A Primal-Dual SGD Algorithm for Distributed Nonconvex Optimization

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

***Abstract*—The distributed nonconvex optimization problem of minimizing a global cost function formed by a sum of n local cost functions by using local information exchange is considered. This problem is an important component of many machine learning techniques with data parallelism, such as deep learning and federated learning. We propose a distributed primal-dual stochastic gradient descent (SGD) algorithm, suitable for arbitrarily connected communication networks and any smooth (possibly nonconvex) cost functions. We show that the proposed algorithm achieves the linear speedup convergence rate $\mathcal{O}(1/\sqrt{nT})$ for general nonconvex cost functions and the linear speedup convergence rate $\mathcal{O}(1/(nT))$ when the global cost function satisfies the Polyak-Łojasiewicz (P-L) condition, where T is the total number of iterations. We also show that the output of the proposed algorithm with fixed parameters linearly converges to a neighborhood of a global optimum. We demonstrate through numerical experiments the efficiency of our algorithm in comparison with the baseline centralized SGD and recently proposed distributed SGD algorithms.**

***Index Terms*—Distributed nonconvex optimization, linear speedup, Polyak-Łojasiewicz condition, primal-dual algorithm, stochastic gradient descent**

I. Introduction

Consider a network of n agents, each of which has a local smooth (possibly nonconvex) cost function $f_i : \mathbb{R}^p \to \mathbb{R}$. All agents collaboratively solve the following optimization problem

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

Each agent i only has information about its local cost function f_i and can communicate with its neighbors through the underlying communication network. The communication network is modeled by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, n\}$ is the agent set, $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$ is the edge set, and $(i, j) \in \mathcal{E}$ if agents i and j can communicate with each other. The set $\mathcal{N}_i = \{j \in [n] : (i, j) \in \mathcal{E}\}$ is the neighboring set of agent i. The optimization problem (1) incorporates many popular machine learning approaches with data parallelism, such as deep learning [1] and federated learning [2]. A star graph is a special undirected graph, in which there is one and only one agent (hub agent) that connects to all of the other agents (leaf agents) and each leaf agent only connects to the hub agent. Such a graph corresponds to the master/worker architecture adopted by some parallel learning algorithms.

In this paper, we consider the case where each agent is able to collect stochastic gradients of its local cost function and propose a distributed stochastic gradient descent (SGD) algorithm to solve (1). In general, SGD algorithms are suitable for scenarios where explicit expressions of the gradients are unavailable or difficult to obtain. For example, in some big data applications, such as empirical risk minimization, the actual gradient is to be calculated from the entire data set, which results in a heavy computational burden. A stochastic gradient can be calculated from a randomly selected subset of the data and is often an efficient way to replace the actual gradient. Other examples when SGD algorithms are suitable include scenarios where data are arriving sequentially such as in online learning [3].

A. Literature Review

When the communication network is a star graph, various parallel SGD algorithms have been proposed to solve (1). A potential performance bottleneck of such algorithms lies on the communication burden of the master. To overcome this issue, a promising strand of research is combining parallel SGD algorithms with communication reduction approaches, e.g., asynchronous parallel SGD algorithms [4]–[8], gradient compression based parallel SGD algorithms [5], [9]–[12], periodic averaging based parallel SGD algorithms [10], [11], [13]–[17], and parallel SGD algorithm with dynamic batch sizes [18]. Convergence properties of these algorithms have been analyzed in detail. In particular, in [10], [14], [16], [18], an $\mathcal{O}(1/\sqrt{nT})$ convergence rate has been established for general nonconvex cost functions, where T is the total number of iterations. This rate is n times faster than the well known $\mathcal{O}(1/\sqrt{T})$ convergence rate established by SGD over a single agent, and thus a linear speedup in the number of agents is achieved. In [17], [18], the convergence rate has been improved to $\mathcal{O}(1/(nT))$ when the global cost function satisfies the P-L condition, which also achieves a linear speedup. In addition to the star architecture restriction, aforementioned parallel SGD algorithms require certain restrictions on the cost functions, such as bounded $\|\nabla f_i\|$ or $\|\nabla f_i - \nabla f\|$.

Distributed algorithms executed over arbitrarily connected communication networks have been suggested to overcome

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, 61991404, 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 and T. Chai are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, 110819, Shenyang, China. {yangtao,tychai}@mail.neu.edu.cn.

communication bottlenecks for parallel SGD algorithms. Various distributed SGD algorithms have been proposed to solve (1), e.g., synchronous distributed SGD algorithms [16], [19]–[21], asynchronous distributed SGD algorithms [22], [23], compression based distributed SGD algorithms [24]–[27], and periodic averaging based distributed SGD algorithm [28]. Convergence properties of these algorithms have been analyzed and the linear speedup convergence rate $\mathcal{O}(1/\sqrt{nT})$ has been established for general nonconvex cost functions [16], [20], [23], [24], [26]–[28]. However, similar to aforementioned parallel SGD algorithms, these distributed algorithms require restrictive assumptions on the cost functions. In order to remove these restrictions, the authors of [29] proposed a variant of the distributed SGD algorithm proposed in [20], named D^2, in which each agent stores the stochastic gradient and its local model in last iteration and linearly combines them with the current stochastic gradient and local model. For this algorithm the authors established linear speedup convergence rate $\mathcal{O}(1/\sqrt{nT})$, but they required that the eigenvalues of the mixing matrix associated with the communication network are strictly greater than $-1/3$. The authors of [30], [31] proposed distributed stochastic gradient tracking algorithms suitable for arbitrarily connected communication networks. However, these algorithms only achieve $\mathcal{O}(1/\sqrt{T})$ convergence rate, which is not a speedup. Moreover, gradient tracking algorithms have the common potential drawback that in order to track the global gradient, at each iteration each agent needs to communicate one additional p-dimensional variable with its neighbors. This results in heavy communication burden when p is large. Note that all aforementioned distributed SGD algorithms converge to stationary points, which may be local or global optima, or saddle points. None of existing studies on distributed SGD algorithms consider finding the global optimum when the global cost function satisfies some additional property, such as the P-L condition studied for the parallel algorithms in [17], [18].

B. Main Contributions

The contributions of this paper are summarized as follows.

(i) We propose a distributed primal-dual SGD algorithm to solve the optimization problem (1). In the proposed algorithm, each agent maintains the primal and dual variable sequences and only communicates the primal variable with its neighbors. This algorithm is suitable for arbitrarily connected communication networks and any smooth (possibly nonconvex) cost functions.

(ii) We show that our algorithm finds a stationary point with the linear speedup convergence rate $\mathcal{O}(1/\sqrt{nT})$ for general nonconvex cost functions. Compared with [10], [14], [16], [18], [20], [23], [24], [26]–[29], we achieve the same convergence rate but under weaker assumptions related to network architectures and/or cost functions, and compared with [30], [31], we not only establish linear speedup but also just use half communication in each iteration.

(iii) We show that our algorithm finds a global optimum with the linear speedup convergence rate $\mathcal{O}(1/(nT))$ when the global cost function satisfies the P-L condition. Compared with [17], [18], [27], [32]–[34] , we achieve the same convergence rate but under weaker assumptions related to network architectures and/or cost functions, and compared with [11], [19], [35]–[39], we not only establish linear speedup but also relax the strong convexity by the P-L condition.

(iv) We show that the output of our algorithm with fixed parameters linearly converges to a neighborhood of a global optimum when the global cost function satisfies the P-L condition. Compared with [19], [39]–[41], which used the strong convexity assumption, we achieve similar convergence result under weaker assumptions on the cost function.

The comparison of this paper to other related studies in the literature is summarized in Table I.

C. Outline

The rest of this paper is organized as follows. Section II presents the new distributed primal-dual SGD algorithm. Section III analyzes its convergence rate. Numerical experiments are given in Section IV. Finally, concluding remarks are offered in Section V. To improve the readability, all the proofs are given in the appendix.

Notations: $\mathbb{N}_0$ and $\mathbb{N}_+$ denote the set of nonnegative and positive integers, respectively. $[n]$ denotes the set $\{1,\dots,n\}$ for any $n\in\mathbb{N}_+$. $\mathbf{1}_n$ ($\mathbf{0}_n$) denotes the column one (zero) vector of dimension n. $\mathrm{col}(z_1,\dots,z_k)$ is the concatenated column vector of vectors $z_i\in\mathbb{R}^{p_i},\ i\in[k]$. $\|\cdot\|$ represents the Euclidean norm for vectors or the induced 2-norm for matrices. Given a differentiable function f, ∇f denotes the gradient of f.

II. Distributed Primal-Dual SGD Algorithm

In this section, we propose a new distributed SGD algorithm based on the primal-dual method.

Denote $\boldsymbol{x}=\mathrm{col}(x_1,\dots,x_n)$, $\tilde{f}(\boldsymbol{x})=\sum_{i=1}^n f_i(x_i)$, and $\boldsymbol{L}=L\otimes\mathbf{I}_p$, where $L=(L_{ij})$ is the weighted Laplacian matrix associated with the undirected communication graph $\mathcal{G}$. Recall that the Laplacian matrix L is positive semi-definite and $\mathrm{null}(L)=\{\mathbf{1}_n\}$ when $\mathcal{G}$ is connected [42]. The optimization problem (1) is equivalent to the following constrained optimization problem:

$$\begin{aligned}\min_{\boldsymbol{x}\in\mathbb{R}^{np}}\quad & \tilde{f}(\boldsymbol{x})\\ \text{s.t.}\quad & \boldsymbol{L}^{1/2}\boldsymbol{x}=\mathbf{0}_{np}.\end{aligned}\tag{2}$$

Let $\boldsymbol{u}\in\mathbb{R}^{np}$ denote the dual variable. Then the augmented Lagrangian function associated with (2) is

$$\mathcal{A}(\boldsymbol{x},\boldsymbol{u})=\tilde{f}(\boldsymbol{x})+\frac{\alpha}{2}\boldsymbol{x}^\top\boldsymbol{L}\boldsymbol{x}+\beta\boldsymbol{u}^\top\boldsymbol{L}^{1/2}\boldsymbol{x},\tag{3}$$

where $\alpha>0$ and $\beta>0$ are parameters to be designed later.

Based on the primal-dual gradient method, a distributed SGD algorithm to solve (2) is

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\eta_k(\alpha_k\boldsymbol{L}\boldsymbol{x}_k+\beta_k\boldsymbol{L}^{1/2}\boldsymbol{u}_k+\boldsymbol{g}^u_k),\tag{4a}$$

$$\boldsymbol{u}_{k+1}=\boldsymbol{u}_k+\eta_k\beta_k\boldsymbol{L}^{1/2}\boldsymbol{x}_k,\ \forall\boldsymbol{x}_0,\ \boldsymbol{u}_0\in\mathbb{R}^{np},\tag{4b}$$

where $\eta_k>0$ is the stepsize at iteration k, $\alpha_k>0$ and $\beta_k>0$ are the values of the parameters α and β at iteration k, respectively, and $\boldsymbol{g}^u_k=\mathrm{col}(g^u_{1,k},\dots,g^u_{n,k})$ with $g^u_{i,k}=g_i(x_{i,k},\xi_{i,k})$

TABLE I: Comparison of this paper to some related works.

Reference	Problem type	Extra assumption	Communication network	Communicated variable	Communication rounds	Convergence rate
[10]	Nonconvex	Bounded $\|\nabla f_i - \nabla f\|$	Star graph	One quantized variable	$\mathcal{O}(n^{5/4}T^{3/4})$	$\mathcal{O}(1/\sqrt{nT})$
[11]	Nonconvex	Identical ∇f_i	Star graph	One quantized variable	$\mathcal{O}(T)$	$\mathcal{O}(1/\sqrt{T})$
	Strongly convex					$\mathcal{O}(1/T)$
[14]	Nonconvex	Bounded $\|\nabla f_i\|$	Star graph	One full-information variable	$\mathcal{O}(n^{3/4}T^{3/4})$	$\mathcal{O}(1/\sqrt{nT})$
[16]	Nonconvex	Bounded $\|\nabla f_i - \nabla f\|$	Star graph	Two full-information variables	$\mathcal{O}(n^{3/4}T^{3/4})$	$\mathcal{O}(1/\sqrt{nT})$
			Connected graph		$\mathcal{O}(T)$	
[17]	P-L condition	Identical ∇f_i	Star graph	One full-information variable	$\mathcal{O}((nT)^{1/3})$	$\mathcal{O}(1/(nT))$
[18]	Nonconvex	Identical ∇f_i, exponentially increasing batch size	Star graph	One full-information variable	$\mathcal{O}(\sqrt{nT}\log(\frac{T}{n}))$	$\mathcal{O}(1/\sqrt{nT})$
	P-L condition				$\mathcal{O}(\log(T))$	$\mathcal{O}(1/(nT))$
[19]	Nonconvex	Bounded $\|\nabla f_i\|$	Connected graph	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/T^{\theta}),\ \forall\theta\in(0,0.5)$
	Strongly convex					$\mathcal{O}(1/T)$ (time-varying stepsize); linearly to a neighbor of the global optimum (fixed stepsize)
[20]	Nonconvex	Bounded $\|\nabla f_i - \nabla f\|$	Connected graph	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/\sqrt{nT})$
[23]	Nonconvex	Bounded $\|\nabla f_i - \nabla f\|$	Uniformly jointly strongly connected digraph	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/\sqrt{nT})$
[24]	Nonconvex	Bounded $\|\nabla f_i - \nabla f\|$	Connected graph	One compressed variable	$\mathcal{O}(T)$	$\mathcal{O}(1/\sqrt{nT})$
[26]	Nonconvex	Bounded $\|\nabla f_i\|$	Strongly connected digraph	One quantized variable	$\mathcal{O}(T)$	$\mathcal{O}(1/\sqrt{nT})$
[27]	Nonconvex	Bounded $\|\nabla f_i\|$	Connected graph	One compressed variable	Event-triggered	$\mathcal{O}(1/\sqrt{nT})$
	Strongly convex					$\mathcal{O}(1/(nT))$
[28]	Nonconvex	Identical ∇f_i	Connected graph	One full-information variable	$\mathcal{O}(n^{3/2}\sqrt{T})$	$\mathcal{O}(1/\sqrt{nT})$
[29]	Nonconvex	The eigenvalues of the mixing matrix are strictly greater than $-1/3$	Connected graph	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/\sqrt{nT})$
[30], [31]	Nonconvex	No	Connected graph	Two full-information variables	$\mathcal{O}(T)$	$\mathcal{O}(1/\sqrt{T})$
[32]	Strongly convex	Bounded $\|\nabla f_i\|$	Star graph	One full-information variable	$\mathcal{O}(\sqrt{T/n})$	$\mathcal{O}(1/(nT))$
[33]	Strongly convex	Bounded $\|\nabla f_i\|$	Connected graph	One compressed variable	$\mathcal{O}(T)$	$\mathcal{O}(1/(nT))$
[34]	Strongly convex	No	Connected graph	Two full-information variables	$\mathcal{O}(T)$	$\mathcal{O}(1/(nT))$
[35]	Strongly convex	Identical ∇f_i	Connected graph	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/T)$
[36]	Strongly convex	No	Connected graph	One full-information variable	$\mathcal{O}(\sqrt{T})$	$\mathcal{O}(1/T)$
[37]	Strongly convex	Bounded $\|\nabla f_i\|$	Uniformly jointly strongly connected digraph	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/T)$
[38]	Strongly convex	No	Connected graph in expectation	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/T)$
[39]	Strongly convex	No	Connected graph	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/T)$ (time-varying stepsize); linearly to a neighbor of the global optimum (fixed stepsize)
[40]	Strongly convex	No	Connected graph	Two full-information variables	$\mathcal{O}(T)$	Linearly to a neighbor of the global optimum (fixed stepsize)
[41]	Strongly convex	No	Strongly connected digraph	Two full-information variables	$\mathcal{O}(T)$	Linearly to a neighbor of the global optimum (fixed stepsize)
This paper	Nonconvex	No	Connected graph	One full-information variable	$\mathcal{O}(T)$	$\mathcal{O}(1/\sqrt{nT})$
	P-L condition					$\mathcal{O}(1/(T^{\theta})),\ \forall\theta\in(0,1)$ (time-varying stepsize); linearly to a neighbor of the global optimum (fixed stepsize)
		Bounded f_i^*				$\mathcal{O}(1/(nT))$

Algorithm 1 Distributed Primal-Dual SGD Algorithm

```
Input: parameters {α_k}, {β_k}, {η_k} ⊆ (0, +∞).
Initialize: x_{i,0} ∈ R^p and v_{i,0} = 0_p, ∀i ∈ [n].
for k = 0, 1, ... do
    for i = 1, ..., n in parallel do
        Broadcast x_{i,k} to N_i and receive x_{j,k} from j ∈ N_i;
        Sample stochastic gradient g_i(x_{i,k}, ξ_{i,k});
        Update x_{i,k+1} by (6a);
        Update v_{i,k+1} by (6b).
    end for
end for
Output: {x_k}.
```

being the stochastic gradient of f_i at $x_{i,k}$ and $\xi_{i,k}$ being a random variable. Denote $\boldsymbol{v}_k = \mathrm{col}(v_{1,k}, \dots, v_{n,k}) = \boldsymbol{L}^{1/2}\boldsymbol{u}_k$. Then the recursion (4) can be rewritten as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k(\alpha_k \boldsymbol{L}\boldsymbol{x}_k + \beta_k \boldsymbol{v}_k + \boldsymbol{g}_k^u), \tag{5a}$$

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \eta_k\beta_k\boldsymbol{L}\boldsymbol{x}_k,\ \forall \boldsymbol{x}_0 \in \mathbb{R}^{np},\ \sum_{j=1}^{n} v_{j,0} = \mathbf{0}_p. \tag{5b}$$

The initialization condition $\sum_{j=1}^{n} v_{j,0} = \mathbf{0}_p$ is derived from $\boldsymbol{v}_0 = \boldsymbol{L}^{1/2}\boldsymbol{u}_0$, and it is easy to be satisfied, for example, $v_{i,0} = \mathbf{0}_p,\ \forall i \in [n]$, or $v_{i,0} = \sum_{j=1}^{n} L_{ij}x_{j,0},\ \forall i \in [n]$. Note that (5) can be written agent-wise as

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

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

This corresponds to our proposed distributed primal-dual SGD algorithm, which is presented in pseudo-code as Algorithm 1.

