T+1)}}\Big), \tag{137}$$

which gives (10c).

From (134b) and (9), similar to the way to get (125), we have

$$\begin{aligned}\frac{1}{T+1}\sum_{k=0}^{T}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] ={}& \frac{1}{n(T+1)}\sum_{k=0}^{T}\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2]\\ &\le 8\Big(\frac{W_{4,0}}{n(T+1)\eta} + \frac{2L_f^2}{n(T+1)}\sum_{k=0}^{T}\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}] + \frac{p\varepsilon_{15}\eta}{n} + \frac{\sqrt{p}\varepsilon_{16}}{\sqrt{n(T+1)}}\Big).\end{aligned} \tag{138}$$

Noting that $\eta = \kappa_2/\beta_k = \sqrt{n}/\sqrt{pT}$, and $n/T < \sqrt{p}/\sqrt{nT}$ due to $T > n^3/p$, from (138) and (137), we have

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] = 8(f(\bar{x}_0) - f^* + 2(\sigma_1^2 + 3\sigma_2^2)L_f + 2L_f^2\kappa_\delta^2)\frac{\sqrt{p}}{\sqrt{nT}} + \mathcal{O}(\frac{n}{T}),$$

which gives (10a).

Taking expectation in $\mathcal{L}_T$, summing (134b) over $k \in [0, T]$, and using (9) yield

$$n(\mathbf{E}[f(\bar{x}_{T+1})] - f^*) = \mathbf{E}[W_{4,T+1}]$$

$$\leq W_{4,0}+\frac{2\sqrt{n}}{\sqrt{pT}}L_f^2\sum_{k=0}^{T}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+2nL_f(\sigma_1^2+3\sigma_2^2)\frac{T+1}{T}+2nL_f^2\kappa_\delta^2\sqrt{\frac{T+1}{T}}. \tag{139}$$

Noting that $W_{4,0}=\mathcal{O}(n)$ and $\sqrt{n}n/\sqrt{pT}<1$ due to $T>n^3/p$, from (137) and (139), we have (10b).

Similar to the proof of (8d), we have (10d).

D. Proof of Theorem 3

In addition to the notations defined in Appendix B, we also denote the following notations.

$$\begin{aligned}
&\varepsilon_{17}=\frac{1}{\kappa_6}\min\Big\{\frac{\varepsilon_4\kappa_0 t_1^\theta}{\kappa_2},\ \frac{\varepsilon_6\kappa_0 t_1^\theta}{\kappa_2},\ \frac{\nu}{8}\Big\},\\
&\varepsilon_{18}=\frac{32\theta 4^\theta L_f(\sigma_1^2+3\sigma_2^2)\kappa_2}{3\nu\kappa_0},\\
&\breve{a}_2=pn(\varepsilon_{11}\kappa_\delta^2+\varepsilon_{12}+\varepsilon_{13}c_g)\frac{\kappa_2^2}{\kappa_0^2},\\
&a_3=\frac{\kappa_2\varepsilon_{17}}{\kappa_0},\\
&a_4=pn(\varepsilon_{11}\kappa_\delta^2+\varepsilon_{12})\frac{\kappa_2^2}{\kappa_0^2}.
\end{aligned}$$

All conditions needed in Lemma 8 are satisfied, so (103a)–(103c) hold.

From (2), we have that

$$\|\bar{\boldsymbol{g}}_k^0\|^2=n\|\nabla f(\bar{x}_k)\|^2\geq 2\nu n(f(\bar{x}_k)-f^*)=2\nu W_{4,k}. \tag{140}$$

From (103a), (140), (119), and (120), we have

$$\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}]&\leq W_k-\varepsilon_4\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\varepsilon_6\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\|^2_{\boldsymbol{K}}-\frac{\eta_k\nu n}{8}W_{4,k}+pn\varepsilon_{12}\eta_k^2+pn\varepsilon_{11}\eta_k\delta_k^2\\
&\leq W_k-\frac{\eta_k}{\kappa_6}\min\Big\{\frac{\varepsilon_4}{\eta_k},\ \frac{\varepsilon_6}{\eta_k},\ \frac{\nu}{8}\Big\}W_k+pn\varepsilon_{12}\eta_k^2+pn\varepsilon_{11}\eta_k\delta_k^2\\
&\leq W_k-\eta_k\varepsilon_{17}W_k+pn\varepsilon_{12}\eta_k^2+pn\varepsilon_{11}\eta_k\delta_k^2,\ \forall k\in\mathbb{N}_0.
\end{aligned} \tag{141}$$

Denote $z_k=\mathbf{E}[W_k]$, $r_{1,k}=\eta_k\varepsilon_{17}$, and $r_{2,k}=pn\varepsilon_{12}\eta_k^2+pn\varepsilon_{11}\eta_k\delta_k^2$. From (141), we have

$$z_{k+1}\leq(1-r_{1,k})z_k+r_{2,k},\ \forall k\in\mathbb{N}_0. \tag{142}$$

From (11), we have

$$r_{1,k} = \eta_k \varepsilon_{17} = \frac{a_3}{(k+t_1)^\theta}, \tag{143}$$

$$r_{2,k} = pn\varepsilon_{12}\eta_k^2 + pn\varepsilon_{11}\eta_k\delta_k^2 \leq \frac{a_4}{(k+t_1)^{2\theta}}. \tag{144}$$

From $\kappa_1 > 1$, we have $\kappa_6 > 1$. From $0 < \kappa_2 < \frac{1}{5}$, we have $\varepsilon_6 = \frac{1}{4}(\kappa_2 - 5\kappa_2^2) \leq \frac{1}{80}$. Thus,

$$r_{1,k} \leq \frac{\varepsilon_6}{\kappa_6} \leq \frac{1}{80}. \tag{145}$$

From (105), we know that

$$a_3 > 0 \text{ and } a_4 > 0. \tag{146}$$

Then, from $\theta \in (0,1)$, (142)–(146), and (47), we have

$$z_k \leq \phi_1(k, t_1, a_3, a_4, \theta, 2\theta, z_0), \ \forall k \in \mathbb{N}_+, \tag{147}$$

where the function ϕ_1 is defined in (48).

From $t_1 \geq (pc_3(\kappa_1, \kappa_2))^{1/\theta}$, we have

$$t_1^\theta = \mathcal{O}(p). \tag{148}$$

From $\kappa_0 \geq c_0(\kappa_1, \kappa_2)/t_1^\theta$, $t_1 \leq (pc_4c_3(\kappa_1, \kappa_2))^{1/\theta}$, $c_0(\kappa_1, \kappa_2) \geq \varepsilon_1 \geq p\kappa_3$, and $c_3(\kappa_1, \kappa_2) = 24\kappa_3/\kappa_2$, we have

$$\frac{\kappa_2}{\kappa_0} \leq \frac{\kappa_2 t_1^\theta}{c_0(\kappa_1, \kappa_2)} \leq \frac{\kappa_2 pc_4c_3(\kappa_1, \kappa_2)}{p\kappa_3} \leq 24c_4. \tag{149}$$

Thus,

$$\phi_1(k, t_1, a_3, a_4, \theta, 2\theta, z_0) = \mathcal{O}(\frac{pn}{(k+t_1)^\theta}). \tag{150}$$

From (119), we have

$$\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + W_{4,k} \leq \hat{V}_k \leq \frac{W_k}{\kappa_7}. \tag{151}$$

From (69g), (147), (150), and (151), we get

$$\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2] = \mathcal{O}(\frac{pn}{(k+t_1)^\theta}), \ \forall k \in \mathbb{N}_+. \tag{152}$$

From (148) and (152), we know that there exists a constant $c_g > 0$, such that

$$\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2] \le nc_g, \ \forall k \in \mathbb{N}_0. \tag{153}$$

From (103b), (153), (129), and (11), we have

$$\breve{z}_{k+1} \le (1-a_1)\breve{z}_k + \frac{\breve{a}_2}{(t+t_1)^{2\theta}}. \tag{154}$$

Using (51), from (131) and (154), we have

$$\breve{z}_k \le \phi_3(k, t_1, a_1, \breve{a}_2, 2\theta, \breve{z}_0), \ \forall k \in \mathbb{N}_+, \tag{155}$$

where the function ϕ_3 is defined in (52). From (155), (129), (52), and (149), we have

$$\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}] \le \frac{1}{\kappa_7}\breve{z}_k \le \frac{1}{\kappa_7}\phi_3(k, t_1, a_1, \breve{a}_2, 2\theta, \breve{z}_0) = \mathcal{O}(\frac{pn}{(k+t_1)^{2\theta}}), \tag{156}$$

which yields (12a).

From (103c), (140), and $\delta_k \le \kappa_\delta \eta_k$ we have

$$\mathbf{E}[W_{4,k+1}] \le \mathbf{E}[W_{4,k}] - \frac{3\nu}{8}\eta_k\mathbf{E}[W_{4,k}] + \|\boldsymbol{x}_k\|^2_{2\eta_k L_f^2 \boldsymbol{K}} + 2pL_f(\sigma_1^2 + 3\sigma_2^2)\eta_k^2 + (n+p)L_f^2\kappa_\delta^2\eta_k^3. \tag{157}$$

Similar to the way to prove (47), from (156) and (157), we have

$$\mathbf{E}[f(\bar{x}_T) - f^*] \le \frac{\varepsilon_{18}p}{n(T+t_1)^\theta} + \mathcal{O}(\frac{p}{(T+t_1)^{2\theta}}). \tag{158}$$

From (149), we have

$$\varepsilon_{18} = \frac{32\theta 4^\theta L_f(\sigma_1^2 + 3\sigma_2^2)\kappa_2}{3\nu\kappa_0} \le \frac{256\theta 4^\theta L_f(\sigma_1^2 + 3\sigma_2^2)c_4}{\nu}. \tag{159}$$

Thus, from (158) and (159), we have (12b).

E. Proof of Theorem 4

In addition to the notations defined in Appendices B and D, we also denote the following notations.

$$\hat{c}_0(\kappa_1, \kappa_2) = \frac{\kappa_2}{8\kappa_6},$$

$$\hat{c}_3(\kappa_0, \kappa_1, \kappa_2) = \max\Big\{\frac{c_0(\kappa_1, \kappa_2)}{\kappa_0}, \ \frac{\kappa_6}{\varepsilon_4}, \ \frac{\kappa_6}{\varepsilon_6}, \ \frac{24\kappa_3}{\kappa_2}, \ p^{\frac{1}{\breve{a}_3}}\Big\},$$

$$\hat{a}_3 = \min\Big\{1,\ \frac{2}{3\kappa_6}\Big\},$$
$$\breve{a}_3 = pn(\varepsilon_{11}\kappa_\delta^2 + \varepsilon_{12} + \varepsilon_{13}\breve{c}_g)\frac{\kappa_2^2}{\kappa_0^2}.$$

From $t_1 > \hat{c}_3(\kappa_0, \kappa_1, \kappa_2) \geq \frac{c_0(\kappa_1,\kappa_2)}{\kappa_0}$, we have

$$\kappa_0 > \frac{c_0(\kappa_1, \kappa_2)}{t_1}.$$

Thus, all conditions needed in Lemma 8 are satisfied, so (142)–(146) still hold when $\theta = 1$.

From rom $t_1 > \hat{c}_3(\kappa_0, \kappa_1, \kappa_2) \geq \kappa_6/\varepsilon_4$, we have

$$\frac{\varepsilon_4 t_1}{\kappa_6} > 1. \tag{160}$$

From $t_1 > \hat{c}_3(\kappa_0, \kappa_1, \kappa_2) \geq \kappa_6/\varepsilon_6$, we have

$$\frac{\varepsilon_6 t_1}{\kappa_6} > 1. \tag{161}$$

From $\kappa_0 \in [3\hat{c}_0\nu\kappa_2/16, 3\nu\kappa_2/16)$, we have

$$\frac{16}{3\nu} < \frac{\kappa_2}{\kappa_0} \leq \frac{16}{3\hat{c}_0\nu}. \tag{162}$$

Thus,

$$\frac{\nu\kappa_2}{8\kappa_6\kappa_0} > \frac{2}{3\kappa_6}. \tag{163}$$

Hence, from (160), (161), and (163), we have

$$a_3 > \hat{a}_3. \tag{164}$$

Then from $\theta = 1$, (142)–(146), (164), and (49), we have

$$z_k \leq \phi_2(k, t_1, a_3, a_4, 2, z_0),\ \forall k \in \mathbb{N}_+, \tag{165}$$

where the function ϕ_2 is defined in (50).

From (164) and (162), we have $\phi_2(k, t_1, a_3, a_4, 2, z_0) = \mathcal{O}(pn/(k+t_1)^{\hat{a}_3})$. Hence, from (69g), (165), and (151), we get

$$\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2] = \mathcal{O}(\frac{pn}{(k+t_1)^{\hat{a}_3}}),\ \forall k \in \mathbb{N}_+. \tag{166}$$

Noting that $t_1 > \hat{c}_3(\kappa_0,\kappa_1,\kappa_2) \geq p^{1/\hat{a}_3}$, from (166), we know that there exists a constant $\breve{c}_g > 0$, such that

$$\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2] \leq n\breve{c}_g,\ \forall k \in \mathbb{N}_0. \tag{167}$$

From (103b), (167), (129), and (11), we have

$$\breve{z}_{k+1} \leq (1-a_1)\breve{z}_k + \frac{\breve{a}_3}{(t+t_1)^2}. \tag{168}$$

Using (51), from (131) and (168), we have

$$\breve{z}_k \leq \phi_3(k, t_1, a_1, \breve{a}_3, 2, \breve{z}_0),\ \forall k \in \mathbb{N}_+, \tag{169}$$

where the function ϕ_3 is defined in (52). From (169), (129), (52), and (162), we have

$$\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}] \leq \frac{1}{\kappa_7}\breve{z}_k \leq \frac{1}{\kappa_7}\phi_3(k, t_1, a_1, \breve{a}_3, 2, \breve{z}_0) = \mathcal{O}(\frac{pn}{(k+t_1)^2}), \tag{170}$$

which yields (14a).

From $\kappa_0 < 3\nu\kappa_2/16$, we have

$$\frac{3\nu\kappa_2}{8\kappa_0} > 2. \tag{171}$$

Same to the way to prove (49), from (170), (171), and (157), we have

$$\mathbf{E}[f(\bar{x}_T) - f^*] \leq \frac{\hat{\varepsilon}_{18}p}{n(T+t_1)} + \mathcal{O}(\frac{p}{(T+t_1)^2}). \tag{172}$$

From (162), we have

$$\hat{\varepsilon}_{18} = \frac{8L_f(\sigma_1^2+3\sigma_2^2)\kappa_2^2}{\kappa_0^2(\frac{3\nu\kappa_2}{8\kappa_0}-1)} \leq \frac{128L_f(\sigma_1^2+3\sigma_2^2)\kappa_2}{3\nu\kappa_0} \leq \frac{2048L_f(\sigma_1^2+3\sigma_2^2)}{9\hat{c}_0\nu^2}. \tag{173}$$

Thus, from (172) and (173), we have (14b).