It should be pointed out that $\{\alpha_k\}$, $\{\beta_k\}$, $\{\eta_k\}$, $\boldsymbol{x}_0$, and $\boldsymbol{v}_0$ used in Algorithm 1 are deterministic, while $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{v}_k\}$ are random variables generated by Algorithm 1. Let $\mathfrak{F}_k$ denote the σ-algebra generated by the random variables $\xi_{1,k}, \dots, \xi_{n,k}$ and let $\mathcal{F}_k = \bigcup_{s=1}^{k} \mathfrak{F}_s$. It is straightforward to see that $\boldsymbol{x}_k$ and $\boldsymbol{v}_k$ depend on $\mathcal{F}_{k-1}$ and are independent of $\mathfrak{F}_s$ for all $s \ge k$.

III. Convergence Rate Analysis

In this section, we analyze the convergence rate of Algorithm 1. The following assumptions are made.

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

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

Assumption 3. *Each local cost function f_i is smooth with constant $L_f > 0$, i.e.,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L_f\|x - y\|,\ \forall x, y \in \mathbb{R}^p. \tag{7}$$

Assumption 4. *The random variables $\{\xi_{i,k},\ i \in [n],\ k \in \mathbb{N}_0\}$ are independent of each other.*

Assumption 5. *The stochastic estimate $g_i(x, \xi_{i,k})$ is unbiased, i.e., for all $i \in [n]$, $k \in \mathbb{N}_0$, and $x \in \mathbb{R}^p$,*

$$\mathbf{E}_{\xi_{i,k}}[g_i(x, \xi_{i,k})] = \nabla f_i(x). \tag{8}$$

Assumption 6. *The stochastic estimate $g_i(x, \xi_{i,k})$ has bounded variance, i.e., there exists a constant σ such that for all $i \in [n]$, $k \in \mathbb{N}_0$, and $x \in \mathbb{R}^p$,*

$$\mathbf{E}_{\xi_{i,k}}[\|g_i(x, \xi_{i,k}) - \nabla f_i(x)\|^2] \le \sigma^2. \tag{9}$$

Remark 1. *The bounded variance assumption (Assumption 6) is weaker than the bounded second moment (or bounded gradient) assumption made in [4]–[6], [8], [12], [14], [19], [21], [26], [27], [32], [33], [37], [43]. Moreover, note that we make no assumption on the boundedness of the deviation between the gradients of local cost functions. In other words, we do not assume that $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f(x)\|^2$ is uniformly bounded, which is commonly done in studies of deep learning, e.g., [10], [14], [16], [20], [22]–[24]. Also, we do not assume that the mean of each local stochastic gradient is the gradient of the global cost function, i.e., $\mathbf{E}_{\xi}[g_i(x, \xi)] = \nabla f(x),\ \forall x \in \mathbb{R}^p,\ \forall i \in [n]$, which is commonly assumed in studies of empirical risk minimization and stochastic optimization, e.g., [7], [9], [11], [13], [15], [17], [18], [25], [28], [35].*

A. Find Stationary Points

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

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

$$\alpha_k = \kappa_1\beta_k,\ \beta_k = \beta,\ \eta_k = \frac{\kappa_2}{\beta_k},\ \forall k \in \mathbb{N}_0, \tag{10}$$

where $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, and $\beta \ge c_0(\kappa_1, \kappa_2)$ with $c_0(\kappa_1, \kappa_2)$, $c_2(\kappa_1) > 0$ defined in Appendix B. Then, for any $T \in \mathbb{N}_+$,

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2\Big] = \mathcal{O}(\frac{1}{T}) + \mathcal{O}(\frac{1}{\beta^2}), \tag{11}$$

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2] = \mathcal{O}(\frac{\beta}{T}) + \mathcal{O}(\frac{1}{n\beta}) + \mathcal{O}(\frac{1}{\beta^2}), \tag{12}$$

$$\mathbf{E}[f(\bar{x}_T)] - f^* = \mathcal{O}(1) + \mathcal{O}(\frac{T}{n\beta^2}) + \mathcal{O}(\frac{T}{\beta^3}), \tag{13}$$

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

Proof: The explicit expressions of the right-hand sides of (11)–(13) and the proof are given in Appendix B. ∎

Corollary 1 (Linear Speedup). *Under the same assumptions as in Theorem 1, let $\beta = \sqrt{T}/\sqrt{n}$. Then, for any $T \ge n(c_0(\kappa_1, \kappa_2))^2$,*

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2\Big] = \mathcal{O}(\frac{n}{T}), \tag{14}$$

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2]=\mathcal{O}(\frac{1}{\sqrt{nT}})+\mathcal{O}(\frac{n}{T}), \quad (15)$$

$$\mathbf{E}[f(\bar{x}_T)]-f^*=\mathcal{O}(1)+\mathcal{O}(\sqrt{\frac{n^3}{T}}). \quad (16)$$

Remark 2. *It should be noted that the same linear speedup result as in* (15) *was also established by the SGD algorithms proposed in [10], [14], [16], [18], [20], [23], [24], [26]–[29]. However, in [10], [16], [20], [23], [24], the additional assumption that the deviation between the gradients of local cost functions is bounded was made; in [14], [26], [27], it was required that each local stochastic gradient has bounded second moment; in [18], [28], it was assumed that the mean of each local stochastic gradient is the gradient of the global cost function; and in [29], it was required that the eigenvalues of the mixing matrix are strictly greater than* $-1/3$. *Moreover, the algorithms proposed in [10], [18] are restricted to a star graph; the distributed momentum SGD algorithm proposed in [16] requires each agent* i *to communicate one additional* p*-dimensional variable besides the communication of* $x_{i,k}$ *with its neighbors at each iteration; and the algorithm proposed in [18] requires an exponentially increasing batch size, which is not favorable in practice. Under the same conditions, the well known* $\mathcal{O}(1/\sqrt{T})$ *convergence rate, which is not a speedup, was achieved by the distributed stochastic gradient tracking algorithm proposed in [30], [31]. Moreover, similar to the distributed momentum SGD algorithm proposed in [16], one potential drawback of the distributed stochastic gradient tracking algorithms is that at each iteration each agent needs to communicate one additional variable. The potential drawbacks of the results stated in Corollary 1 are that (i) we do not consider communication efficiency, which was considered in [10], [14], [18], [24], [26]–[28]; and (ii) we use time-invariant undirected graphs rather than directed graphs as considered in [23], [26]. We leave the extension to the time-varying directed graphs with communication efficiency as future research directions.*

B. Find Global Optimum

Let us next consider cases when Algorithm 1 finds global optima. The following assumption is crucial.

Assumption 7. *The global cost function* $f(x)$ *satisfies the Polyak-Łojasiewicz (P-L) condition with constant* $\nu>0$, *i.e.,*

$$\frac{1}{2}\|\nabla f(x)\|^2\geq\nu(f(x)-f^*),\ \forall x\in\mathbb{R}^p. \quad (17)$$

It is straightforward to see that every (essentially or weakly) strongly convex function satisfies the P-L condition. The P-L condition implies that every stationary point is a global minimizer, i.e., $\mathbb{X}^*=\{x\in\mathbb{R}^p:\nabla f(x)=\mathbf{0}_p\}$. But unlike (essentially or weakly) strong convexity, the P-L condition alone does not imply convexity of f. Moreover, it does not imply that $\mathbb{X}^*$ is a singleton either [44], [45].

Many practical applications, such as least squares and logistic regression, do not always have strongly convex cost functions. The cost function in least squares problems has the form

$$f(x)=\frac{1}{2}\|Ax-b\|^2,$$

where $A\in\mathbb{R}^{m\times p}$ and $b\in\mathbb{R}^m$. Note that if A has full column rank, then $f(x)$ is strongly convex. However, if A is rank deficient, then $f(x)$ is not strongly convex, but it is convex and satisfies the P-L condition. Examples of nonconvex functions which satisfy the P-L condition can be found in [44], [45].

Although it is difficult to precisely characterize the general class of functions for which the P-L condition is satisfied, in [44], one important special class was given as follows:

Lemma 1. *Let* $f(x)=g(Ax)$, *where* $g:\mathbb{R}^p\to\mathbb{R}$ *is a strongly convex function and* $A\in\mathbb{R}^{p\times p}$ *is a matrix, then* f *satisfies the P-L condition.*

We have the following global convergence results.

Theorem 2. *Suppose Assumptions 1–7 hold. For any given* $T\geq(c_0(\kappa_1,\kappa_2))^{1/\theta}$, *let* $\{\boldsymbol{x}_0,\dots,\boldsymbol{x}_T\}$ *be the sequence generated by Algorithm 1 with*

$$\alpha_k=\kappa_1\beta_k,\ \beta_k=(T+1)^\theta,\ \eta_k=\frac{\kappa_2}{\beta_k},\ \forall k\leq T, \quad (18)$$

where $\theta\in(0,1)$, $\kappa_1>c_1$, $\kappa_2\in(0,c_2(\kappa_1))$. *Then,*

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,T}-\bar{x}_T\|^2\Big]=\mathcal{O}(\frac{1}{T^{2\theta}}), \quad (19a)$$

$$\mathbf{E}[f(\bar{x}_T)-f^*]=\mathcal{O}(\frac{1}{nT^\theta})+\mathcal{O}(\frac{1}{T^{2\theta}}). \quad (19b)$$

Proof: The explicit expression of the right-hand side of (104) and proof are given in Appendix C. ■

From Theorem 2, we see that the convergence rate is strictly greater than $\mathcal{O}(1/T)$. In the following we show that the linear speedup convergence rate $\mathcal{O}(1/(nT))$ can be achieved if the P-L constant ν is known in advance and each $f_i^*>-\infty$, where $f_i^*=\min_{x\in\mathbb{R}^p}f_i(x)$. The total number of iterations T is not needed.

Theorem 3 (Linear Speedup). *Suppose Assumptions 1–7 hold, and the Polyak-Łojasiewicz constant* ν *is known in advance, and each* $f_i^*>-\infty$. *Let* $\{\boldsymbol{x}_k\}$ *be the sequence generated by Algorithm 1 with*

$$\alpha_k=\kappa_1\beta_k,\ \beta_k=\kappa_0(k+t_1),\ \eta_k=\frac{\kappa_2}{\beta_k},\ \forall k\in\mathbb{N}_0, \quad (20)$$

where $\kappa_0\in(0,\nu\hat{c}_0(\kappa_1,\kappa_2))$, $\kappa_1>c_1$, $\kappa_2\in(0,\hat{c}_2(\kappa_1))$, *and* $t_1>\hat{c}_3(\kappa_0,\kappa_1,\kappa_2)$ *with* $\hat{c}_0(\kappa_1,\kappa_2)$, $\hat{c}_2(\kappa_1)$, $\hat{c}_3(\kappa_0,\kappa_1,\kappa_2)$ *defined in Appendix D. Then, for any* $T\in\mathbb{N}_+$,

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,T}-\bar{x}_T\|^2\Big]=\mathcal{O}(\frac{1}{T^2}), \quad (21a)$$

$$\mathbf{E}[f(\bar{x}_T)-f^*]=\mathcal{O}(\frac{1}{nT})+\mathcal{O}(\frac{1}{T^2}). \quad (21b)$$

Proof: The explicit expressions of the right-hand sides of (21a) and (21b), and the proof are given in Appendix D. ■

Remark 3. *Note that it has been shown in [43] that* $\mathcal{O}(1/T)$ *convergence rate is optimal for centralized strongly convex* *optimization. This rate has been established by various distributed SGD algorithms when each local cost function is strongly convex, e.g., [11], [19], [35]–[39]. In contrast, the linear speedup convergence rate $\mathcal{O}(1/(nT))$ established in Theorem 3 only requires that the global cost function satisfies the P-L condition, but no convexity assumption is required neither on the global cost function nor on the local cost functions. The SGD algorithms in [17], [18], [27], [32]–[34] also achieve the linear speedup convergence. However, the algorithms in [17], [18], [32] are restricted to a star graph, while our algorithm is applicable to an arbitrarily connected graph. Moreover, [17], [18] assumed that the mean of each local stochastic gradient is the gradient of the global cost function, and T has to be known to choose the algorithm parameters. The algorithm in [18] furthermore requires an exponentially increasing batch size, which is not favorable in practice. In [32], it was assumed that the global cost function is strongly convex. In [27], [32], it was assumed that each local stochastic gradient has bounded second moment. In [27], [33], [34], it was assumed that each local cost function is strongly convex. It is one of our future research directions to establish the linear speedup convergence with reduced communication rounds and communication efficiency for an arbitrarily connected graph.*

Theorem 3 show that the convergence rate to a global optimum is sublinear when we allow the algorithm parameters to be time-varying. The following theorem establishes that the output of Algorithm 1 with fixed algorithm parameters linearly converges to a neighborhood of a global optimum.

Theorem 4. *Suppose Assumptions 1–7 hold. Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Algorithm 1 with*

$$\alpha_k = \kappa_1\beta,\ \beta_k = \beta,\ \eta_k = \frac{\kappa_2}{\beta},\ \forall k \in \mathbb{N}_0, \tag{22}$$

where $\kappa_1 > c_1$, $\kappa_2 \in (0, c_2(\kappa_1))$, and $\beta \ge c_0(\kappa_1,\kappa_2)$ with $c_0(\kappa_1,\kappa_2),\ c_2(\kappa_1) > 0$ defined in Appendix B. Then,

$$\begin{aligned}&\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k}-\bar{x}_k\|^2 + f(\bar{x}_k) - f^*\Big]\\&\le (1-\eta\varepsilon)^k c_1 + c_2\eta\sigma^2,\ \forall k \in \mathbb{N}_+,\end{aligned} \tag{23}$$

where $\varepsilon \in (0, 1/\eta)$, c_1, $c_2 > 0$ are constants defined in Appendix E.