F. Proof of Theorem 5

In addition to the notations defined in Appendices B, D, and E, we also denote the following notations.

$$\breve{c}_0(\kappa_1,\kappa_2) = \max\Big\{\varepsilon_1,\ \frac{2\varepsilon_5}{\varepsilon_4},\ \Big(\frac{2p\varepsilon_7}{\varepsilon_4}\Big)^{\frac{1}{2}},\ \frac{\varepsilon_8}{2\varepsilon_6},\ 4p\kappa_2\varepsilon_{10}\Big\},$$

$$
\begin{aligned}
\check{c}_3(\kappa_0,\kappa_1,\kappa_2) = \max\Big\{&\frac{\check{c}_0(\kappa_1,\kappa_2)}{\kappa_0},\ \frac{\kappa_6}{\varepsilon_4},\ \frac{\kappa_6}{\varepsilon_6},\ \Big(\frac{16L_f\kappa_3}{\nu\kappa_2}\Big)^{\frac{1}{3}},\ \Big(\frac{16L_f\kappa_4}{\nu\kappa_0\kappa_2}\Big)^{\frac{1}{3}},\ \frac{64pL_f\kappa_2\varepsilon_{10}}{\nu\kappa_0},\ p^{\frac{1}{\hat{a}_3}}\Big\},\\
\check{\varepsilon}_{12} &= 2\varepsilon_{10}\sigma_1^2 + \frac{1}{p}\varepsilon_9\tilde{\sigma}_2^2 + 6\varepsilon_{10}\tilde{\sigma}_2^2,\\
\tilde{\sigma}_2^2 &= 2L_f f^* - 2L_f\frac{1}{n}\sum_{i=1}^n f_i^*.
\end{aligned}
$$

To prove Theorem 5, the following lemma is used.

Lemma 10. *Suppose Assumptions 1–4 hold and each $f_i^* > -\infty$. Suppose $\alpha_k = \kappa_1\beta_k$, $\beta_k = \kappa_0(k+t_1)^\theta$, and $\eta_k = \frac{\kappa_2}{\beta_k}$, where $\theta \in [0,1]$, $\kappa_0 \ge \check{c}_0(\kappa_1,\kappa_2)/t_1^\theta$, $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, and $t_1 \ge 1$. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1, then*

$$
\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[W_{k+1}] \le{}& W_k - \varepsilon_4\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \varepsilon_6\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}} - \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2\\
&+ \frac{4}{3}L_f(b_{3,k} + 6pb_{4,k})\eta_k^2W_{4,k} + pn\check{\varepsilon}_{12}\eta_k^2 + pn\varepsilon_{11}\eta_k\delta_k^2, \qquad (174a)\\
\mathbf{E}_{\mathfrak{L}_k}[\check{W}_{k+1}] \le{}& \check{W}_k - \varepsilon_4\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \varepsilon_6\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}\\
&+ \frac{4}{3}L_fp\varepsilon_{13}\eta_k^2W_{4,k} + pn\check{\varepsilon}_{12}\eta_k^2 + pn\varepsilon_{11}\eta_k\delta_k^2, \qquad (174b)\\
\mathbf{E}_{\mathfrak{L}_k}[W_{4,k+1}] \le{}& W_{4,k} + \|\boldsymbol{x}_k\|^2_{2\eta_kL_f^2\boldsymbol{K}} - \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2 + \frac{8p}{n}L_f^2\eta_k^2W_{4,k}\\
&+ 2p\eta_k^2L_f(\sigma_1^2 + 2\tilde{\sigma}_2^2) + (n+p)L_f^2\eta_k\delta_k^2. \qquad (174c)
\end{aligned}
$$

Proof: We know that (69a)–(69g) and (81) still hold since Assumptions 3 and 4 hold.

We have

$$
\|\boldsymbol{g}_k^0\|^2 = \sum_{i=1}^n\|\nabla f_i(\bar{x}_k)\|^2 \le \sum_{i=1}^n 2L_f(f_i(\bar{x}_k) - f_i^*) = 2L_fn(f(\bar{x}_k) - f^*) + n\tilde{\sigma}_2^2, \qquad (175)
$$

where the inequality holds due to (38).

We have

$$
\|\boldsymbol{g}_k\|^2 = \|\boldsymbol{g}_k - \boldsymbol{g}_k^0 + \boldsymbol{g}_k^0\|^2 \le 2(\|\boldsymbol{g}_k - \boldsymbol{g}_k^0\|^2 + \|\boldsymbol{g}_k^0\|^2) \le 2(L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + 2L_fW_{4,k} + n\tilde{\sigma}_2^2), \quad (176)
$$

where the first inequality holds due to the Cauchy–Schwarz inequality; and the last inequality holds due to (72) and (175).

From (81) and (176), we have

$$
\mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2] \le 16pL_fW_{4,k} + 8pL_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + 4np\sigma_1^2 + 8np\tilde{\sigma}_2^2 + 0.5np^2L_f^2\delta_k^2. \qquad (177)
$$

From the Cauchy–Schwarz inequality, (69e), and (175), we have

$$\|\boldsymbol{g}^0_{k+1}\|^2 = \|\boldsymbol{g}^0_{k+1} - \boldsymbol{g}^0_k + \boldsymbol{g}^0_k\|^2 \le 2(\|\boldsymbol{g}^0_{k+1} - \boldsymbol{g}^0_k\|^2 + \|\boldsymbol{g}^0_k\|^2) \le 2(\eta_k^2 L_f^2 \|\boldsymbol{g}^e_k\|^2 + 2L_f W_{4,k} + n\tilde{\sigma}_2^2). \tag{178}$$

Then, similar to the way to get Lemma 8, from (69a)–(69g), (177), and (178), we get Lemma 10. ■

Now we are ready to prove Theorem 5

From $t_1 > \check{c}_3(\kappa_0, \kappa_1, \kappa_2) \ge \frac{\check{c}_0(\kappa_1,\kappa_2)}{\kappa_0}$, we have

$$\kappa_0 > \frac{\check{c}_0(\kappa_1, \kappa_2)}{t_1}.$$

Thus, all conditions needed in Lemma 10 are satisfied, so (174a)–(174c) still hold when $\theta = 1$.

Similar to the way to get (114), from $t_1 \ge \max\{(\frac{16L_f\kappa_3}{\nu\kappa_2})^{\frac{1}{3}}, (\frac{16L_f\kappa_4}{\nu\kappa_0\kappa_2})^{\frac{1}{3}}, \frac{64pL_f\kappa_2\varepsilon_{10}}{\nu\kappa_0}\}$, we have

$$\frac{1}{2} - \frac{4}{3\nu} L_f(b_{3,k} + 6pb_{4,k})\eta_k \ge \frac{1}{8}. \tag{179}$$

From (174a), (140), (179), (119), and (120), we know that (141) still holds when ε_{12} is replaced by $\breve{\varepsilon}_{12}$.

Then, similar to the way to get (14a) and (14b), we have (16a) and (16b).