Proof: The proof is given in Appendix E. ■

Remark 4. *It should be highlighted that we do not need to know the P-L constant ν in advance. Similar convergence result as stated in* (23) *was achieved by the distributed SGD algorithms proposed in [19], [39]–[41] when each local cost function is strongly convex, which obviously is stronger than the P-L condition assumed in Theorem 4. In addition to the strong convexity condition, in [19], it was also assumed that each local cost function is Lipschitz-continuous. Some information related to the Lyapunov function and global parameters, which may be difficult to get, were furthermore needed to design the stepsize. Moreover, in [39]–[41], the strong convexity constant was needed to design the stepsize and in [40], [41], a p-dimensional auxiliary variable, which is used to track the global gradient, was communicated between agents. The potential drawbacks of the results stated in Theorem 4 are that (i) we use undirected graphs rather than directed graphs as considered in [41]; and (ii) we do not analyze the robustness level to gradient noise as [39] did. We leave the extension to the (time-varying) directed graphs and the robustness level analysis as future research directions.*

IV. Simulations

In this section, we evaluate the performance of the proposed distributed primal-dual SGD algorithm through numerical experiments. All algorithms and agents are implemented and simulated in MATLAB R2018b, run on a desktop with `Intel Core i5-9600K processor, Nvidia RTX 2070 super, 32 GB RAM, Ubuntu 16.04`.

A. Neural Networks

We consider the training of neural networks (NN) for image classification tasks of the database MNIST [46]. The same NN is adopted as in [21] for each agent and the communication graph is generated randomly. The graph is shown in Fig. 1 and the corresponding Laplacian matrix $\boldsymbol{L}$ is given in (24). The corresponding mixing matrix $\boldsymbol{W}$ is constructed by metropolis weight, which is given in (25).

$$\boldsymbol{L} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ -1 & 3 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & -1 & 3 & -1 & 0 & 0 & -1 & 0 & 0 & 0\\ 0 & -1 & -1 & 4 & -1 & -1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -1 & -1 & 2 & 0 & 0 & 0 & 0\\ 0 & 0 & -1 & 0 & 0 & 0 & 2 & -1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}. \tag{24}$$

$$\boldsymbol{W} = \begin{bmatrix} 3/4 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 3/10 & 1/4 & 1/5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 3/10 & 1/5 & 0 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/5 & 7/15 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/5 & 1/3 & 7/15 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 0 & 0 & 0 & 5/12 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 2/3 \end{bmatrix}. \tag{25}$$

Each local neural network consists of a single hidden layer of 50 neurons, followed by a sigmoid activation layer, followed by the output layer of 10 neurons and another sigmoid activation layer. In this experiment, we use a subset of MNIST data set. Each agent is assigned 2500 data points randomly, and at each iteration, only one data point is picked up by the agent following a uniform distribution.

We compare our proposed distributed primal-dual SGD algorithm with time-varying and fixed parameters (DPD-SGD-T and DPD-SGD-F) with state-of-the-art algorithms: the distributed momentum SGD algorithm (DM-SGD) [16], the distributed SGD algorithm (D-SGD-1) [19], [20], the distributed SGD algorithm (D-SGD-2) [21], D^2 [29], the distributed stochastic gradient tracking algorithm (D-SGT-1) [30], [41], the distributed stochastic gradient tracking algorithm (D-SGT-2) [31], [40], and the baseline centralized SGD algorithm (C-SGD). We list all the parameters[1] we choose in the NN experiment for each algorithm in Table II.

We demonstrate the result in terms of the empirical risk function [47], which is given as

$$R(\boldsymbol{z}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_n}\sum_{j=1}^{m_n}\sum_{k=0}^{9}(t_k \ln y_k(\boldsymbol{x},\boldsymbol{z}) + (1-t_k)\ln(1-y_k(\boldsymbol{x},\boldsymbol{z})))$$

where m_n indicates the size of data set for each agent, t_k denotes the target (ground truth) of digit k corresponding to a single image, $\boldsymbol{x}$ is a single image input, $\boldsymbol{z} = (z^{(1)}, z^{(2)})$ with $z^{(1)}$ and $z^{(2)}$ being the weights in the 2 layers separately, and $y_k \in [0,1]$ is the output which expresses the probability of digit $k = 0, \ldots, 9$. The mapping from input to output is given as:

$$y_k(\boldsymbol{x},\boldsymbol{z}) = \sigma\left(\sum_{j=0}^{50} z_{k,j}^{(2)}\sigma\left(\sum_{i=0}^{28\times 28} z_{j,i}^{(1)}x_i\right)\right),$$

where $\sigma(s) = \frac{1}{1+\exp(-s)}$ is the sigmoid function.

Fig. 2 shows that the proposed distributed primal-dual SGD algorithm with time-varying parameters converges almost as fast as the distributed SGD algorithm in [19], [20] and faster than the distributed SGD algorithms in [21], [29]–[31], [40], [41] and the centralized SGD algorithm. Note that our algorithm converges slower than the distributed momentum SGD algorithm [16]. This is reasonable since that algorithm is an accelerated algorithm with extra requirement on the cost functions, i.e., the deviations between the gradients of local cost functions is bounded, and it requires each agent to communicate two p-dimensional variables with its neighbors at each iteration. The slope of the curves are however almost the same. The accuracy of each algorithm is given in Table III.

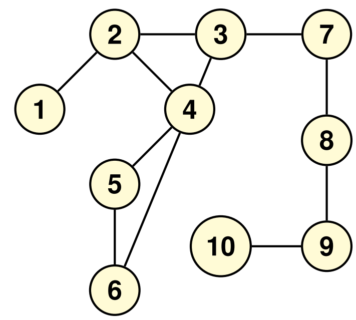

Fig. 1: Connection Topology.

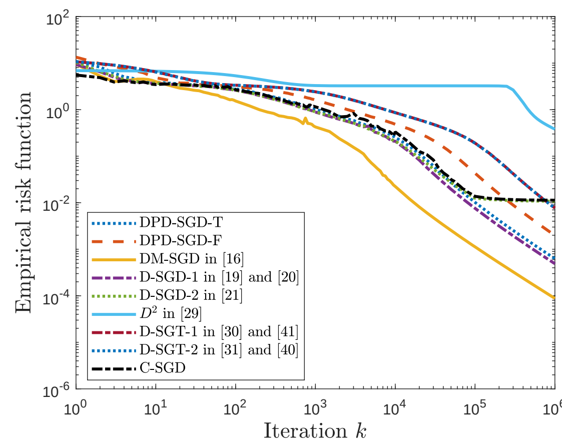

Fig. 2: Empirical Risk.

B. Convolutional Neural Networks

Let us consider the training of a convolutional neural networks (CNN) model. We build a CNN model for each agent with five 3×3 convolutional layers using ReLU as activation function, one average pooling layer with filters of size 2×2, one sigmoid layer with dimension 360, another sigmoid layer with dimension 60, one softmax layer with dimension 10. In this experiment, we use the whole MNIST data set. We use

[1]Note: the parameter names are different in each paper.

TABLE II: Parameters in each algorithm in NN experiment.

Algorithm	η_k	α_k	β_k
DPD-SGD-T	$0.08/k^{10^{-5}}$	$4k^{10^{-5}}$	$3k^{10^{-5}}$
DPD-SGD-F	0.03	5	20
DM-SGD [16]	0.1	✗	0.8
D-SGD-1 [19], [20]	0.1	✗	✗
D-SGD-2 [21]	✗	$0.1/(10^{-5}k+1)$	$0.2/(10^{-5}k+1)^{0.3}$
D^2 [29]	0.01	✗	✗
D-SGT-1 [30], [41]	0.01	✗	✗
D-SGT-2 [31], [40]	0.01	✗	✗
C-SGD	0.1	✗	✗

TABLE III: Accuracy on each algorithm in NN experiment.

Algorithm	Accuracy
DPD-SGD-T	93.04%
DPD-SGD-F	92.76%
DM-SGD [16]	93.44%
D-SGD-1 [19], [20]	92.96%
D-SGD-2 [21]	92.88%
D^2 [29]	90.44%
D-SGT-1 [30], [41]	92.88%
D-SGT-2 [31], [40]	92.96%
C-SGD	93%

the same communication graph as in above NN experiment. Each agent is assigned 6000 data points randomly. We set the batch size as 20, which means at each iteration, 20 data points are chosen by the agent to update the gradient, which is also following a uniform distribution. For each algorithm, we do 10 epochs to train the CNN model.

We compare our algorithms DPD-SGD-T and DPD-SGD-F with the fastest one above: DM-SGD, D-SGD-1, and C-SGD. We list all the parameters we choose in the CNN experiment for each algorithm in Table IV.

We demonstrate the training loss and the test accuracy of each algorithm in Fig. 3 and Fig. 4. Here we use Categorical Cross-Entropy loss, which is a softmax activation plus a Cross-Entropy loss. We can see that our algorithms perform almost the same as the DM-SGD and better than the D-SGD-1 and the centralized C-SGD. The accuracy of each algorithm is given in Table V.

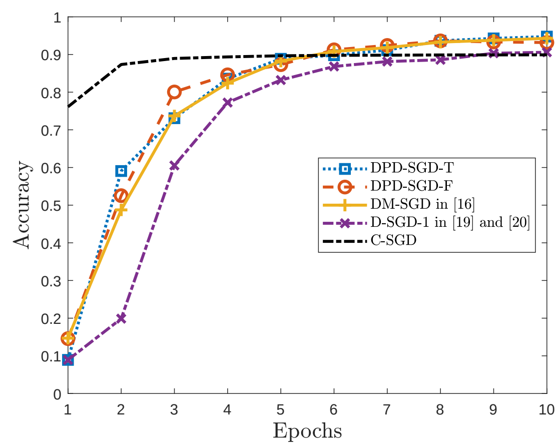

Fig. 4: CNN accuracy.

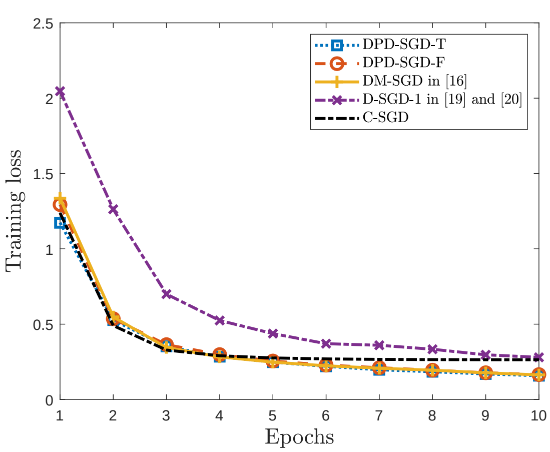

Fig. 3: CNN training loss.

V. Conclusions

In this paper, we studied distributed nonconvex optimization. We proposed a distributed primal-dual SGD algorithm and derived its convergence rate. More specifically, the linear speedup convergence rate $\mathcal{O}(1/\sqrt{nT})$ was established for smooth nonconvex cost functions under arbitrarily connected communication networks. The convergence rate was improved to the linear speedup convergence rate $\mathcal{O}(1/(nT))$ when the global cost function satisfies the P-L condition. It was also shown that the output of the proposed algorithm with fixed parameters linearly converges to a neighborhood of a global optimum. Interesting directions for future work include establishing linear speedup convergence under the P-L condition while considering communication reduction with asynchronous, periodic, or compressed communication.

TABLE IV: Parameters in each algorithm in CNN experiment.

Algorithm	η_k	α_k	β_k
DPD-SGD-T	$0.5/k^{10^{-5}}$	$0.5k^{10^{-5}}$	$0.1k^{10^{-5}}$
DPD-SGD-F	0.5	0.5	0.1
DM-SGD [16]	0.1	✗	0.8
D-SGD [19], [20]	0.1	✗	✗
C-SGD	0.1	✗	✗

TABLE V: Accuracy on each algorithm in CNN experiment.

Algorithm	Accuracy
DPD-SGD-T	94.75%
DPD-SGD-F	93.17%
DM-SGD [16]	94.29%
D-SGD [19], [20]	92.96%
C-SGD	89.91%

References

[1] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *International Conference on Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[3] J. Langford, L. Li, and T. Zhang, "Sparse online learning via truncated gradient," *Journal of Machine Learning Research*, vol. 10, pp. 777–801, 2009.

[4] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2011, pp. 693–701.

[5] C. M. De Sa, C. Zhang, K. Olukotun, and C. Ré, "Taming the wild: A unified analysis of hogwild-style algorithms," in *Advances in Neural Information Processing Systems*, 2015, pp. 2674–2682.

[6] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Advances in Neural Information Processing Systems*, 2015, pp. 2737–2745.

[7] 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.

[8] Z. Zhou, P. Mertikopoulos, N. Bambos, P. Glynn, Y. Ye, L.-J. Li, and F.-F. Li, "Distributed asynchronous optimization with unbounded delays: How slow can you go?" in *International Conference on Machine Learning*, 2018, pp. 5970–5979.

[9] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning*, 2018, pp. 560–569.

[10] P. Jiang and G. Agrawal, "A linear speedup analysis of distributed deep learning with sparse and quantized communication," in *Advances in Neural Information Processing Systems*, 2018, pp. 2525–2536.

[11] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," *arXiv preprint arXiv:1909.13014*, 2019.

[12] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 668–14 679.

[13] J. Wang and G. Joshi, "Adaptive communication strategies to achieve the best error-runtime trade-off in local-update SGD," in *Conference on Machine Learning and Systems*, 2019.

[14] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5693–5700.

[15] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. Cadambe, "Trading redundancy for communication: Speeding up distributed SGD for non-convex optimization," in *International Conference on Machine Learning*, 2019, pp. 2545–2554.

[16] 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.

[17] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. Cadambe, "Local SGD with periodic averaging: Tighter analysis and adaptive synchronization," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 080–11 092.

[18] H. Yu and R. Jin, "On the computation and communication complexity of parallel SGD with dynamic batch sizes for stochastic non-convex optimization," in *International Conference on Machine Learning*, 2019, pp. 7174–7183.

[19] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, "Collaborative deep learning in fixed topology networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 5904–5914.

[20] 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.

[21] J. George, T. Yang, H. Bai, and P. Gurram, "Distributed stochastic gradient method for non-convex problems with applications in supervised learning," in *IEEE Conference on Decision and Control*, 2019, pp. 5538–5543.

[22] X. Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous decentralized parallel stochastic gradient descent," in *International Conference on Machine Learning*, 2018, pp. 3043–3052.

[23] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *International Conference on Machine Learning*, 2019, pp. 344–353.

[24] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," in *Advances in Neural Information Processing Systems*, 2018, pp. 7652–7662.

[25] A. Reisizadeh, H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Robust and communication-efficient collaborative learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 8386–8397.

[26] H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized push-sum for gossip and decentralized optimization over directed graphs," *arXiv preprint arXiv:2002.09964*, 2020.

[27] N. Singh, D. Data, J. George, and S. Diggavi, "SQuARM-SGD: Communication-efficient momentum SGD for decentralized optimization," *arXiv preprint arXiv:2005.07041*, 2020.

[28] J. Wang and G. Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms," *arXiv preprint arXiv:1808.07576*, 2018.

[29] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D^2: Decentralized training over decentralized data," in *International Conference on Machine Learning*, 2018, pp. 4848–4856.

[30] S. Lu, X. Zhang, H. Sun, and M. Hong, "GNSD: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization," in *IEEE Data Science Workshop*, 2019, pp. 315–321.

[31] J. Zhang and K. You, "Decentralized stochastic gradient tracking for empirical risk minimization," *arXiv preprint arXiv:1909.02712*, 2019.

[32] S. U. Stich, "Local SGD converges fast and communicates little," in *International Conference on Learning Representations*, 2019.

[33] 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.
[34] A. Olshevsky, I. C. Paschalidis, and S. Pu, "A non-asymptotic analysis of network independence for distributed stochastic gradient descent," *arXiv preprint arXiv:1906.02702*, 2019.
[35] M. Rabbat, "Multi-agent mirror descent for decentralized stochastic optimization," in *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2015, pp. 517–520.
[36] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *Mathematical Programming*, pp. 1–48, 2018.
[37] D. Yuan, Y. Hong, D. W. Ho, and G. Jiang, "Optimal distributed stochastic mirror descent for strongly convex optimization," *Automatica*, vol. 90, pp. 196–203, 2018.
[38] D. Jakovetic, D. Bajovic, A. K. Sahu, and S. Kar, "Convergence rates for distributed stochastic optimization over random networks," in *IEEE Conference on Decision and Control*, 2018, pp. 4238–4245.
[39] A. Fallah, M. Gurbuzbalaban, A. Ozdaglar, U. Simsekli, and L. Zhu, "Robust distributed accelerated stochastic gradient methods for multi-agent networks," *arXiv preprint arXiv:1910.08701*, 2019.
[40] S. Pu and A. Nedić, "A distributed stochastic gradient tracking method," in *IEEE Conference on Decision and Control*, 2018, pp. 963–968.
[41] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *IEEE Conference on Decision and Control*, 2019, pp. 8353–8358.
[42] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, 2010.
[43] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *International Conference on Machine Learning*, 2012, pp. 1571–1578.
[44] 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.
[45] 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.
[46] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," *Available: http://yann. lecun. com/exdb/mnist*, 2010.
[47] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 421–436.
[48] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer International Publishing, 2018.
[49] Y. Tang, J. Zhang, and N. Li, "Distributed zero-order algorithms for non-convex multi-agent optimization," *arXiv preprint arXiv:1908.11444v3*, 2020.
[50] 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.

Appendix

A. Notations and Useful Lemmas

$\mathbf{I}_n$ is the n-dimensional identity matrix. 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, $M\ge N$ 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 $x^\top Ax$. $\lceil\cdot\rceil$ 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. For any $n\in\mathbb{N}_0$, $n!$ is the factorial of n.

Denote $K_n=\mathbf{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, $\boldsymbol{K}=K_n\otimes\mathbf{I}_p$, $\boldsymbol{H}=\frac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top\otimes\mathbf{I}_p)$, $\bar{x}_k=\frac{1}{n}(\mathbf{1}_n^\top\otimes\mathbf{I}_p)\boldsymbol{x}_k$, $\bar{\boldsymbol{x}}_k=\mathbf{1}_n\otimes\bar{x}_k$, $\boldsymbol{g}_k=\nabla\tilde{f}(\boldsymbol{x}_k)$, $\bar{\boldsymbol{g}}_k=\boldsymbol{H}\boldsymbol{g}_k$, $\boldsymbol{g}_k^0=\nabla\tilde{f}(\bar{\boldsymbol{x}}_k)$, $\bar{\boldsymbol{g}}_k^0=\boldsymbol{H}\boldsymbol{g}_k^0=\mathbf{1}_n\otimes\nabla f(\bar{x}_k)$, and $\bar{\boldsymbol{g}}_k^u=\boldsymbol{H}\boldsymbol{g}_k^u$.

The following results are used in the proofs.

Lemma 2. *(Lemma 1.2.3 in [48] and Lemma 3 in [49]) If the function $f(x):\ \mathbb{R}^p\mapsto\mathbb{R}$ is smooth with constant $L_f>0$, then*

$$|f(y)-f(x)-(y-x)^\top\nabla f(x)|\le\frac{L_f}{2}\|y-x\|^2, \tag{26a}$$
$$\|\nabla f(x)\|^2\le 2L_f(f(x)-f^*),\ \forall x,y\in\mathbb{R}^p, \tag{26b}$$

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

Lemma 3. *(Lemmas 1 and 2 in [50]) Let L be the Laplacian matrix of the graph $\mathcal{G}$. If Assumption 1 holds, then L is positive semi-definite,* $\mathrm{null}(L)=\mathrm{null}(K_n)=\{\mathbf{1}_n\}$, $L\le\rho(L)\mathbf{I}_n$, $\rho(K_n)=1$,

$$K_nL=LK_n=L, \tag{27}$$
$$0\le\rho_2(L)K_n\le L\le\rho(L)K_n. \tag{28}$$

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

$$R\Lambda_1^{-1}R^\top L=LR\Lambda_1^{-1}R^\top=K_n, \tag{29}$$
$$\frac{1}{\rho(L)}K_n\le R\Lambda_1^{-1}R^\top\le\frac{1}{\rho_2(L)}K_n, \tag{30}$$

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.

Lemma 4. *Let $a\in(0,1)$ be a constant, then*

$$(1-a)^T\le\frac{k!}{(aT)^k},\ \forall k,T\in\mathbb{N}_0. \tag{31}$$

Proof: For any constant $a\in(0,1)$, we have $\ln(1-a)\le-a$. Thus,

$$(1-a)^T\le e^{-aT},\ \forall T\in\mathbb{N}_0. \tag{32}$$

For any constant $x>0$, we have $e^x>\frac{x^k}{k!}$, $\forall k\in\mathbb{N}_0$. This result together with (32) yields (31). ■

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

$$z_k\ge0, \tag{33a}$$
$$z_{k+1}\le(1-r_{1,k})z_k+r_{2,k}, \tag{33b}$$
$$1>r_{1,k}\ge\frac{a_1}{(k+t_1)^\delta}, \tag{33c}$$
$$r_{2,k}\le\frac{a_2}{(k+t_1)^2},\ \forall k\in\mathbb{N}_0, \tag{33d}$$

where $\delta\ge0$, $a_1>0$, and $a_2>0$ are constants.

(i) If $\delta=1$ and $a_1>1$, then

$$z_k\le\phi_1(k,t_1,a_1,a_2,z_0),\ \forall k\in\mathbb{N}_+, \tag{34}$$

where

$$\phi_1(k,t_1,a_1,a_2,z_0)=\frac{t_1^{a_1}z_0}{(k+t_1)^{a_1}}+\frac{a_2}{(k+t_1-1)^2}+\frac{4a_2}{(a_1-1)(k+t_1)}. \tag{35}$$

(ii) If $\delta=0$, then

$$z_k \le \phi_2(k,t_1,a_1,a_2,z_0),\ \forall k\in\mathbb{N}_+, \tag{36}$$

where

$$\begin{aligned}
&\phi_2(k,t_1,a_1,a_2,z_0)\\
&=(1-a_1)^k z_0+a_2(1-a_1)^{k+t_1-1}\Big([t_2-t_1]_+ s_3(t_1)\\
&\quad+([t_3-t_1]_+-[t_2-t_1]_+)s_3(t_3)\Big)\\
&\quad+\frac{\mathbf{1}_{(k+t_1-1\ge t_3)}2a_2}{-\ln(1-a_1)(k+t_1)^2(1-a_1)}
\end{aligned} \tag{37}$$

with $s_3(k)=\frac{1}{k^2(1-a_1)^k}$, $t_2=\lceil\frac{-2}{\ln(1-a_1)}\rceil$, *and* $t_3=\lceil\frac{-4}{\ln(1-a_1)}\rceil$.

Proof: (i) From (33a)–(33c), for any $k\in\mathbb{N}_+$, it holds that

$$\begin{aligned}
z_k \le& \prod_{\tau=0}^{k-1}(1-r_{1,\tau})z_0+r_{2,k-1}\\
&+\sum_{l=0}^{k-2}\prod_{\tau=l+1}^{k-1}(1-r_{1,\tau})r_{2,l}.
\end{aligned} \tag{38}$$

For any $t\in[0,1]$, it holds that $1-t\le e^{-t}$ since $s_1(t)=1-t-e^{-t}$ is a non-increasing function in the interval $[0,1]$. Thus, for any $k>l\ge 0$, it holds that

$$\prod_{\tau=l}^{k-1}(1-r_{1,\tau})\le e^{-\sum_{\tau=l}^{k-1}r_{1,\tau}}. \tag{39}$$

We also have

$$\begin{aligned}
&\sum_{\tau=l}^{k-1}r_{1,\tau}\ge\sum_{\tau=l}^{k-1}\frac{a_1}{\tau+t_1}=\sum_{\tau=l+t_1}^{k-1+t_1}\frac{a_1}{\tau}\\
&\ge\int_{t=l+t_1}^{k+t_1}\frac{a_1}{t}dt=a_1(\ln(k+t_1)-\ln(l+t_1)),
\end{aligned} \tag{40}$$

where the first inequality holds due to (33c) and the second inequality holds since $s_2(t)=a_1/t$ is a decreasing function in the interval $[1,+\infty)$.

Hence, (39) and (40) yield

$$\prod_{\tau=l}^{k-1}(1-r_{1,\tau})\le e^{-\sum_{\tau=l}^{k-1}r_{1,\tau}}\le\frac{(l+t_1)^{a_1}}{(k+t_1)^{a_1}}. \tag{41}$$

We have

$$\begin{aligned}
&\sum_{l=0}^{k-2}\prod_{\tau=l+1}^{k-1}(1-r_{1,\tau})r_{2,l}\\
&\le\sum_{l=0}^{k-2}\frac{(l+t_1+1)^{a_1}}{(k+t_1)^{a_1}}\frac{a_2}{(l+t_1)^2}\\
&\le\sum_{l=0}^{k-2}\frac{(l+t_1+1)^{a_1}}{(k+t_1)^{a_1}}\frac{a_2}{(\frac{t_1}{t_1+1}l+t_1)^2}\\
&=\frac{(\frac{t_1+1}{t_1})^2a_2}{(k+t_1)^{a_1}}\sum_{l=0}^{k-2}\frac{(l+t_1+1)^{a_1}}{(l+t_1+1)^2}\\
&=\frac{4a_2}{(k+t_1)^{a_1}}\sum_{l=t_1+1}^{k+t_1-1}l^{a_1-2}\le\frac{4a_2}{(a_1-1)(k+t_1)},
\end{aligned} \tag{42}$$

where the first inequality holds due to (41) and (33d).

From (38), (41), and (42), we have (34).

(ii) Denote $a=1-a_1$. From (33c) and $\delta=0$, we know that $a_1\in(0,1)$. Thus, $a\in(0,1)$.

From (33a)–(33d) and $\delta_1=0$, for any $k\in\mathbb{N}_+$, it holds that

$$\begin{aligned}
z_k\le&(1-a_1)^kz_0+\sum_{\tau=0}^{k-1}(1-a_1)^{k-1-\tau}r_{2,\tau}\\
\le&a^kz_0+a_2a^{k+t_1-1}\sum_{\tau=0}^{k-1}\frac{1}{(\tau+t_1)^2a^{\tau+t_1}}.
\end{aligned} \tag{43}$$

We have

$$\begin{aligned}
&\sum_{\tau=0}^{k-1}\frac{1}{(\tau+t_1)^2a^{\tau+t_1}}=\sum_{\tau=t_1}^{k+t_1-1}\frac{1}{\tau^2a^\tau}\\
&=\sum_{\tau=t_1}^{t_2-1}s_3(\tau)+\sum_{\tau=t_2}^{t_3-1}s_3(\tau)+\sum_{\tau=t_3}^{k+t_1-1}s_3(\tau).
\end{aligned} \tag{44}$$

We know that $s_3(t)=1/(t^2a^t)$ is decreasing and increasing in the intervals $[1,t_2-1]$ and $[t_2,+\infty)$, respectively, since

$$\begin{aligned}
\frac{ds_3(t)}{dt}=&-s_3(t)\Big(\frac{2}{t}+\ln(a)\Big)\le 0,\ \forall t\in\Big(0,\frac{-2}{\ln(a)}\Big],\\
\frac{ds_3(t)}{dt}=&-s_3(t)\Big(\frac{2}{t}+\ln(a)\Big)\ge 0,\ \forall t\in\Big[\frac{-2}{\ln(a)},+\infty\Big).
\end{aligned}$$

Thus, we have

$$\sum_{\tau=k_1}^{t_2-1}s_3(\tau)\le(t_2-k_1)s_3(k_1),\ \forall k_1\in[1,t_2-1], \tag{45a}$$

$$\sum_{\tau=k_2}^{t_3-1}s_3(\tau)\le(t_3-k_2)s_3(t_3),\ \forall k_2\in[t_2,t_3-1], \tag{45b}$$

$$\sum_{\tau=t_3}^{k_3}s_3(\tau)\le\int_{t_3}^{k_3+1}s_3(t)dt,\ \forall k_3\ge t_3. \tag{45c}$$

Denote $b=1/a$. We have

$$\begin{aligned}
&\int_{t_3}^{k_3+1}s_3(t)dt=\int_{t_3}^{k_3+1}\frac{b^t}{t^2}dt=\int_{t_3}^{k_3+1}\frac{1}{\ln(b)t^2}db^t\\
&=\frac{b^{k_3+1}}{\ln(b)(k_3+1)^2}-\frac{b^{t_3}}{\ln(b)t_3^2}+\int_{t_3}^{k_3+1}\frac{2b^t}{\ln(b)t^3}dt\\
&\le\frac{b^{k_3+1}}{\ln(b)(k_3+1)^2}+\int_{t_3}^{k_3+1}\frac{2}{\ln(b)t}s_3(t)dt\\
&\le\frac{b^{k_3+1}}{\ln(b)(k_3+1)^2}+\frac{2}{\ln(b)t_3}\int_{t_3}^{k_3+1}s_3(t)dt\\
&\le\frac{b^{k_3+1}}{\ln(b)(k_3+1)^2}+\frac{1}{2}\int_{t_3}^{k_3+1}s_3(t)dt,
\end{aligned} \tag{46}$$

where the last inequality holds due to $t_3=\lceil\frac{-4}{\ln(1-a_1)}\rceil\ge\frac{-4}{\ln(1-a_1)}=\frac{4}{\ln(b)}$.

From (45c) and (46), we have

$$\sum_{\tau=t_3}^{k_3} s_3(\tau) \le \frac{-2}{\ln(a)(k_3+1)^2 a^{k_3+1}},\ \forall k_3 \ge t_3. \tag{47}$$

From (43), (44), (45a), (45b), and (47), we get (36). ■

Lemma 6. *Suppose Assumptions 1 and 3–6 hold. Then the following holds for Algorithm 1*

$$\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{1,k+1}]\\
&\le W_{1,k} - \|\boldsymbol{x}_k\|^2_{\eta_k\alpha_k\boldsymbol{L}-\frac{1}{2}\eta_k\boldsymbol{K}-\frac{3}{2}\eta_k^2\alpha_k^2\boldsymbol{L}^2-\frac{1}{2}\eta_k(1+5\eta_k)L_f^2\boldsymbol{K}}\\
&\quad - \eta_k\beta_k\boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\frac{3}{2}\eta_k^2\beta_k^2\boldsymbol{K}}\\
&\quad + 2n\sigma^2\eta_k^2,
\end{aligned} \tag{48}$$

where $W_{1,k}=\frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}$.