G. Proof of Theorem 6

In addition to the notations defined in Appendix C, we also denote the following notations.

$$\varepsilon = \frac{1}{2} + \frac{1}{2}\max\{1 - \tilde{\varepsilon}_{17},\ \hat{\varepsilon}\},$$

$$\tilde{\varepsilon}_{17} = \frac{1}{4\kappa_6}\min\{4\varepsilon_4,\ 8\varepsilon_6,\ \eta\nu\},$$

$$c_4 = \frac{1}{\varepsilon_4}\Big(\frac{W_0}{n} + \frac{p\tilde{\varepsilon}_{11}\kappa_\delta^2\eta}{1-\hat{\varepsilon}}\Big),$$

$$c_5 = \frac{2p\tilde{\varepsilon}_{10}}{\varepsilon_4},$$

$$c_6 = 8\Big(\frac{W_0}{n} + \frac{p\tilde{\varepsilon}_{11}\kappa_\delta^2\eta}{1-\hat{\varepsilon}}\Big),$$

$$c_7 = 16p\tilde{\varepsilon}_{10},$$

$$c_8 = \frac{W_0}{n} + \frac{p\tilde{\varepsilon}_{11}\kappa_\delta^2\eta}{\varepsilon - \hat{\varepsilon}},$$

$$c_9 = \frac{2p\tilde{\varepsilon}_{10}\eta}{\tilde{\varepsilon}_{17}}.$$

All conditions needed in Lemma 9 are satisfied, so (134a) still holds.

(i) Taking expectation in $\mathcal{L}_T$, summing (134a) over $k \in [0, T]$, and using $\delta_{i,k} \in (0, \kappa_\delta \hat{\varepsilon}^{k/2}]$ yield

$$\mathbf{E}[W_{T+1}] + \varepsilon_4 \sum_{k=0}^{T} \|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{8}\eta \sum_{k=0}^{T} \|\bar{\boldsymbol{g}}^0_k\|^2 \le W_0 + 2pn(\sigma_1^2 + 3\sigma_2^2)\tilde{\varepsilon}_{10}\eta^2(T+1) + \frac{pn\tilde{\varepsilon}_{11}\kappa_\delta^2\eta}{1-\hat{\varepsilon}},$$

which gives (18a)–(18b).

(ii) If Assumption 6 also holds, then (140) holds. From (134a), (140), and (120), for any $k \in \mathbb{N}_0$, we have

$$\begin{aligned}\mathbf{E}[W_{k+1}] &\le W_k - \varepsilon_4\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - 2\varepsilon_6\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}^0_k\|^2_{\boldsymbol{K}} \\ &\quad - \frac{\eta\nu n}{4}(f(\bar{x}_k) - f^*) + 2pn(\sigma_1^2 + 3\sigma_2^2)\tilde{\varepsilon}_{10}\eta^2 + pn\tilde{\varepsilon}_{11}\eta\delta_k^2 \\ &\le W_k - \tilde{\varepsilon}_{17}W_k + 2pn(\sigma_1^2 + 3\sigma_2^2)\tilde{\varepsilon}_{10}\eta^2 + pn\tilde{\varepsilon}_{11}\eta\delta_k^2. \end{aligned} \tag{180}$$

From (145)

$$0 < \tilde{\varepsilon}_{17} \le \frac{2\varepsilon_6}{\kappa_6} \le \frac{1}{40}. \tag{181}$$

From (180), (119), (181), and $\delta_{i,k} \in (0, \kappa_\delta \hat{\varepsilon}^{\frac{k}{2}}]$, we have

$$\begin{aligned}\mathbf{E}[W_{k+1}] &\le (1 - \tilde{\varepsilon}_{17})^{k+1}W_0 + 2pn(\sigma_1^2 + 3\sigma_2^2)\tilde{\varepsilon}_{10}\eta^2 \sum_{\tau=0}^{k}(1 - \tilde{\varepsilon}_{17})^\tau \\ &\quad + pn\tilde{\varepsilon}_{11}\kappa_\delta^2\eta \sum_{\tau=0}^{k}(1 - \tilde{\varepsilon}_{17})^\tau \hat{\varepsilon}^{k-\tau}, \ \forall k \in \mathbb{N}_0. \end{aligned} \tag{182}$$

From (182), (42), and $\varepsilon > \max\{1 - \tilde{\varepsilon}_{17},\ \hat{\varepsilon}\}$, we have

$$\mathbf{E}[W_{k+1}] \le \epsilon^{k+1}c_8 + n\eta(\sigma_1^2 + 3\sigma_2^2)c_9, \ \forall k \in \mathbb{N}_0, \tag{183}$$

which gives (19).

H. Proof of Theorem 7

We denote the following notations.

$$\begin{aligned}
d_1 &= \frac{\rho_2(L)}{2\rho(L^2)},\\
d_2(\gamma) &= \min\Big\{\frac{4\epsilon_1}{9L_f^2},\ \frac{1}{48p(2\epsilon_2+L_f)}\Big\},\\
\epsilon_1 &= \frac{1}{2}\gamma\rho_2(L)-\gamma^2\rho(L^2),\\
\epsilon_2 &= \frac{1+2\gamma\rho_2(L)}{2\gamma\rho_2(L)},\\
\epsilon_3 &= 2\Big(2\epsilon_2+\frac{1}{n}L_f\Big)(\sigma_1^2+3\sigma_2^2),\\
\epsilon_4 &= \frac{1}{4}L_f^2\Big(\frac{1}{48}+\frac{4}{p}\Big),\\
\epsilon_5 &= \frac{W_{1,0}+W_{4,0}}{n}+\frac{2\theta p(\epsilon_3+\kappa_\delta^2\epsilon_4)\kappa_\eta^2}{2\theta-1},\\
\epsilon_6 &= pn\kappa_\eta^2(24L_f\epsilon_2\epsilon_5G_f^2+4\epsilon_2(\sigma_1^2+3\sigma_2^2)+\epsilon_4\kappa_\delta^2).
\end{aligned}$$

To prove Theorem 7, the following lemma is used.

Lemma 11. *Suppose Assumptions 1–5 hold. Suppose $\gamma\in(0,d_1)$ and $\eta_k\in(0,d_2(\gamma)]$. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 2, then*

$$\mathbf{E}_{\mathfrak{L}_k}[W_{1,k+1}+W_{4,k+1}]\le W_{1,k}+W_{4,k}-\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\epsilon_1\boldsymbol{K}}-\frac{1}{8}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2+pn\epsilon_3\eta_k^2+pn\epsilon_4\eta_k\delta_k^2, \tag{184a}$$

$$\mathbf{E}_{\mathfrak{L}_k}[W_{1,k+1}]\le W_{1,k}-\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\epsilon_1\boldsymbol{K}}+12p\epsilon_2\eta_k^2\|\bar{\boldsymbol{g}}_k^0\|^2+4pn\epsilon_2(\sigma_1^2+3\sigma_2^2)\eta_k^2+pn\epsilon_4\eta_k\delta_k^2, \tag{184b}$$

$$\mathbf{E}_{\mathfrak{L}_k}[W_{4,k+1}]\le W_{4,k}+\|\boldsymbol{x}_k\|^2_{2L_f^2\eta_k\boldsymbol{K}}-\frac{1}{8}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2+2pL_f(\sigma_1^2+3\sigma_2^2)\eta_k^2+(p+n)L_f^2\eta_k\delta_k^2. \tag{184c}$$

Proof: It is straightforward to see that for $\{\boldsymbol{x}_k\}$ generated by Algorithm 2, Lemma 6 and (97) still hold. Thus, (116) still holds.