Proof: Noting that $\nabla\tilde{f}$ is Lipschitz-continuous with constant $L_f>0$ since Assumption 3 is satisfied, we have that

$$\|\boldsymbol{g}_k^0-\boldsymbol{g}_k\|^2 \le L_f^2\|\bar{\boldsymbol{x}}_k-\boldsymbol{x}_k\|^2 = L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}. \tag{49}$$

From Assumptions 4–6, we know that

$$\mathbf{E}_{\mathfrak{F}_k}[\boldsymbol{g}_k^u]=\boldsymbol{g}_k, \tag{50}$$

$$\mathbf{E}_{\mathfrak{F}_k}[\|\boldsymbol{g}_k^u-\boldsymbol{g}_k\|^2]\le n\sigma^2. \tag{51}$$

From (49), (51), and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbf{E}_{\mathfrak{F}_k}[\|\boldsymbol{g}_k^0-\boldsymbol{g}_k^u\|^2] &= \mathbf{E}_{\mathfrak{F}_k}[\|\boldsymbol{g}_k^0-\boldsymbol{g}_k+\boldsymbol{g}_k-\boldsymbol{g}_k^u\|^2]\\
&\le 2\|\boldsymbol{g}_k^0-\boldsymbol{g}_k\|^2+2\mathbf{E}_{\mathfrak{U}_k}[\|\boldsymbol{g}_k-\boldsymbol{g}_k^u\|^2]\\
&\le 2L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+2n\sigma^2.
\end{aligned} \tag{52}$$

We have

$$\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{1,k+1}] = \mathbf{E}_{\mathfrak{F}_k}\Big[\frac{1}{2}\|\boldsymbol{x}_{k+1}\|^2_{\boldsymbol{K}}\Big]\\
&= \mathbf{E}_{\mathfrak{F}_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^u)\|^2_{\boldsymbol{K}}\Big]\\
&= \mathbf{E}_{\mathfrak{F}_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(\mathbf{I}_{np}-\eta_k\alpha_k\boldsymbol{L})\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^u\Big)\\
&\quad +\frac{1}{2}\eta_k^2\beta_k^2\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^u\Big\|^2_{\boldsymbol{K}}\Big]\\
&= \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{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(\mathbf{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-\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&\quad +\frac{1}{2}\eta_k^2\beta_k^2\mathbf{E}_{\mathfrak{F}_k}\Big[\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0+\frac{1}{\beta_k}\boldsymbol{g}_k^u-\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}\\
&\quad -\eta_k\beta_k\boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&\quad +\frac{\eta_k}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+\frac{\eta_k}{2}\|\boldsymbol{g}_k-\boldsymbol{g}_k^0\|^2\\
&\quad +\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}_k^0\Big\|^2_{\boldsymbol{K}}\\
&\quad +\frac{1}{2}\eta_k^2\alpha_k^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{L}^2}+\frac{1}{2}\eta_k^2\|\boldsymbol{g}_k-\boldsymbol{g}_k^0\|^2\\
&\quad +\eta_k^2\beta_k^2\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}+\eta_k^2\mathbf{E}_{\mathfrak{F}_k}[\|\boldsymbol{g}_k^u-\boldsymbol{g}_k^0\|^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}\\
&\quad +\frac{\eta_k}{2}(1+\eta_k)\|\boldsymbol{g}_k-\boldsymbol{g}_k^0\|^2+\eta_k^2\mathbf{E}_{\mathfrak{F}_k}[\|\boldsymbol{g}_k^u-\boldsymbol{g}_k^0\|^2]\\
&\quad -\eta_k\beta_k\boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)+\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}},
\end{aligned} \tag{53}$$

where the second equality holds due to (5a); the third equality holds due to (27) in Lemma 3; the fourth equality holds since $\boldsymbol{x}_k$ and $\boldsymbol{v}_k$ are independent of $\mathfrak{F}_k$ and (50); and the inequality holds due to the Cauchy-Schwarz inequality and $\rho(\boldsymbol{K})=1$.

Then, from (49), (52), and (53), we have (48). ■

Lemma 7. *Suppose Assumptions 1 and 3 hold, and $\{\beta_k\}$ is non-decreasing. Then the following holds for Algorithm 1*

$$\begin{aligned}
&W_{2,k+1}\\
&\le W_{2,k}+(1+\omega_k)\eta_k\beta_k\boldsymbol{x}_k^\top(\boldsymbol{K}+\kappa_1\boldsymbol{L})\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&\quad +\frac{1}{2}(\eta_k+\omega_k+\eta_k\omega_k)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}\\
&\quad +\|\boldsymbol{x}_k\|^2_{(1+\omega_k)\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)}\\
&\quad +\frac{\eta_k}{\beta_k^2}\Big(\eta_k+\frac{1}{2}\Big)(1+\omega_k)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)L_f^2\|\bar{\boldsymbol{g}}_k^u\|^2\\
&\quad +\frac{1}{2}\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)(\omega_k+\omega_k^2)\|\boldsymbol{g}_{k+1}^0\|^2,
\end{aligned} \tag{54}$$

where $W_{2,k}=\frac{1}{2}\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}$, $\boldsymbol{Q}=R\Lambda_1^{-1}R^\top\otimes\mathbf{I}_p$ with matrices R and Λ_1^{-1} given in Lemma 3, $\omega_k=\frac{1}{\beta_k}-\frac{1}{\beta_{k+1}}$, and $\kappa_1>0$ is a constant.

Proof: Denote $\bar{v}_k=\frac{1}{n}(\mathbf{1}_n^\top\otimes\mathbf{I}_p)\boldsymbol{v}_k$. Then, from (5b), we know that

$$\bar{v}_{k+1}=\bar{v}_k. \tag{55}$$

Then, from (55) and $\sum_{i=1}^n v_{i,0}=\mathbf{0}_p$, we know that

$$\bar{v}_k=\mathbf{0}_p, \tag{56}$$

Then, from (56) and (5a), we know that

$$\bar{\boldsymbol{x}}_{k+1}=\bar{\boldsymbol{x}}_k-\eta_k\bar{\boldsymbol{g}}_k^u. \tag{57}$$

Since $\nabla\tilde{f}$ is Lipschitz-continuous and (57), we have

$$\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2\le L_f^2\|\bar{\boldsymbol{x}}_{k+1}-\bar{\boldsymbol{x}}_k\|^2=\eta_k^2L_f^2\|\bar{\boldsymbol{g}}_k^u\|^2. \tag{58}$$

We know that $\omega_k\ge 0$ since $\{\beta_k\}$ is non-decreasing. We have

$$\begin{aligned}
W_{2,k+1}&=\frac{1}{2}\Big\|\boldsymbol{v}_{k+1}+\frac{1}{\beta_{k+1}}\boldsymbol{g}_{k+1}^0\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&=\frac{1}{2}\Big\|\boldsymbol{v}_{k+1}+\frac{1}{\beta_k}\boldsymbol{g}_{k+1}^0+\Big(\frac{1}{\beta_{k+1}}-\frac{1}{\beta_k}\Big)\boldsymbol{g}_{k+1}^0\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&\le\frac{1}{2}(1+\omega_k)\Big\|\boldsymbol{v}_{k+1}+\frac{1}{\beta_k}\boldsymbol{g}_{k+1}^0\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&\quad+\frac{1}{2}(\omega_k+\omega_k^2)\|\boldsymbol{g}_{k+1}^0\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}},
\end{aligned} \tag{59}$$

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

For the first term in the right-hand side of (59), 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}}\\
&=\frac{1}{2}\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&\quad+\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}\|\boldsymbol{g}^0_{k+1}-\boldsymbol{g}^0_k\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&\quad+\frac{1}{\beta_k}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k+\eta_k\beta_k\boldsymbol{L}\boldsymbol{x}_k\Big)^\top(\boldsymbol{Q}+\kappa_1\boldsymbol{K})(\boldsymbol{g}^0_{k+1}-\boldsymbol{g}^0_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}^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}\|\boldsymbol{g}^0_{k+1}-\boldsymbol{g}^0_k\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&\quad+\frac{\eta_k}{2}\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}+\frac{1}{2\eta_k\beta_k^2}\|\boldsymbol{g}^0_{k+1}-\boldsymbol{g}^0_k\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\
&\quad+\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}^0_{k+1}-\boldsymbol{g}^0_k\|^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_{\eta_k^2\beta_k^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)}+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\frac{1}{2}\eta_k(\boldsymbol{Q}+\kappa_1\boldsymbol{K})}\\
&\quad+\frac{1}{\beta_k^2}\Big(1+\frac{1}{2\eta_k}\Big)\|\boldsymbol{g}^0_{k+1}-\boldsymbol{g}^0_k\|^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}^0_k\Big)\\
&\quad+\|\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}^0_k\Big\|^2_{\frac{1}{2}\eta_k(\boldsymbol{Q}+\kappa_1\boldsymbol{K})}\\
&\quad+\frac{1}{\beta_k^2}\Big(1+\frac{1}{2\eta_k}\Big)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\|\boldsymbol{g}^0_{k+1}-\boldsymbol{g}^0_k\|^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}^0_k\Big)\\
&\quad+\|\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}^0_k\Big\|^2_{\frac{1}{2}\eta_k(\boldsymbol{Q}+\kappa_1\boldsymbol{K})}\\
&\quad+\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}}^u_k\|^2,
\end{aligned}\tag{60}
$$

where the first equality holds due to (5b); the second equality holds due to (27) and (29) in Lemma 3; the first inequality holds due to the Cauchy-Schwarz inequality; the last equality holds due to (27) and (29) in Lemma 3; the second inequality holds due to $\rho(\boldsymbol{Q}+\kappa_1\boldsymbol{K})\le\rho(\boldsymbol{Q})+\kappa_1\rho(\boldsymbol{K})$, (30), $\rho(\boldsymbol{K})=1$; and the last inequality holds due to (58).

For the second term in the right-hand side of (59), we have

$$
\|\boldsymbol{g}^0_{k+1}\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\le\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\|\boldsymbol{g}^0_{k+1}\|^2. \tag{61}
$$

Also note that

$$
\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\le\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}. \tag{62}
$$

Then, from (59)–(62), we have (54). ■

Lemma 8. *Suppose Assumptions 1 and 3–6 hold, and $\{\beta_k\}$ in non-decreasing. Then the following holds for Algorithm 1*

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{3,k+1}]\\
&\le W_{3,k}-(1+\omega_k)\eta_k\alpha_k\boldsymbol{x}_k^\top\boldsymbol{L}(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k)\\
&\quad+\|\boldsymbol{x}_k\|^2_{\eta_k(\beta_k\boldsymbol{L}+\frac{1}{2}\boldsymbol{K})+\eta_k^2(\frac{1}{2}\alpha_k^2-\alpha_k\beta_k+\beta_k^2)\boldsymbol{L}^2}\\
&\quad+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\omega_k\eta_k\alpha_k\boldsymbol{L}^2+\frac{1}{2}\eta_k(1+3\eta_k)L_f^2\boldsymbol{K}}\\
&\quad+\frac{\eta_k}{2\beta_k^2}(1+3\eta_k)L_f^2\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}^u_k\|^2]+n\sigma^2\eta_k^2\\
&\quad-\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\eta_k(\beta_k-\frac{1}{2}-\eta_k\beta_k^2-\frac{1}{2}\omega_k\alpha_k)\boldsymbol{K}}\\
&\quad+\frac{1}{2}\omega_k\mathbf{E}_{\mathfrak{F}_k}[2W_{1,k+1}+\|\boldsymbol{g}^0_{k+1}\|^2].
\end{aligned}\tag{63}
$$

where $W_{3,k}=\boldsymbol{x}_k^\top\boldsymbol{K}(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k)$.

Proof: We have

$$
\begin{aligned}
W_{3,k+1}&=\boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1}+\frac{1}{\beta_{k+1}}\boldsymbol{g}^0_{k+1}\Big)\\
&=\boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1}+\frac{1}{\beta_k}\boldsymbol{g}^0_{k+1}+\Big(\frac{1}{\beta_{k+1}}-\frac{1}{\beta_k}\Big)\boldsymbol{g}^0_{k+1}\Big)\\
&=\boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1}+\frac{1}{\beta_k}\boldsymbol{g}^0_{k+1}\Big)-\omega_k\boldsymbol{x}_{k+1}^\top\boldsymbol{K}\boldsymbol{g}^0_{k+1}\\
&\le\boldsymbol{x}_{k+1}^\top\boldsymbol{K}\Big(\boldsymbol{v}_{k+1}+\frac{1}{\beta_k}\boldsymbol{g}^0_{k+1}\Big)\\
&\quad+\frac{1}{2}\omega_k(\|\boldsymbol{x}_{k+1}\|^2_{\boldsymbol{K}}+\|\boldsymbol{g}^0_{k+1}\|^2).
\end{aligned}\tag{64}
$$

For the first term in the right-hand side of (64), we have

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

$$
\begin{aligned}
&+\frac{1}{2}\eta_k^2\alpha_k^2\|\boldsymbol{L}\boldsymbol{x}_k\|^2+\frac{1}{2\beta_k^2}\mathbf{E}_{\mathfrak{F}_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{F}_k}[\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2]\\
&+\frac{1}{2}\eta_k\|\boldsymbol{g}_k-\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-\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{F}_k}[\|\boldsymbol{g}_k^u-\boldsymbol{g}_k^0\|^2]+\frac{1}{2\beta_k^2}\mathbf{E}_{\mathfrak{F}_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-\boldsymbol{g}_k^0\|^2+\frac{1}{2}\eta_k^2\mathbf{E}_{\mathfrak{F}_k}[\|\boldsymbol{g}_k^u-\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{F}_k}[\|\boldsymbol{g}_{k+1}^0-\boldsymbol{g}_k^0\|^2]\\
&-\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\|^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+\frac{1}{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{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]+n\sigma^2\eta_k^2\\
&-\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\|^2_{\eta_k(\beta_k-\frac{1}{2}-\eta_k\beta_k^2)\boldsymbol{K}},
\end{aligned}\tag{65}
$$

where the second equality holds due to (5); the second equality holds since (27) in Lemma 3, $\boldsymbol{x}_k$ and $\boldsymbol{v}_k$ are independent of $\mathfrak{F}_k$, and (50); the first inequality holds due to the Cauchy-Schwarz inequality, (27), $\rho(\boldsymbol{K})=1$, and the Jensen's inequality; and the last inequality holds due to (49), (52), and (58). For the third term in the right-hand side of (65), we have

$$
\begin{aligned}
&\omega_k\eta_k\alpha_k\boldsymbol{x}_k^\top\boldsymbol{L}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&=\omega_k\eta_k\alpha_k\boldsymbol{x}_k^\top\boldsymbol{L}\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big)\\
&\le\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\omega_k\eta_k\alpha_k\boldsymbol{L}^2}+\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\frac{1}{2}\omega_k\eta_k\alpha_k\boldsymbol{K}}.
\end{aligned}\tag{66}
$$

Then, from (64)–(66), we have (63). ■

Lemma 9. *Suppose Assumptions 2–5 hold. Then the following holds for Algorithm 1*

$$
\begin{aligned}
\mathbf{E}_{\mathfrak{F}_k}[W_{4,k+1}]\le&W_{4,k}-\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k\|^2+\|\boldsymbol{x}_k\|^2_{\frac{\eta_k}{2}L_f^2\boldsymbol{K}}\\
&-\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k^0\|^2+\frac{1}{2}\eta_k^2L_f\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2],
\end{aligned}\tag{67}
$$

where $W_{4,k}=n(f(\bar{x}_k)-f^*)=\tilde{f}(\bar{\boldsymbol{x}}_k)-\tilde{f}^*$.