We have

$$\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[W_{1,k+1}] &= \mathbf{E}_{\mathfrak{L}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_{k+1}\|^2_{\boldsymbol{K}}\Big]\\
&= \mathbf{E}_{\mathfrak{L}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_k-(\gamma\boldsymbol{L}\boldsymbol{x}_k+\eta_k\boldsymbol{g}_k^e)\|^2_{\boldsymbol{K}}\Big]\\
&= \mathbf{E}_{\mathfrak{L}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\gamma\|\boldsymbol{x}_k\|^2_{\boldsymbol{L}}+\frac{1}{2}\gamma^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{L}^2}
\end{aligned}$$

$$\begin{aligned}
&- \eta_k \boldsymbol{x}_k^\top (\boldsymbol{I}_{np} - \gamma \boldsymbol{L}) \boldsymbol{K} \boldsymbol{g}_k^e + \frac{1}{2}\eta_k^2 \|\boldsymbol{g}_k^e\|^2_{\boldsymbol{K}}\Big] \\
\leq\, & \mathbf{E}_{\mathfrak{L}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \gamma \|\boldsymbol{x}_k\|^2_{\boldsymbol{L}} + \frac{1}{2}\gamma^2 \|\boldsymbol{x}_k\|^2_{\boldsymbol{L}^2} \\
&+ \frac{1}{2}\gamma \rho_2(L) \|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{2\gamma\rho_2(L)}\eta_k^2 \|\boldsymbol{g}_k^e\|^2 \\
&+ \frac{1}{2}\gamma^2 \|\boldsymbol{x}_k\|^2_{\boldsymbol{L}^2} + \frac{1}{2}\eta_k^2 \|\boldsymbol{g}_k^e\|^2 + \frac{1}{2}\eta_k^2 \|\boldsymbol{g}_k^e\|^2\Big] \\
\leq\, & \mathbf{E}_{\mathfrak{L}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \gamma \|\boldsymbol{x}_k\|^2_{\rho_2(L)\boldsymbol{K}} + \gamma^2 \|\boldsymbol{x}_k\|^2_{\rho(L^2)\boldsymbol{K}} \\
&+ \frac{1}{2}\gamma \rho_2(L) \|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1 + 2\gamma\rho_2(L)}{2\gamma\rho_2(L)}\eta_k^2 \|\boldsymbol{g}_k^e\|^2\Big] \\
=\, & \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \|\boldsymbol{x}_k\|^2_{\epsilon_1 \boldsymbol{K}} + \epsilon_2 \eta_k^2 \mathbf{E}_{\mathfrak{L}_k}[\|\boldsymbol{g}_k^e\|^2] \\
\leq\, & \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \|\boldsymbol{x}_k\|^2_{\epsilon_1 \boldsymbol{K}} + \epsilon_2 \eta_k^2 (12p \|\bar{\boldsymbol{g}}_k^0\|^2 + 12pL_f^2 \|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} \\
&+ 4np\sigma_1^2 + 12np\sigma_2^2 + 0.5np^2 L_f^2 \delta_k^2) \\
=\, & \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \|\boldsymbol{x}_k\|^2_{\epsilon_1 \boldsymbol{K} - 12pL_f^2 \epsilon_2 \eta_k^2 \boldsymbol{K}} + \epsilon_2 \eta_k^2 (12p \|\bar{\boldsymbol{g}}_k^0\|^2 \\
&+ 4np\sigma_1^2 + 12np\sigma_2^2 + 0.5np^2 L_f^2 \delta_k^2),
\end{aligned} \tag{185}$$

where the second equality holds due to (20); the third equality holds due to (34a); the first inequality holds due to the Cauchy–Schwarz inequality and $\rho(\boldsymbol{K}) = 1$; the second inequality holds due to (34b); the second last equality holds since that $x_{i,k}$ is independent of $\mathfrak{L}_k$; and the last inequality holds due to (70a).

From (116) and (185), we have

$$\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[W_{1,k+1} + W_{4,k+1}] \leq\, & W_{1,k} + W_{4,k} - \|\boldsymbol{x}_k\|^2_{\epsilon_1 \boldsymbol{K} - (L_f^2 \eta_k + 12pL_f^2 \epsilon_2 \eta_k^2 + \frac{6p}{n} L_f^3 \eta_k^2)\boldsymbol{K}} \\
&- \frac{1}{4}\Big(1 - 48p\epsilon_2 \eta_k - \frac{24p}{n} L_f \eta_k\Big)\eta_k \|\bar{\boldsymbol{g}}_k^0\|^2 - \frac{1}{4}(1 - 2L_f \eta_k)\eta_k \|\bar{\boldsymbol{g}}_k^s\|^2 \\
&+ 2pn\Big(2\epsilon_2 + \frac{1}{n}L_f\Big)(\sigma_1^2 + 3\sigma_2^2)\eta_k^2 + \frac{1}{4}pnL_f^2\Big(2p\epsilon_2\eta_k + \frac{1}{n}pL_f\eta_k + \frac{4}{p}\Big)\eta_k \delta_k^2.
\end{aligned} \tag{186}$$

From $\gamma \in (0, d_1)$ and $\rho_2(L) \leq \rho(L)$, we have

$$0 < \epsilon_1 < \frac{1}{16}. \tag{187}$$

From $\eta_k \leq d_2(\gamma) \leq 1/(48p(2\epsilon_2 + L_f))$, we have

$$48p\epsilon_2\eta_k + \frac{24p}{n}L_f\eta_k \leq 24p(2\epsilon_2 + L_f)d_2(\gamma) \leq \frac{1}{2}, \tag{188a}$$

$$2L_f\eta_k \leq \frac{2L_f}{48p(2\epsilon_2 + L_f)} < \frac{1}{24p} < 1, \tag{188b}$$

$$\frac{1}{4}L_f^2\Big(2p\epsilon_2\eta_k + \frac{1}{n}pL_f\eta_k + \frac{4}{p}\Big) \leq \epsilon_4. \tag{188c}$$

From $\eta_k \leq d_2(\gamma) \leq 4\epsilon_1/(9L_f^2)$ and (188a), we have

$$L_f^2\eta_k + 12pL_f^2\epsilon_2\eta_k^2 + \frac{6p}{n}L_f^3\eta_k^2 \leq (1 + 6p(2\epsilon_2 + L_f)d_2(\gamma))L_f^2d_2(\gamma) \leq \frac{9}{8}L_f^2d_2(\gamma) \leq \frac{1}{2}\epsilon_1. \tag{189}$$

From (186)–(189), we have (184a).

Similarly, we get (184b) and (184c). ■

Now it is ready to prove Theorem 7.

From $\kappa_\eta \in (0, d_2(\gamma)t_1^\theta]$ and $\eta_k = \kappa_\eta/(k + t_1)^\theta$, we have $\eta_k \leq d_2(\gamma)$. Thus, all conditions needed in Lemma 11 are satisfied. So (184a) and (184b) hold.

Taking expectation in $\mathcal{L}_T$, summing (184a) over $k \in [0, T]$, and using (43) and $\eta_k = \kappa_\eta/(k + t_1)^\theta$ and $\delta_k \leq \kappa_\delta\sqrt{\eta_k}$ as stated in (21), yield

$$\begin{aligned}
&\mathbf{E}[W_{1,T+1} + W_{4,T+1}] + \sum_{k=0}^{T}\mathbf{E}\Big[\frac{1}{2}\epsilon_1\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{8}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2\Big]\\
&\leq W_{1,0} + W_{4,0} + pn(\epsilon_3 + \kappa_\delta^2\epsilon_4)\kappa_\eta^2\sum_{k=0}^{T}\frac{1}{(k + t_1)^{2\theta}} \leq n\epsilon_5.
\end{aligned} \tag{190}$$

Noting that $t_1^\theta = \mathcal{O}(\sqrt{p})$, we have

$$\kappa_\eta = \mathcal{O}(\frac{t_1^\theta}{p}) = \mathcal{O}(\frac{1}{\sqrt{p}}). \tag{191}$$

From $W_{1,0} + W_{4,0} = \mathcal{O}(n)$ and (191), we have

$$\epsilon_5 = \frac{W_{1,0} + W_{4,0}}{n} + \frac{2\theta p(\epsilon_3 + \kappa_\delta^2\epsilon_4)\kappa_\eta^2}{2\theta - 1} = \mathcal{O}(1). \tag{192}$$

From (190), (187), and $\sum_{k=0}^{T}\eta_k = \sum_{k=0}^{T}\frac{\kappa_\eta}{(k+t_1)^\theta} \geq \frac{\kappa_\eta(T+t_1)^{1-\theta}}{1-\theta}$, we have

$$\frac{\sum_{k=0}^{T}\eta_k\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2]}{\sum_{k=0}^{T}\eta_k} = \frac{\sum_{k=0}^{T}\eta_k\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2]}{n\sum_{k=0}^{T}\eta_k} \leq \frac{8(1-\theta)\epsilon_5}{\kappa_\eta(T + t_1)^{1-\theta}}. \tag{193}$$

From (191)–(193), we have (22a).

From (190) and (187), we have

$$\mathbf{E}[f(\bar{x}_{T+1})]-f^*\le\frac{1}{n}W_{4,T+1}\le\epsilon_5. \tag{194}$$

From (194) and (192), we have (22b).

From (190) and (187), we have

$$\sum_{k=0}^{T}\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}]\le\frac{2n\epsilon_5}{\epsilon_1}. \tag{195}$$

From (69g) and (194), we have

$$\|\bar{\boldsymbol{g}}^0_k\|^2\le 2nL_f\epsilon_5. \tag{196}$$

From (70a), (195), and (196), we know that $\mathbf{E}[\|\boldsymbol{g}^e_k\|^2]$ is bounded. Then, same as the proof of the first part of Theorem 1 in [57], we have (22d).

From (184b), (196), and (21), we have

$$\mathbf{E}[W_{1,k+1}]\le(1-\epsilon_1)\mathbf{E}[W_{1,k}]+\frac{\epsilon_6}{(t+t_1)^{2\theta}}. \tag{197}$$

From (197), (187), and (51), we have

$$\mathbf{E}[W_{1,k}]\le\phi_3(k,t_1,\epsilon_1,\epsilon_6,2\theta,W_{1,0}),\ \forall k\in\mathbb{N}_+, \tag{198}$$

where the function ϕ_3 is defined in (52).

Noting that $\phi_3(k,t_1,\epsilon_1,\epsilon_6,2\theta,W_{1,0})=\mathcal{O}(n/k^{2\theta})$, from (198), we have (22c).

I. *Proof of Theorem 8*

We use the notations defined in Appendix H.

From $\eta_k=\eta=\sqrt{n}/\sqrt{pT}$ and $T\ge n/(pd_2^2(\gamma))$, we have $\eta_k\le d_2(\gamma)$. Thus, all conditions needed in Lemma 11 are satisfied. So (184a) and (184c) hold.

From (184a), $\eta_k=\eta=\sqrt{n}/\sqrt{pT}$, and $\delta_{i,k}\le\kappa_\delta/(pn(k+1))^{1/4}$ as stated in (23), similar to the way to get (195) and (194), we have

$$\frac{1}{T+1}\sum_{k=0}^{T}\mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k}-\bar{x}_k\|^2]\le\frac{2}{\epsilon_1}\Big(\frac{W_{1,0}+W_{4,0}}{n(T+1)}+\frac{n\epsilon_3}{T}+\frac{2n\kappa_\delta^2\epsilon_4}{\sqrt{T(T+1)}}\Big), \tag{199}$$

which gives (24c).

From (184c) and $\eta_k = \eta$, we have

$$\mathbf{E}_{\mathfrak{L}_k}[W_{4,k+1}] \le W_{4,k} + \|\boldsymbol{x}_k\|^2_{2L_f^2\eta\boldsymbol{K}} - \frac{1}{8}\eta\|\bar{\boldsymbol{g}}_k^0\|^2 + 2pL_f(\sigma_1^2 + 3\sigma_2^2)\eta^2 + (p+n)L_f^2\eta\delta_k^2. \tag{200}$$

From (200) and $\delta_{i,k} \le \kappa_\delta/(pn(k+1))^{1/4}$ as stated in (23), similar to the way to get (193), we have

$$\begin{aligned}&\frac{1}{T+1}\sum_{k=0}^{T}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] = \frac{1}{n(T+1)}\sum_{k=0}^{T}\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2]\\ &\le 8\Big(\frac{W_{4,0}}{n(T+1)\eta} + \frac{2L_f^2}{n(T+1)}\sum_{k=0}^{T}\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}] + \frac{2pL_f(\sigma_1^2+3\sigma_2^2)\eta}{n} + \frac{2\sqrt{p}L_f^2\kappa_\delta^2}{\sqrt{n(T+1)}}\Big).\end{aligned} \tag{201}$$

Noting that $\eta = \sqrt{n}/\sqrt{pT}$ and $T \ge n^3/p$, from (199) and (201), we have

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] = 8(f(\bar{x}_0) - f^* + 2(\sigma_1^2 + 3\sigma_2^2)L_f + 2L_f^2\kappa_\delta^2)\frac{\sqrt{p}}{\sqrt{nT}} + \mathcal{O}(\frac{n}{T}),$$

which gives (24a).

Taking expectation in $\mathcal{L}_T$, summing (200) over $k \in [0, T]$, and using $\delta_{i,k} \le \kappa_\delta/(pn(k+1))^{1/4}$ yield

$$\begin{aligned}&n(\mathbf{E}[f(\bar{x}_{T+1})] - f^*) = \mathbf{E}[W_{4,T+1}]\\ &\le W_{4,0} + 2\eta L_f^2\sum_{k=0}^{T}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + (T+1)2p\eta^2L_f(\sigma_1^2+3\sigma_2^2) + 2\sqrt{pn}L_f^2\eta\sqrt{T+1}.\end{aligned} \tag{202}$$

Noting that $W_{4,0} = \mathcal{O}(n)$, $\eta = \sqrt{n}/\sqrt{pT}$, and $T \ge n^3/p$, from (199) and (202), we have (24b).

Similar to the proof of (22d), we have (24d).

J. Proof of Theorem 9

In addition to the notations defined in Appendix H, we also denote the following notations.

$$\begin{aligned}&\tilde{\epsilon}_6 = pn\kappa_\eta^2(12L_f\epsilon_2d_g + 4\epsilon_2(\sigma_1^2+3\sigma_2^2) + \epsilon_4\kappa_\delta^2),\\ &\epsilon_7 = \min\Big\{\frac{\epsilon_1t_1^\theta}{\kappa_\eta},\ \frac{\nu}{4}\Big\},\\ &b_1 = \epsilon_7\kappa_\eta,\\ &b_2 = pn(\epsilon_3 + \epsilon_4\kappa_\delta^2)\kappa_\eta^2.\end{aligned}$$

All conditions needed in Lemma 11 are satisfied, so (184a)–(184c) hold.