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

From (49) and $\rho(\boldsymbol{H})=1$, we have that

$$
\begin{aligned}
\|\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k\|^2&=\|\boldsymbol{H}(\boldsymbol{g}_k^0-\boldsymbol{g}_k)\|^2\\
&\le\|\boldsymbol{g}_k^0-\boldsymbol{g}_k\|^2\le L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}.
\end{aligned}\tag{68}
$$

From (50), we have

$$
\mathbf{E}_{\mathfrak{F}_k}[\bar{\boldsymbol{g}}_k^u]=\mathbf{E}_{\mathfrak{F}_k}[\boldsymbol{H}\boldsymbol{g}_k^u]=\boldsymbol{H}\mathbf{E}_{\mathfrak{F}_k}[\boldsymbol{g}_k^u]=\bar{\boldsymbol{g}}_k.\tag{69}
$$

We have

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{4,k+1}]=\mathbf{E}_{\mathfrak{F}_k}[\tilde{f}(\bar{\boldsymbol{x}}_{k+1})-\tilde{f}^*]\\
&=\mathbf{E}_{\mathfrak{F}_k}[\tilde{f}(\bar{\boldsymbol{x}}_k)-\tilde{f}^*+\tilde{f}(\bar{\boldsymbol{x}}_{k+1})-\tilde{f}(\bar{\boldsymbol{x}}_k)]\\
&\le\mathbf{E}_{\mathfrak{F}_k}[\tilde{f}(\bar{\boldsymbol{x}}_k)-\tilde{f}^*-\eta_k(\bar{\boldsymbol{g}}_k^u)^\top\boldsymbol{g}_k^0+\frac{1}{2}\eta_k^2L_f\|\bar{\boldsymbol{g}}_k^u\|^2]\\
&=\tilde{f}(\bar{\boldsymbol{x}}_k)-\tilde{f}^*-\eta_k\bar{\boldsymbol{g}}_k^\top\boldsymbol{g}_k^0+\frac{1}{2}\eta_k^2L_f\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]\\
&=\tilde{f}(\bar{\boldsymbol{x}}_k)-\tilde{f}^*-\eta_k\bar{\boldsymbol{g}}_k^\top\bar{\boldsymbol{g}}_k^0+\frac{1}{2}\eta_k^2L_f\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]\\
&=W_{4,k}-\frac{\eta_k}{2}\bar{\boldsymbol{g}}_k^\top(\bar{\boldsymbol{g}}_k+\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k)\\
&\quad-\frac{\eta_k}{2}(\bar{\boldsymbol{g}}_k-\bar{\boldsymbol{g}}_k^0+\bar{\boldsymbol{g}}_k^0)^\top\bar{\boldsymbol{g}}_k^0+\frac{1}{2}\eta_k^2L_f\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]\\
&\le W_{4,k}-\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k\|^2+\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k\|^2-\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k^0\|^2\\
&\quad+\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k\|^2+\frac{1}{2}\eta_k^2L_f\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]\\
&=W_{4,k}-\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k\|^2+\frac{\eta_k}{2}\|\bar{\boldsymbol{g}}_k^0-\bar{\boldsymbol{g}}_k\|^2\\
&\quad-\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k^0\|^2+\frac{1}{2}\eta_k^2L_f\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2],
\end{aligned}\tag{70}
$$

where the first inequality holds since that $\tilde{f}$ is smooth, (26a) and (57); the third equality holds since $\boldsymbol{x}_k$ and $\boldsymbol{v}_k$ are independent of $\mathfrak{F}_k$ and (69); the fourth equality holds due to $\bar{\boldsymbol{g}}_k^\top\boldsymbol{g}_k^0=\boldsymbol{g}_k^\top\boldsymbol{H}\boldsymbol{g}_k^0=\boldsymbol{g}_k^\top\boldsymbol{H}\boldsymbol{H}\boldsymbol{g}_k^0=\bar{\boldsymbol{g}}_k^\top\bar{\boldsymbol{g}}_k^0$; and the second inequality holds due to the Cauchy-Schwarz inequality.

Then, from (68) and (70), we have (67). ■

B. *Proof of Theorem 1*

We denote the following notations

$$
\begin{aligned}
c_0(\kappa_1,\kappa_2)&=\max\{4\kappa_2\varepsilon_5,\ \varepsilon_6\},\\
c_1&=\frac{1}{\rho_2(L)}+1,\\
c_2(\kappa_1)&=\min\Big\{\frac{\varepsilon_1}{\varepsilon_2},\ \frac{1}{5}\Big\},\\
\kappa_3&=\frac{1}{\rho_2(L)}+\kappa_1+1,\\
\kappa_4&=\frac{1}{\rho_2(L)}+\kappa_1+\frac{3}{2},\\
\kappa_5&=\frac{\kappa_1+1}{2}+\frac{1}{2\rho_2(L)},\\
\kappa_6&=\min\Big\{\frac{1}{2\rho(L)},\ \frac{\kappa_1-1}{2\kappa_1}\Big\},\\
\kappa_7&=\frac{\varepsilon_3}{L_f^2},\\
\varepsilon_1&=(\kappa_1-1)\rho_2(L)-1,\\
\varepsilon_2&=\rho(L)+(2\kappa_1^2+1)\rho(L^2)+1,
\end{aligned}
$$

$$
\begin{aligned}
\varepsilon_3 =&\varepsilon_1\kappa_2-\varepsilon_2\kappa_2^2,\\
\varepsilon_4 =&\frac{1}{2}(\kappa_2-5\kappa_2^2),\\
\varepsilon_5 =&L_f+\frac{1}{\kappa_2\varepsilon_6}\kappa_3L_f^2+\frac{2}{\varepsilon_6^2}\kappa_4L_f^2,\\
\varepsilon_6 =&\max\Big\{\frac{1}{2}(2+3L_f^2),\ \kappa_3\Big\}.
\end{aligned}
$$

To prove Theorem 1, we need the following lemma.

Lemma 10. *Suppose Assumptions 1–6 hold. Suppose $\alpha_k=\alpha=\kappa_1\beta$, $\beta_k=\beta\ge c_0(\kappa_1,\kappa_2)$, and $\eta_k=\eta=\kappa_2/\beta$, where $\kappa_1>c_1$ and $\kappa_2\in(0,c_2(\kappa_1))$. Then, for any $k\in\mathbb{N}_0$ the following holds for Algorithm 1*

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\
&\le W_k-\|\boldsymbol{x}_k\|^2_{\varepsilon_3\boldsymbol{K}}-\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}_k^0\Big\|^2_{\varepsilon_4\boldsymbol{K}}\\
&\quad-\frac{1}{4}\eta\|\bar{\boldsymbol{g}}_k^0\|^2+(\varepsilon_5+3n)\sigma^2\eta^2, \qquad (71)\\
&\mathbf{E}_{\mathfrak{F}_k}[\breve{W}_{k+1}]\\
&\le \breve{W}_k-\|\boldsymbol{x}_k\|^2_{\varepsilon_3\boldsymbol{K}}-\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}_k^0\Big\|^2_{\varepsilon_4\boldsymbol{K}}+2\varepsilon_5\eta^2\|\bar{\boldsymbol{g}}_k^0\|^2\\
&\quad+2L_f^2\varepsilon_5\eta^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+(\varepsilon_5+3n)\sigma^2\eta^2, \qquad (72)\\
&\mathbf{E}_{\mathfrak{F}_k}[W_{4,k+1}]\\
&\le W_{4,k}-\frac{1}{4}\eta\|\bar{\boldsymbol{g}}_k^0\|^2+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\eta L_f^2\boldsymbol{K}}+L_f\sigma^2\eta^2, \qquad (73)\\
&\mathbf{E}_{\mathfrak{F}_k}[\tilde{W}_{k+1}]\\
&\le \tilde{W}_k-\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\varepsilon_3\boldsymbol{K}}-\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}_k^0\Big\|^2_{\varepsilon_4\boldsymbol{K}}-\frac{1}{4}\kappa_7\|\bar{\boldsymbol{g}}_k^0\|^2\\
&\quad+(\varepsilon_5+3n)\sigma^2\eta^2+\kappa_7L_f\sigma^2\eta, \qquad (74)
\end{aligned}
$$

where $W_k=\sum_{i=1}^4 W_{i,k}$, $\breve{W}_k=\sum_{i=1}^3 W_{i,k}$, and $\tilde{W}_k=W_k+\frac{\kappa_7}{\eta}W_{4,k}$.

Proof: (i) Noting that $\alpha_k=\alpha=\kappa_1\beta$, $\beta_k=\beta$, $\eta_k=\eta$, and $\omega_k=\frac{1}{\beta_k}-\frac{1}{\beta_{k+1}}=0$, from (48), (54), (63), and (67), we have

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\
&\le W_k+\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}_k^0\Big\|^2_{\frac{3}{2}\eta^2\beta^2\boldsymbol{K}}+2n\sigma^2\eta^2\\
&\quad-\|\boldsymbol{x}_k\|^2_{\eta\alpha\boldsymbol{L}-\frac{1}{2}\eta\boldsymbol{K}-\frac{3}{2}\eta^2\alpha^2\boldsymbol{L}^2-\frac{1}{2}\eta(1+5\eta)L_f^2\boldsymbol{K}}\\
&\quad+\|\boldsymbol{x}_k\|^2_{\eta^2\beta^2(\boldsymbol{L}+\kappa_1\boldsymbol{L}^2)}+\frac{1}{2}\eta\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}}\\
&\quad+\frac{\eta}{\beta^2}\Big(\eta+\frac{1}{2}\Big)\Big(\frac{1}{\rho_2(L)}+\kappa_1\Big)L_f^2\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]\\
&\quad+\|\boldsymbol{x}_k\|^2_{\eta(\beta\boldsymbol{L}+\frac{1}{2}\boldsymbol{K})+\eta^2(\frac{1}{2}\alpha^2-\alpha\beta+\beta^2)\boldsymbol{L}^2+\frac{1}{2}\eta(1+3\eta)L_f^2\boldsymbol{K}}\\
&\quad+\frac{\eta}{2\beta^2}(1+3\eta)L_f^2\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]+n\sigma^2\eta^2\\
&\quad-\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}_k^0\Big\|^2_{\eta(\beta-\frac{1}{2}-\eta\beta^2)\boldsymbol{K}}-\frac{1}{4}\eta\|\bar{\boldsymbol{g}}_k\|^2\\
&\quad+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\eta L_f^2\boldsymbol{K}}-\frac{1}{4}\eta\|\bar{\boldsymbol{g}}_k^0\|^2+\frac{1}{2}\eta^2L_f\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]. \qquad (75)
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u\|^2]=\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u-\bar{\boldsymbol{g}}_k+\bar{\boldsymbol{g}}_k\|^2]\\
&\le 2\mathbf{E}_{\mathfrak{F}_k}[\|\bar{\boldsymbol{g}}_k^u-\bar{\boldsymbol{g}}_k\|^2]+2\|\bar{\boldsymbol{g}}_k\|^2\\
&=2n\mathbf{E}_{\mathfrak{F}_k}\Big[\Big\|\frac{1}{n}\sum_{i=1}^n(g_{i,k}^u-g_{i,k})\Big\|^2\Big]+2\|\bar{\boldsymbol{g}}_k\|^2\\
&=\frac{2}{n}\mathbf{E}_{\mathfrak{F}_k}\Big[\Big\|\sum_{i=1}^n(g_{i,k}^u-g_{i,k})\Big\|^2\Big]+2\|\bar{\boldsymbol{g}}_k\|^2\\
&=\frac{2}{n}\sum_{i=1}^n\mathbf{E}_{\mathfrak{F}_k}[\|g_{i,k}^u-g_{i,k}\|^2]+2\|\bar{\boldsymbol{g}}_k\|^2\\
&\le 2\sigma^2+2\|\bar{\boldsymbol{g}}_k\|^2, \qquad (76)
\end{aligned}
$$

where the first inequality holds due to the Cauchy-Schwarz inequality; the last equality holds since $\{g_{i,k}^u,\ i\in[n]\}$ are independent of each other as assumed in Assumption 4, $\boldsymbol{x}_k$ and $\boldsymbol{v}_k$ are independent of $\mathfrak{F}_k$, and $\mathbf{E}_{\mathfrak{F}_k}[g_{i,k}^u]=g_{i,k}$ as assumed in Assumption 5; and the last inequality holds due to (51).

From (75), (76), and $\alpha=\kappa_1\beta$, we have

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\
&\le W_k-\|\boldsymbol{x}_k\|^2_{\eta\boldsymbol{M}_1-\eta^2\boldsymbol{M}_2}-\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}_k^0\Big\|^2_{b_{1,k}\boldsymbol{K}}\\
&\quad-b_{2,k}\eta\|\bar{\boldsymbol{g}}_k\|^2-\frac{1}{4}\eta\|\bar{\boldsymbol{g}}_k^0\|^2+b_{3,k}\sigma^2\eta^2+3n\sigma^2\eta^2, \qquad (77)
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{M}_1 =&(\alpha-\beta)\boldsymbol{L}-\frac{1}{2}(2+3L_f^2)\boldsymbol{K},\\
\boldsymbol{M}_2 =&\beta^2\boldsymbol{L}+(2\alpha^2+\beta^2)\boldsymbol{L}^2+4L_f^2\boldsymbol{K},\\
b_{1,k} =&\frac{1}{2}(2\beta-\kappa_3)\eta-\frac{5}{2}\beta^2\eta^2,\\
b_{2,k} =&\frac{1}{4}-b_{3,k}\eta,\\
b_{3,k} =&L_f+\frac{1}{\beta^2\eta}\kappa_3L_f^2+\frac{2}{\beta^2}\kappa_4L_f^2.
\end{aligned}
$$

From (28), $\alpha=\kappa_1\beta$, $\kappa_1>c_1>1$, $\eta=\kappa_2/\beta$, and $\beta\ge c_0(\kappa_1,\kappa_2)\ge\varepsilon_6\ge(2+3L_f^2)/2$, we have

$$\eta\boldsymbol{M}_1\ge\varepsilon_1\kappa_2\boldsymbol{K}. \qquad (78)$$

From (28), $\alpha=\kappa_1\beta$, and $\beta\ge\frac{1}{2}(2+3L_f^2)>2L_f$, we have

$$\eta^2\boldsymbol{M}_2\le\varepsilon_2\kappa_2^2\boldsymbol{K}. \qquad (79)$$

From $\beta\ge\kappa_3$, we have

$$b_{1,k}\ge\varepsilon_4. \qquad (80)$$

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

$$\varepsilon_1>0. \qquad (81)$$

From (81) and $\kappa_2\in(0,\min\{\frac{\varepsilon_1}{\varepsilon_2},\ \frac{1}{5}\})$, we have

$$
\begin{aligned}
\varepsilon_3>0, \qquad (82)\\
\varepsilon_4>0. \qquad (83)
\end{aligned}
$$

From (82), (83), and $\beta\ge4\kappa_2\varepsilon_5$, we have

$$b_{3,k}=L_f+\frac{1}{\beta^2\eta_k}\kappa_3L_f^2+\frac{2}{\beta^2}\kappa_4L_f^2\le\varepsilon_5, \qquad (84)$$

$$b_{2,k}=\frac{1}{4}-b_{3,k}\eta\ge\frac{1}{4}-\frac{\kappa_2}{\beta}\varepsilon_5\ge0. \qquad (85)$$

From (77)–(80), (84), and (85), we have (71).

(ii) Similar to the way to get (71), we have

$$\begin{aligned}&\mathbf{E}_{\mathfrak{F}_k}[\breve{W}_{k+1}]\\&\le \breve{W}_k-\|\boldsymbol{x}_k\|^2_{\varepsilon_3\boldsymbol{K}}-\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}^0_k\Big\|^2_{\varepsilon_4\boldsymbol{K}}\\&\quad+\varepsilon_5\eta^2\|\bar{\boldsymbol{g}}_k\|^2+(\varepsilon_5+3n)\sigma^2\eta^2,\end{aligned}\tag{86}$$

We have

$$\begin{aligned}&\|\bar{\boldsymbol{g}}_k\|^2=\|\bar{\boldsymbol{g}}_k-\bar{\boldsymbol{g}}^0_k+\bar{\boldsymbol{g}}^0_k\|^2\\&\le 2\|\bar{\boldsymbol{g}}_k-\bar{\boldsymbol{g}}^0_k\|^2+2\|\bar{\boldsymbol{g}}^0_k\|^2\le 2L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+2\|\bar{\boldsymbol{g}}^0_k\|^2,\end{aligned}\tag{87}$$

where the last inequality holds due to (68).

From (86) and (87), we have (72).

(iii) From (67) and (76), we have

$$\begin{aligned}&\mathbf{E}_{\mathfrak{F}_k}[W_{4,k+1}]\\&\le W_{4,k}-\frac{1}{4}\eta\|\bar{\boldsymbol{g}}_k\|^2+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}\eta L_f^2\boldsymbol{K}}\\&\quad-\frac{1}{4}\eta\|\bar{\boldsymbol{g}}^0_k\|^2+\eta^2L_f(\sigma^2+\|\bar{\boldsymbol{g}}_k\|^2),\end{aligned}\tag{88}$$

From $\eta=\kappa_2/\beta$ and $\beta\ge 4\kappa_2\varepsilon_5>4\kappa_2L_f$, we have

$$\eta L_f<\frac{1}{4}\tag{89}$$

From (88) and (89), we have (73).

(iv) From (71), (73), and $\kappa_7=\varepsilon_3/L_f^2$, we have (74). ■

Now we are ready to prove Theorem 1.