Denote $\check{W}_k = W_{1,k} + W_{4,k}$. From (184a) and (140), we have

$$\begin{aligned}\mathbf{E}_{\mathfrak{L}_k}[\check{W}_{k+1}] &\le \check{W}_k - \|\boldsymbol{x}_k\|^2_{\frac{1}{2}\epsilon_1\boldsymbol{K}} - \frac{\nu}{4}\eta_k W_{4,k} + pn\epsilon_3\eta_k^2 + pn\epsilon_4\eta_k\delta_k^2\\ &\le \Big(1 - \eta_k \min\Big\{\frac{\epsilon_1}{\eta_k},\ \frac{\nu}{4}\Big\}\Big)\check{W}_k + pn\epsilon_3\eta_k^2 + pn\epsilon_4\eta_k\delta_k^2\\ &\le (1-\eta_k\epsilon_7)\check{W}_k + pn\epsilon_3\eta_k^2 + pn\epsilon_4\eta_k\delta_k^2,\ \forall k \in \mathbb{N}_0. \end{aligned} \tag{203}$$

Denote $\check{z}_k = \mathbf{E}[\check{W}_k]$, $s_{1,k} = \eta_k\epsilon_7$, and $s_{2,k} = pn\epsilon_3\eta_k^2 + pn\epsilon_4\eta_k\delta_k^2$. From (203), we have

$$\check{z}_{k+1} \le (1-s_{1,k})\check{z}_k + s_{2,k},\ \forall k \in \mathbb{N}_0. \tag{204}$$

From (25), we have

$$s_{1,k} = \eta_k\epsilon_7 = \frac{b_1}{(k+t_1)^\theta}, \tag{205}$$

$$s_{2,k} = pn\epsilon_3\eta_k^2 + pn\epsilon_4\eta_k\delta_k^2 \le \frac{b_2}{(k+t_1)^{2\theta}}. \tag{206}$$

From (187), we have

$$0 < s_{1,k} \le \epsilon_1 \le \frac{1}{16}. \tag{207}$$

Then, from $\theta \in (0,1)$, (204)–(207), and (47), we have

$$\check{z}_k \le \phi_1(k, t_1, b_1, b_2, \theta, 2\theta, \check{z}_0),\ \forall k \in \mathbb{N}_+, \tag{208}$$

where the function ϕ_1 is defined in (48).

Noting that $t_1^\theta = \mathcal{O}(p)$, we have

$$\kappa_\eta = \mathcal{O}(\frac{t_1^\theta}{p}) = \mathcal{O}(1). \tag{209}$$

From (69g), (208), and (209), we get

$$\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2] = \mathcal{O}(\frac{pn}{(k+t_1)^\theta}),\ \forall k \in \mathbb{N}_+. \tag{210}$$

From (148) and (210), we know that there exists a constant $d_g > 0$, such that

$$\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2] \le nd_g,\ \forall k \in \mathbb{N}_0. \tag{211}$$

From (184b), (211), and (25), we have

$$\mathbf{E}[W_{1,k+1}] \leq (1-\epsilon_1)\mathbf{E}[W_{1,k}] + \frac{\tilde{\epsilon}_6}{(t+t_1)^{2\theta}}. \tag{212}$$

Using (51), from (187) and (212), we have

$$\mathbf{E}[W_{1,k}] \leq \phi_3(k, t_1, \epsilon_1, \tilde{\epsilon}_6, 2\theta, W_{0,k}),\ \forall k \in \mathbb{N}_+, \tag{213}$$

where the function ϕ_3 is defined in (52). From (213), (52), and (209), we have

$$\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}] \leq 2\mathbf{E}[W_{1,k}] \leq 2\phi_3(k, t_1, \epsilon_1, \tilde{\epsilon}_6, 2\theta, W_{0,k}) = \mathcal{O}(\frac{pn}{(k+t_1)^{2\theta}}), \tag{214}$$

which yields (26a).

From (184c), (140), and $\delta_k \leq \kappa_\delta \eta_k$ we have

$$\begin{aligned}\mathbf{E}[W_{4,k+1}] \leq{}& \mathbf{E}[W_{4,k}] - \frac{\nu}{4}\eta_k\mathbf{E}[W_{4,k}] + \|\boldsymbol{x}_k\|^2_{2L_f^2\eta_k\boldsymbol{K}} \\ &+ 2pL_f(\sigma_1^2 + 3\sigma_2^2)\eta_k^2 + (p+n)L_f^2\kappa_\delta^2\eta_k^3. \end{aligned} \tag{215}$$

Similar to the way to prove (47), from (214) and (215), we have (26b).

K. Proof of Theorem 10

In addition to the notations defined in Appendices H and J, we also denote

$$\hat{d}_2(\gamma) = \max\Big\{\frac{1}{\epsilon_1},\ \frac{\kappa_\eta}{d_2(\gamma)}\Big\}.$$

From $t_1 > \hat{d}_2(\gamma) \geq \frac{\kappa_\eta}{d_2(\gamma)}$, we have

$$\eta_k = \frac{\kappa_\eta}{k+t_1} \leq \frac{\kappa_\eta}{t_1} < d_2(\gamma).$$

Thus, all conditions needed in Lemma 11 are satisfied, so (204)–(207) still hold when $\theta = 1$.

From $t_1 > \hat{d}_2(\gamma) \geq \frac{1}{\epsilon_1}$, we have

$$\epsilon_1 t_1 > 1. \tag{216}$$

From $\kappa_\eta > 4/\nu$, we have

$$\frac{\nu\kappa_\eta}{4} > 1. \tag{217}$$

Hence, from (216) and (217), we have

$$b_1 = \epsilon_6 \kappa_\eta > 1. \tag{218}$$

Then from $\theta = 1$, (204)–(207), (218), and (49), we have

$$\check{z}_k \le \phi_2(k, t_1, b_1, b_2, 2, \check{z}_0),\ \forall k \in \mathbb{N}_+, \tag{219}$$

where the function ϕ_2 is defined in (50).

From $\kappa_\eta > 4/\nu$, we know $\kappa_\eta = \mathcal{O}(1)$, thus $\phi_2(k, t_1, b_1, b_2, 2, \check{z}_0) = \mathcal{O}(pn/k)$. Hence, from (69g) and (219), we get

$$\mathbf{E}[\|\bar{\boldsymbol{g}}_k^0\|^2] = \mathcal{O}(\frac{pn}{k+t_1}),\ \forall k \in \mathbb{N}_+. \tag{220}$$

Then, similar to the way to get (26a) and (26b), we get (28a) and (28b).

L. Proof of Theorem 11

In addition to the notations defined in Appendices H, J, and K, we also denote

$$\begin{aligned}
\tilde{d}_2(\gamma) &= \min\Big\{\frac{\epsilon_1}{4L_f^2},\ \frac{1}{4p(2\epsilon_2 + L_f)}\Big\},\\
\check{d}_2(\gamma) &= \max\Big\{\frac{1}{\epsilon_1},\ \frac{\kappa_\eta}{\tilde{d}_2(\gamma)},\ \frac{\kappa_\eta}{8\nu\epsilon_8}\Big\},\\
\check{\epsilon}_3 &= 2\Big(2\epsilon_2 + \frac{1}{n}L_f\Big)(\sigma_1^2 + 2\tilde{\sigma}_2^2),\\
\check{\epsilon}_4 &= \frac{1}{4}L_f^2\Big(\frac{1}{8} + \frac{4}{p}\Big),\\
\epsilon_8 &= 8p(2\epsilon_2 + L_f)L_f
\end{aligned}$$

To prove Theorem 11, the following lemma is used.