Denote

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

We have

$$\begin{aligned}&W_k\\&=\frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+\frac{1}{2}\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{Q}+\kappa_1\boldsymbol{K}}\\&\quad+\boldsymbol{x}_k^\top\boldsymbol{K}\Big(\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}^0_k\Big)+n(f(\bar{x}_k)-f^*)\\&\ge\frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+\frac{1}{2}\Big(\frac{1}{\rho(L)}+\kappa_1\Big)\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}\\&\quad-\frac{1}{2\kappa_1}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\frac{\kappa_1}{2}\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}+n(f(\bar{x}_k)-f^*)\\&\ge\kappa_6\Big(\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+\Big\|\boldsymbol{v}_k+\frac{1}{\beta}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}\Big)+n(f(\bar{x}_k)-f^*)\qquad(90)\\&\ge\kappa_6\hat{V}_k\ge 0,\qquad(91)\end{aligned}$$

where the first inequality holds due to (30) and the Cauchy-Schwarz inequality; and the last inequality holds due to $0<\kappa_6<0.5$. Similarly, we have

$$W_k\le\kappa_5\hat{V}_k.\tag{92}$$

From (71) and (83), we have

$$\begin{aligned}&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\&\le W_k-\varepsilon_3\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\frac{\kappa_2}{4\beta}\|\bar{\boldsymbol{g}}^0_k\|^2+\frac{(\varepsilon_5+3n)\kappa_2^2\sigma^2}{\beta^2}.\end{aligned}\tag{93}$$

Then, taking expectation in $\mathcal{F}_T$ and summing (93) over $k\in[0,T]$ yield

$$\begin{aligned}&\mathbf{E}[W_{T+1}]+\sum_{k=0}^{T}\mathbf{E}\Big[\varepsilon_3\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+\frac{\kappa_2}{4\beta}\|\bar{\boldsymbol{g}}^0_k\|^2\Big]\\&\le W_0+\frac{(T+1)(\varepsilon_5+3n)\kappa_2^2\sigma^2}{\beta^2}.\end{aligned}\tag{94}$$

From (94), (91), and (82), we have

$$\begin{aligned}&\frac{1}{T+1}\sum_{k=0}^{T}\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k}-\bar{x}_k\|^2\Big]\\&=\frac{1}{n(T+1)}\sum_{k=0}^{T}\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}]\\&\le\frac{W_0}{n\varepsilon_3(T+1)}+\frac{(\varepsilon_5+3n)\kappa_2^2\sigma^2}{n\varepsilon_3\beta^2}.\end{aligned}\tag{95}$$

Noting that $W_0=\mathcal{O}(n)$, from (95), we have (11).

Taking expectation in $\mathcal{F}_T$ and summing (74) over $k\in[0,T]$ yield

$$\begin{aligned}&\frac{1}{4}n\kappa_7\sum_{k=0}^{T}\mathbf{E}[\|\nabla f(\bar{x}_k)\|^2]=\frac{1}{4}\kappa_7\sum_{k=0}^{T}\mathbf{E}[\|\bar{\boldsymbol{g}}^0_k\|^2]\\&\le\tilde{W}_0+(T+1)(\varepsilon_5+3n)\sigma^2\eta^2+(T+1)\kappa_7L_f\sigma^2\eta.\end{aligned}\tag{96}$$

Noting that $\tilde{W}_0=\mathcal{O}(n/\eta)$ and $\eta=\kappa_2/\beta$, from (96), we have (12).

Taking expectation in $\mathcal{F}_T$ and summing (73) over $k\in[0,T]$ yield

$$\begin{aligned}&n(\mathbf{E}[f(\bar{x}_{T+1})]-f^*)=\mathbf{E}[W_{4,T+1}]\\&\le W_{4,0}+\frac{1}{2}\eta L_f^2\sum_{k=0}^{T}\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}]+L_f\sigma^2\eta^2(T+1).\end{aligned}\tag{97}$$

Noting that $W_{4,0}=\mathcal{O}(n)$ and $\eta=\kappa_2/\beta$, from (94) and (97), we have (13).

C. Proof of Theorem 2

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

$$\varepsilon_7=\frac{1}{\kappa_5}\min\Big\{\varepsilon_3,\ \varepsilon_4,\ \frac{\nu\kappa_2}{2(T+1)^\theta}\Big\}.$$

From the conditions in Theorem 2, we know that all conditions needed in Lemma 10 are satisfied, so (71)–(73) still hold.

From Assumptions 2 and 7 as well as (17), we have that

$$\|\bar{\boldsymbol{g}}^0_k\|^2=n\|\nabla f(\bar{x}_k)\|^2\ge 2\nu n(f(\bar{x}_k)-f^*)=2\nu W_{4,k}.\tag{98}$$

From (91), we have

$$\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+W_{4,k}\le\hat{V}_k\le\frac{W_k}{\kappa_6}.\tag{99}$$

From (71), (98), (91), (92), and (18), we have

$$\begin{aligned}&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\&\le W_k-\varepsilon_3\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\varepsilon_4\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}\eta\nu W_{4,k}+(\varepsilon_5+3n)\sigma^2\eta^2\\
&\leq W_k-\frac{1}{\kappa_5}\min\Big\{\varepsilon_3,\ \varepsilon_4,\ \frac{\nu\eta}{2}\Big\}W_k+(\varepsilon_5+3n)\sigma^2\eta^2.
\end{aligned}
\tag{100}
$$

From (100) and (18), we have

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\
&\leq W_k-\varepsilon_7W_k+\frac{\kappa_2^2(\varepsilon_5+3n)\sigma^2}{(T+1)^{2\theta}},\ \forall k\leq T.
\end{aligned}
\tag{101}
$$

From $\kappa_1>1$, we have $\kappa_5>1$. From $0<\kappa_2<1/5$, we have $\varepsilon_4=(\kappa_2-5\kappa_2^2)/2\leq 1/40$. Thus,

$$
0<\varepsilon_7\leq\frac{\varepsilon_4}{\kappa_5}\leq\frac{1}{40}.
\tag{102}
$$

Then, from (101), (91), and (102), we have

$$
\begin{aligned}
&\mathbf{E}[W_{k+1}]\\
&\leq(1-\varepsilon_7)^{k+1}W_0+\frac{\kappa_2^2(\varepsilon_5+3n)\sigma^2}{(T+1)^{2\theta}}\sum_{l=0}^{k}(1-\varepsilon_7)^l\\
&\leq W_0e^{-\varepsilon_7(k+1)}+\frac{\kappa_2^2(\varepsilon_5+3n)\sigma^2}{\varepsilon_7(T+1)^{2\theta}},\ \forall k\leq T.
\end{aligned}
\tag{103}
$$

Then, noting that $\varepsilon_7=\mathcal{O}(1/(T+1)^\theta)$ and $\theta\in(0,1)$, from (103), (31), and (99), we have

$$
\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+W_{4,k}]=\mathcal{O}(\frac{n}{T^\theta}),\ \forall k\leq T.
\tag{104}
$$

Thus, there exists a constant $c_f>0$ such that

$$
\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+W_{4,k}]\leq nc_f,\ \forall k\leq T.
\tag{105}
$$

From (90) and (92), we have

$$
0\leq 2\kappa_6(W_{1,k}+W_{2,k})\leq\breve{W}_k\leq 2\kappa_5(W_{1,k}+W_{2,k}).
\tag{106}
$$

From (26b), we have

$$
\|\bar{\boldsymbol{g}}^0_k\|^2=n\|\nabla f(\bar{x}_k)\|^2\leq 2L_fn(f(\bar{x}_k)-f^*)=2W_{4,k}
\tag{107}
$$

Denote $\breve{z}_k=\mathbf{E}[\breve{W}_k]$. From (72) and (105)–(107), we have

$$
\breve{z}_{k+1}\leq(1-a_1)\breve{z}_k+a_2\eta^2,\ \forall k\leq T,
\tag{108}
$$

where $a_1=\min\{\varepsilon_3,\ \varepsilon_4\}/\kappa_5$ and $a_2=n(4\varepsilon_5L_fc_f+2\varepsilon_5L_f^2c_f+(\varepsilon_5+3)\sigma^2)$.

From (102), we have

$$
a_1\leq\frac{\varepsilon_4}{\kappa_5}\leq\frac{1}{40}.
\tag{109}
$$

From (108) and (109), we have

$$
\breve{z}_{k+1}\leq(1-a_1)^{k+1}\breve{z}_0+\frac{a_2\eta^2}{a_1},\ \forall k\leq T,
$$

which yields (19a).

From (73) and (98), we have

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{4,k+1}]\\
&\leq\Big(1-\frac{\nu\eta}{2}\Big)W_{4,k}+\frac{1}{2}\eta L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+L_f\sigma^2\eta^2\\
&\leq\Big(1-\frac{\nu\eta}{2}\Big)^{k+1}W_{4,0}+\frac{1}{\nu}(L_f^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+2L_f\sigma^2\eta).
\end{aligned}
\tag{110}
$$

Noting $\eta=\kappa_2/(T+1)^\theta$, from (110), (31), and (19a), we have (19b).

D. Proof of Theorem 3

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

$$
\begin{aligned}
\hat{c}_0(\kappa_1,\kappa_2)=&\frac{\kappa_2}{8\kappa_5},\\
\tilde{c}_0(\kappa_1,\kappa_2)=&\max\Big\{4\varepsilon_{11},\ \varepsilon_6,\ \frac{\varepsilon_{10}}{\varepsilon_4}\Big\},\\
\hat{c}_2(\kappa_1)=&\min\Big\{\frac{\varepsilon_1}{\varepsilon_2},\ \frac{\varepsilon_8}{\varepsilon_9},\ \frac{1}{5}\Big\},\\
\hat{c}_3(\kappa_0,\kappa_1,\kappa_2)=&\max\Big\{\frac{\tilde{c}_0(\kappa_1,\kappa_2)}{\kappa_0},\ \frac{\kappa_5}{\varepsilon_3},\ \frac{2\kappa_5}{\varepsilon_4},\ \frac{8L_f\kappa_3}{\nu\kappa_2}\\
&\frac{16L_f(\kappa_3-1)}{\nu\kappa_0\kappa_2}\Big\},\\
\tilde{\sigma}^2=&2L_ff^*-2L_f\frac{1}{n}\sum_{i=1}^{n}f_i^*,\\
\varepsilon_8=&\kappa_1\rho_2(L)-1,\\
\varepsilon_9=&\frac{1}{2}(3\kappa_1+2)\kappa_1\rho(L^2)+\rho(L)+1,\\
\varepsilon_{10}=&\kappa_1+\kappa_3-1+\kappa_1\kappa_2+3\kappa_2^2+\kappa_2(\kappa_3-1),\\
\varepsilon_{11}=&\kappa_2L_f+(2\kappa_3-1+\kappa_2(10\kappa_3-4))L_f^2,\\
\varepsilon_{12}=&3+L_f+\frac{\kappa_3L_f^2}{\kappa_0\kappa_2t_1}+\frac{2\kappa_4L_f^2}{\kappa_0^2t_1^2}+\frac{2+2\kappa_3L_f^2}{\kappa_0t_1^2}\\
&+\frac{(\kappa_3-1)L_f^2}{\kappa_0^2\kappa_2t_1^3}+\frac{(\kappa_3-1)L_f^2}{\kappa_0^2t_1^4}\Big(\frac{2}{\kappa_0}+2\Big),\\
\varepsilon_{13}=&\frac{\kappa_0\kappa_3}{\kappa_2^2}+\frac{\kappa_3-1}{\kappa_2^2t_1^2},\\
\varepsilon_{14}=&\varepsilon_{12}\sigma^2+\varepsilon_{13}\tilde{\sigma}^2,\\
\varepsilon_{15}=&\frac{1}{\kappa_5}\min\Big\{\frac{\varepsilon_3\kappa_0t_1}{\kappa_2},\ \frac{\varepsilon_4\kappa_0t_1}{2\kappa_2},\ \frac{\nu}{8}\Big\}.
\end{aligned}
$$

To prove Theorem 3, we need the following lemma.

Lemma 11. *Suppose Assumptions 1–6 hold. Suppose $\alpha_k=\kappa_1\beta_k$, $\beta_k=\kappa_0(k+t_1)$, and $\eta_k=\kappa_2/\beta_k$, where $\kappa_0\geq\tilde{c}_0(\kappa_1,\kappa_2)/t_1$, $\kappa_1>c_1$, $\kappa_2\in(0,\hat{c}_2(\kappa_1))$, and $t_1\geq 1$. Then, for any $k\in\mathbb{N}_0$ the following holds for Algorithm 1*

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\
&\leq W_k-\varepsilon_3\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\frac{1}{2}\varepsilon_4\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}-\frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}^0_k\|^2\\
&\quad+2L_fb_{8,k}\eta_k^2W_{4,k}+n\varepsilon_{14}\eta_k^2,
\end{aligned}
\tag{111}
$$

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[\breve{W}_{k+1}]\\
&\leq\breve{W}_k-\varepsilon_3\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\frac{1}{2}\varepsilon_4\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}+n\varepsilon_{14}\eta_k^2\\
&\quad+2\varepsilon_{12}L_f^2\eta_k^2\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+2(2\varepsilon_{12}+\varepsilon_{13})L_f\eta_k^2W_{4,k},
\end{aligned}
\tag{112}
$$

$$
\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{4,k+1}]\\
&\leq W_{4,k}-\frac{\eta_k}{4}\|\bar{\boldsymbol{g}}^0_k\|^2+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}L_f^2\eta_k\boldsymbol{K}}+\eta_k^2L_f\sigma^2,
\end{aligned}
\tag{113}
$$

where $b_{8,k}=\kappa_3\frac{\omega_k}{\eta_k^2}+(\kappa_3-1)\frac{\omega_k^2}{\eta_k^2}$.

Proof: (i) We have

$$\|\boldsymbol{g}_k^0\|^2 = \sum_{i=1}^n \|\nabla f_i(\bar{x}_k)\|^2 \le \sum_{i=1}^n 2L_f(f_i(\bar{x}_k) - f_i^*) = 2L_f n(f(\bar{x}_k) - f^*) + n\tilde{\sigma}^2, \quad (114)$$

where the inequality holds due to (26b).