Lemma 12. *Suppose Assumptions 1–4 hold and each $f_i^* > -\infty$. Suppose $\gamma \in (0, d_1)$ and $\eta_k \in (0, \tilde{d}_2(\gamma)]$. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 2, then*

$$\begin{aligned}
\mathbf{E}_{\mathfrak{L}_k}[W_{1,k+1} + W_{4,k+1}] \le{}& W_{1,k} + W_{4,k} - \|\boldsymbol{x}_k\|^2_{\frac{1}{2}\epsilon_1 \boldsymbol{K}} - \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2\\
&+ \epsilon_8\eta_k^2 W_{4,k} + pn\check{\epsilon}_3\eta_k^2 + pn\check{\epsilon}_4\eta_k\delta_k^2,
\end{aligned} \tag{221a}$$

$$\mathbf{E}_{\mathfrak{L}_k}[W_{1,k+1}] \le W_{1,k} - \|\boldsymbol{x}_k\|^2_{\frac{1}{2}\epsilon_1 \boldsymbol{K}} + 16p\epsilon_2 L_f \eta_k^2 W_{4,k}$$

$$+ 4pn\epsilon_2(\sigma_1^2 + 2\tilde{\sigma}_2^2)\eta_k^2 + pn\check{\epsilon}_4\eta_k\delta_k^2, \tag{221b}$$

$$\mathbf{E}_{\mathfrak{L}_k}[W_{4,k+1}] \le W_{4,k} + \|\boldsymbol{x}_k\|^2_{2L_f^2\eta_k\boldsymbol{K}} - \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2 + \frac{8p}{n}L_f^2W_{4,k}$$
$$+ 2pL_f(\sigma_1^2 + 2\tilde{\sigma}_2^2)\eta_k^2 + (p+n)L_f^2\eta_k\delta_k^2. \tag{221c}$$

Proof: We know that (69a)–(69g) and (177) still hold since Assumptions 3 and 4 hold, and each $f_i^* > -\infty$. Then, similar to the way to get Lemma 11, we get Lemma 12. ■

Now we are ready to prove Theorem 11.

From $t_1 > \check{d}_2(\gamma) \ge \max\{\kappa_\eta/\tilde{d}_2(\gamma),\ \kappa_\eta/(4\nu\epsilon_8)\}$, we have

$$\eta_k = \frac{\kappa_\eta}{k+t_1} \le \frac{\kappa_\eta}{t_1} < \min\Big\{\tilde{d}_2(\gamma),\ \frac{1}{4\nu\epsilon_8}\Big\}. \tag{222}$$

Thus, all conditions needed in Lemma 12 are satisfied, so (221a)–(221c) hold

From (221a), (140), and (222), we know that (203) still holds when ϵ_3 and ϵ_4 are replaced by $\check{\epsilon}_3$ and $\check{\epsilon}_4$, respectively.

Then, similar to the way to get (28a) and (28b), we have (30a) and (30b).

M. Proof of Theorem 12

In addition to the notations defined in Appendix H, we also denote the following notations.

$$\begin{aligned}
&\epsilon = 0.5 + 0.5\max\{1 - \tilde{\epsilon}_7,\ \hat{\epsilon}\},\\
&\tilde{\epsilon}_7 = \min\Big\{\epsilon_1,\ \frac{1}{4}\nu\eta\Big\},\\
&d_3 = \frac{2}{\epsilon_1}\Big(\frac{W_{1,0} + W_{4,0}}{n} + \frac{p\epsilon_4\kappa_\delta^2\eta}{1-\hat{\epsilon}}\Big),\\
&d_4 = \frac{4p}{\epsilon_1}\Big(2\epsilon_2 + \frac{1}{n}L_f\Big),\\
&d_5 = 8\Big(\frac{W_{1,0} + W_{4,0}}{n} + \frac{p\epsilon_4\kappa_\delta^2\eta}{1-\hat{\epsilon}}\Big),\\
&d_6 = 16p\Big(2\epsilon_2 + \frac{1}{n}L_f\Big),\\
&d_7 = \frac{W_{1,0} + W_{4,0}}{n} + \frac{p\epsilon_4\kappa_\delta^2\eta}{\epsilon-\hat{\epsilon}},\\
&d_8 = \frac{2p\eta}{\tilde{\epsilon}_7}\Big(2\epsilon_2 + \frac{1}{n}L_f\Big).
\end{aligned}$$

All conditions needed in Lemma 11 are satisfied, so (184a) still holds.

(i) Taking expectation in $\mathcal{L}_T$, summing (184a) over $k \in [0,T]$, and using $\eta_k = \eta$ and $\delta_{i,k} \in (0, \kappa_\delta \hat{\epsilon}^{k/2}]$ yield

$$\begin{aligned}&\mathbf{E}[W_{1,T+1} + W_{4,T+1}] + \frac{1}{2}\epsilon_1 \sum_{k=0}^{T} \|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{8}\eta \sum_{k=0}^{T} \|\bar{\boldsymbol{g}}_k^0\|^2\\ &\le W_{1,0} + W_{4,0} + pn\epsilon_3\eta^2(T+1) + \frac{pn\epsilon_4\kappa_\delta^2\eta}{1-\hat{\epsilon}},\end{aligned}$$

which gives (32a)–(32b).

(ii) If Assumption 6 also holds, then (140) holds. Thus, (203) also holds when $\eta_k = \eta$. From (203) and $\eta_k = \eta$, for all $k \in \mathbb{N}_0$, we have

$$\mathbf{E}_{\mathfrak{L}_k}[\check{W}_{k+1}] \le (1-\tilde{\epsilon}_7)\check{W}_k + pn\epsilon_3\eta^2 + pn\epsilon_4\eta\delta_k^2. \tag{223}$$

From (187)

$$0 < \tilde{\epsilon}_7 \le \epsilon_1 < \frac{1}{16}. \tag{224}$$

From (223), (224), and $\delta_{i,k} \in (0, \kappa_\delta \hat{\epsilon}^{\frac{k}{2}}]$, we have

$$\begin{aligned}&\mathbf{E}[\check{W}_{k+1}]\\ &\le (1-\tilde{\epsilon}_7)^{k+1}\check{W}_0 + pn\epsilon_3\eta^2 \sum_{\tau=0}^{k}(1-\tilde{\epsilon}_7)^{\tau} + pn\epsilon_4\kappa_\delta^2\eta \sum_{\tau=0}^{k}(1-\tilde{\epsilon}_7)^{\tau}\hat{\epsilon}^{k-\tau},\ \forall k \in \mathbb{N}_0.\end{aligned} \tag{225}$$

From (225), (42), and $\epsilon > \max\{1-\tilde{\epsilon}_7,\ \hat{\epsilon}\}$, we have

$$\mathbf{E}[\check{W}_{k+1}] \le \epsilon^{k+1}d_7 + n(\sigma_1^2 + 3\sigma_2^2)d_8,\ \forall k \in \mathbb{N}_0, \tag{226}$$

which gives (33).