From the Cauchy-Schwarz inequality, (58), and (114), we have

$$\|\boldsymbol{g}_{k+1}^0\|^2 = \|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0 + \boldsymbol{g}_k^0\|^2 \le 2(\|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0\|^2 + \|\boldsymbol{g}_k^0\|^2) \le 2(\eta_k^2 L_f^2 \|\bar{\boldsymbol{g}}_k^u\|^2 + 2L_f W_{4,k} + n\tilde{\sigma}^2). \quad (115)$$

From (48), (54), (63), (67), (76), (115), $\alpha_k = \kappa_1\beta_k$, and $\eta_k = \kappa_2/\beta_k$, we have

$$\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}] \\
&\le W_k - \|\boldsymbol{x}_k\|^2_{\eta_k \boldsymbol{M}_{3,k} - \eta_k^2 \boldsymbol{M}_{4,k} - \frac{1}{2}\kappa_1\kappa_2\omega_k + \eta_k\omega_k \boldsymbol{M}_{5,k} - \eta_k^2\omega_k \boldsymbol{M}_{6,k}} \\
&\quad - \Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{b_{4,k}^0 \boldsymbol{K}} - \eta_k b_{5,k}\|\bar{\boldsymbol{g}}_k\|^2 - \frac{1}{4}\eta_k\|\bar{\boldsymbol{g}}_k^0\|^2 \\
&\quad + \eta_k^2(b_{6,k} + b_{7,k}n)\sigma^2 + \eta_k^2 b_{8,k}(2L_f W_{4,k} + n\tilde{\sigma}^2), \quad (116)
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{M}_{3,k} &= (\alpha_k - \beta_k)\boldsymbol{L} - \frac{1}{2}(2 + 3L_f^2)\boldsymbol{K}, \\
\boldsymbol{M}_{4,k} &= \beta_k^2\boldsymbol{L} + (2\alpha_k^2 + \beta_k^2)\boldsymbol{L}^2 + 4L_f^2\boldsymbol{K}, \\
\boldsymbol{M}_{5,k} &= \alpha_k\boldsymbol{L} - \frac{1}{2}(1 + L_f^2)\boldsymbol{K}, \\
\boldsymbol{M}_{6,k} &= \frac{3}{2}\alpha_k^2\boldsymbol{L}^2 + \beta_k^2(\boldsymbol{L} + \kappa_1\boldsymbol{L}^2) + \frac{5}{2}L_f^2\boldsymbol{K}, \\
b_{4,k}^0 &= \frac{1}{2}\eta_k(2\beta_k - \kappa_3) - \frac{5}{2}\eta_k^2\beta_k^2 - \frac{1}{2}\omega_k\eta_k(\kappa_3 - 1) \\
&\quad - \frac{1}{2}\omega_k(\eta_k\alpha_k + \kappa_3 - 1 + 3\eta_k^2\beta_k^2), \\
b_{5,k} &= \frac{1}{4} - \eta_k b_{6,k}, \\
b_{6,k} &= L_f + \frac{1}{\beta_k^2\eta_k}\kappa_3 L_f^2 + \frac{2}{\beta_k^2}\kappa_4 L_f^2 + 2\kappa_3 L_f^2\omega_k \\
&\quad + \omega_k\Big(\frac{1}{\beta_k^2\eta_k} + \frac{2}{\beta_k^2} + 2\omega_k\Big)(\kappa_3 - 1)L_f^2, \\
b_{7,k} &= 3 + 2\omega_k.
\end{aligned}$$

From (28), $\alpha_k = \kappa_1\beta_k$, $\kappa_1 > 1$, $\beta_k \ge \kappa_0 t_1 \ge \tilde{c}_0(\kappa_1, \kappa_2) \ge \varepsilon_6 \ge (2 + 3L_f^2)/2$, and $\eta_k = \kappa_2/\beta_k$, we have

$$\eta_k \boldsymbol{M}_{3,k} \ge \varepsilon_1\kappa_2\boldsymbol{K}. \quad (117)$$

From (28), $\alpha_k = \kappa_1\beta_k$, $\beta_k \ge (2 + 3L_f^2)/2 > 2L_f$, and $\eta_k = \kappa_2/\beta_k$, we have

$$\eta_k^2 \boldsymbol{M}_{4,k} \le \varepsilon_2\kappa_2^2\boldsymbol{K}. \quad (118)$$

From (28), $\alpha_k = \kappa_1\beta_k$, $\beta_k \ge (2 + 3L_f^2)/2 > (1 + L_f^2)/2$, and $\eta_k = \kappa_2/\beta_k$, we have

$$\eta_k \boldsymbol{M}_{5,k} \ge \varepsilon_8\kappa_2\boldsymbol{K}. \quad (119)$$

From (28), $\alpha_k = \kappa_1\beta_k$, $\beta_k > 2L_f > \sqrt{10}L_f/2$, and $\eta_k = \kappa_2/\beta_k$, we have

$$\eta_k^2 \boldsymbol{M}_{6,k} \le \varepsilon_9\kappa_2^2\boldsymbol{K}. \quad (120)$$

From $\alpha_k = \kappa_1\beta_k$, $\beta_k \ge \kappa_3$, and $\eta_k = \kappa_2/\beta_k$, we have

$$b_{4,k}^0 \ge b_{4,k}, \quad (121)$$

where $b_{4,k} = \varepsilon_4 - \frac{1}{2}\omega_k(\kappa_1 + \kappa_3 - 1 + \kappa_1\kappa_2 + 3\kappa_2^2) - \frac{1}{2}\omega_k\eta_k(\kappa_3 - 1)$.

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

$$\varepsilon_1 > 0 \text{ and } \varepsilon_8 > 0. \quad (122)$$

From (122) and $\kappa_2 \in (0, \min\{\frac{\varepsilon_1}{\varepsilon_2}, \frac{\varepsilon_8}{\varepsilon_9}, \frac{1}{5}\})$, we have

$$\varepsilon_3 > 0, \quad (123)$$

$$\varepsilon_8\kappa_2 - \varepsilon_9\kappa_2^2 > 0, \quad (124)$$

$$\varepsilon_4 > 0. \quad (125)$$

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

$$\begin{aligned}
\omega_k &= \frac{1}{\beta_k} - \frac{1}{\beta_{k+1}} = \frac{1}{\kappa_0}\Big(\frac{1}{k + t_1} - \frac{1}{k + t_1 + 1}\Big) \\
&= \frac{1}{\kappa_0(k + t_1)(k + t_1 + 1)} \le \frac{\kappa_0}{\beta_k^2}. \quad (126)
\end{aligned}$$

From (123)–(126), and $\kappa_0 \ge \max\{\frac{4\varepsilon_{11}}{t_1}, \frac{\varepsilon_{10}}{\varepsilon_4 t_1}\}$, we have

$$b_{4,k} \ge \varepsilon_4 - \frac{\varepsilon_{10}}{2\kappa_0 t_1^2} \ge \varepsilon_4 - \frac{\varepsilon_{10}}{2\kappa_0 t_1} \ge \frac{1}{2}\varepsilon_4 > 0, \quad (127)$$

$$b_{5,k} \ge \frac{1}{4} - \frac{\varepsilon_{11}}{\kappa_0 t_1} \ge 0. \quad (128)$$

From (126) and $\beta_k = \kappa_0(k + t_1) \ge \kappa_0 t_1$, we have

$$b_{6,k} + b_{7,k} \le \varepsilon_{12}, \quad (129)$$

$$b_{8,k} \le \varepsilon_{13}, \quad (130)$$

From (116)–(121), (123)–(125), and (127)–(130), we have (111).

(ii) Similarly to the way to get (111), we have

$$\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[\breve{W}_{k+1}] \\
&\le \breve{W}_k - \varepsilon_3\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \frac{1}{2}\varepsilon_4\Big\|\boldsymbol{v}_k + \frac{1}{\beta_k}\boldsymbol{g}_k^0\Big\|^2_{\boldsymbol{K}} + \varepsilon_{12}\eta_k^2\|\bar{\boldsymbol{g}}_k\|^2 \\
&\quad + 2L_f\varepsilon_{13}\eta_k^2 W_{4,k} + n\varepsilon_{14}\eta_k^2, \ \forall k \in \mathbb{N}_0, \quad (131)
\end{aligned}$$

From (131), (87), and (107), we have (112).

(iii) From (67), (76), and (107), we have

$$\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{4,k+1}] \\
&\le W_{4,k} - \frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k\|^2 + \|\boldsymbol{x}_k\|^2_{\frac{1}{2}L_f^2\eta_k\boldsymbol{K}} \\
&\quad - \frac{\eta_k}{4}\|\bar{\boldsymbol{g}}_k^0\|^2 + \eta_k^2 L_f(\sigma^2 + \|\bar{\boldsymbol{g}}_k\|^2). \quad (132)
\end{aligned}$$

From $0 < \eta_k \le \kappa_2/(\kappa_0 t_1)$ and $\kappa_0 t_1 \ge \tilde{c}_0(\kappa_1, \kappa_2) \ge 4\varepsilon_{11} > 4\kappa_2 L_f$, we have

$$\eta_k^2 L_f < \frac{1}{4}\eta_k. \quad (133)$$

From (132) and (133), we have (113). ■

Now we are ready to prove Theorem 3.

From $t_1 > \hat{c}_3(\kappa_0, \kappa_1, \kappa_2) \ge \tilde{c}_0(\kappa_1, \kappa_2)/\kappa_0$, we have

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

Thus, all conditions needed in Lemma 11 are satisfied, so (111)–(113) hold.

From (111) and (98), we have

$$\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\
&\le W_k-\varepsilon_3\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\frac{1}{2}\varepsilon_4\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}-\frac{\eta_k\nu}{2}W_{4,k}\\
&\quad+2L_fb_{8,k}\eta_k^2W_{4,k}+n\varepsilon_{14}\eta_k^2+n\varepsilon_{14}\eta_k^2\\
&=W_k-\varepsilon_3\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}-\frac{1}{2}\varepsilon_4\Big\|\boldsymbol{v}_k+\frac{1}{\beta_k}\boldsymbol{g}^0_k\Big\|^2_{\boldsymbol{K}}\\
&\quad-2\Big(\frac{1}{4}-\frac{1}{\nu}L_fb_{8,k}\eta_k\Big)\nu\eta_kW_{4,k},\ \forall k\in\mathbb{N}_0. \qquad (134)
\end{aligned}$$

From $t_1\ge 8L_f\kappa_3/(\nu\kappa_2)$, we have

$$\frac{1}{4}-\frac{L_f\kappa_3}{\nu\kappa_2t_1}\ge\frac{1}{8}. \qquad (135)$$

From (126), (135), and $\kappa_0\ge\frac{16L_f(\kappa_3-1)}{\nu\kappa_2t_1}$, we have

$$\begin{aligned}
&\frac{1}{4}-\frac{1}{\nu}L_fb_{8,k}\eta_k\ge\frac{1}{4}-\frac{L_f\kappa_0\kappa_3}{\nu\kappa_2\beta_k}-\frac{L_f\kappa_0^2(\kappa_3-1)}{\nu\kappa_2\beta_k^3}\\
&\ge\frac{1}{4}-\frac{L_f\kappa_3}{\nu\kappa_2t_1}-\frac{L_f(\kappa_3-1)}{\nu\kappa_2\kappa_0t_1^3}\ge\frac{1}{8}-\frac{L_f(\kappa_3-1)}{\nu\kappa_2\kappa_0t_1}\ge\frac{1}{16}. \qquad (136)
\end{aligned}$$

From (134), (91), and (92), we have

$$\begin{aligned}
&\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}]\\
&\le W_k-\frac{\eta_k}{\kappa_5}\min\Big\{\frac{\varepsilon_3}{\eta_k},\ \frac{\varepsilon_4}{2\eta_k},\ \frac{\nu}{8}\Big\}W_k+n\varepsilon_{14}\eta_k^2\\
&\le W_k-\varepsilon_{15}\eta_kW_k+n\varepsilon_{14}\eta_k^2,\ \forall k\in\mathbb{N}_0. \qquad (137)
\end{aligned}$$

Denote $z_k=\mathbf{E}[W_k]$, $r_{1,k}=\varepsilon_{15}\eta_k$, and $r_{2,k}=n\varepsilon_{14}\eta_k^2$, then from (137), we have

$$z_{k+1}\le(1-r_{1,k})z_k+r_{2,k},\ \forall k\in\mathbb{N}_0. \qquad (138)$$

From (20), we have

$$r_{1,k}=\eta_k\varepsilon_{15}=\frac{a_3}{k+t_1}, \qquad (139)$$

$$r_{2,k}=\eta_k^2\varepsilon_{14}n\sigma^2=\frac{a_4}{(k+t_1)^2}, \qquad (140)$$

where $a_3=\kappa_2\varepsilon_{15}/\kappa_0$ and $a_4=n\kappa_2^2\varepsilon_{14}/\kappa_0^2$.

From (102), we have

$$r_{1,k}\le\frac{\varepsilon_4}{2\kappa_5}\le\frac{1}{80}. \qquad (141)$$

From rom $t_1>\hat{c}_3(\kappa_0,\kappa_1,\kappa_2)\ge\kappa_5/\varepsilon_3$, we have

$$\frac{\varepsilon_3t_1}{\kappa_5}>1. \qquad (142)$$

From $t_1>\hat{c}_3(\kappa_0,\kappa_1,\kappa_2)\ge 2\kappa_5/\varepsilon_4$, we have

$$\frac{\varepsilon_4t_1}{2\kappa_5}>1. \qquad (143)$$

From $\kappa_0\in(0,\nu\hat{c}_0(\kappa_1,\kappa_2))$, $\hat{c}_0(\kappa_1,\kappa_2)=\kappa_2/(8\kappa_5)$, we have

$$\frac{\nu\kappa_2}{8\kappa_5\kappa_0}>1. \qquad (144)$$

Hence, from (142)–(144), we have

$$a_3=\frac{\kappa_2\varepsilon_{15}}{\kappa_0}>1. \qquad (145)$$

Then, from (138)–(141), (145), and (34), we have

$$z_k\le\phi_1(k,t_1,a_3,a_4,z_0),\ \forall k\in\mathbb{N}_+, \qquad (146)$$

where the function ϕ_1 is defined in (35).

Noting that $\phi_1(k,t_1,a_3,a_4,z_0)=\mathcal{O}(n/k)$, from (146) and (99), we get

$$\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+W_{4,k}]=\mathcal{O}(\frac{n}{k}),\ \forall k\in\mathbb{N}_+. \qquad (147)$$

Thus, there exists a constant $c_f>0$ such that

$$\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}+W_{4,k}]\le nc_f. \qquad (148)$$

From (112), (148), (106), and (20), we have

$$\breve{z}_{k+1}\le(1-a_5)\breve{z}_k+\frac{a_6}{(t+t_1)^2}, \qquad (149)$$

where $a_5=\min\{\varepsilon_3,\ \varepsilon_4/2\}/\kappa_5$ and $a_6=n(2\varepsilon_{12}L_f^2c_f+2(2\varepsilon_{12}+\varepsilon_{13})L_fc_f+\varepsilon_{14})\kappa_2^2/\kappa_0^2$.

From (102), we have

$$a_5\le\frac{\varepsilon_4}{2\kappa_5}\le\frac{1}{80}. \qquad (150)$$

From (123) and (125), we know that

$$a_5>0\text{ and }a_6>0. \qquad (151)$$

From (149)–(151) and (36), we have

$$\breve{z}_k\le\phi_2(k,t_1,a_5,a_6,\breve{z}_0)=\mathcal{O}(\frac{n}{k^2}),\ \forall k\in\mathbb{N}_+, \qquad (152)$$

where the function ϕ_2 is defined in (37).

From (37), (106), and (152), we have

$$\mathbf{E}[\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}]\le\frac{1}{\kappa_6}\breve{z}_k\le\frac{1}{\kappa_6}\phi_2(k,t_1,a_5,a_6,\breve{z}_0)=\mathcal{O}(\frac{n}{k^2}). \qquad (153)$$

From (153), we have (21a).

From (113) and (98), we have

$$\begin{aligned}
&\mathbf{E}[W_{4,k+1}]\\
&\le(1-\frac{\nu}{2}\eta_k)\mathbf{E}[W_{4,k}]+\|\boldsymbol{x}_k\|^2_{\frac{1}{2}L_f^2\eta_k\boldsymbol{K}}+L_f\sigma^2\eta_k^2. \qquad (154)
\end{aligned}$$

From (144) and $\kappa_5>1$, we have

$$\frac{\nu\kappa_2}{2\kappa_0}>\frac{\nu\kappa_2}{2\kappa_5\kappa_0}>4. \qquad (155)$$

Similar to the way to prove (34), from (153), (154), and (155), we have (21b).

E. Proof of Theorem 4

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

$$\begin{aligned}
\varepsilon&=\frac{1}{\kappa_5}\min\Big\{\frac{\varepsilon_3}{\eta},\ \frac{\varepsilon_4}{\eta},\ \frac{\nu}{2}\Big\},\\
c_1&=\frac{W_0}{n\kappa_6},\\
c_2&=\frac{\varepsilon_5+3n}{n\varepsilon\kappa_6}.
\end{aligned}$$

From the conditions in Theorem 4, we know that (100) holds. Thus,

$$\mathbf{E}_{\mathfrak{F}_k}[W_{k+1}] \leq W_k - \eta\varepsilon W_k + (\varepsilon_5 + 3n)\sigma^2\eta^2. \tag{156}$$

Similar to the way to get (102), we have

$$0 < \eta\varepsilon < 1. \tag{157}$$

From (156) and (157), we have

$$\begin{aligned}
&\mathbf{E}[W_{k+1}] \\
&\leq (1-\eta\varepsilon)\mathbf{E}[W_k] + (\varepsilon_5 + 3n)\sigma^2\eta^2 \\
&\leq (1-\eta\varepsilon)^{k+1}W_0 + (\varepsilon_5 + 3n)\sigma^2\eta^2 \sum_{\tau=0}^{k}(1-\eta\varepsilon)^{\tau} \\
&\leq (1-\eta\varepsilon)^{k+1}W_0 + \frac{\eta(\varepsilon_5 + 3n)\sigma^2}{\varepsilon}.
\end{aligned} \tag{158}$$

Hence, (158) and (99) give (23).