Distributed stochastic inertial-accelerated methods with delayed derivatives for nonconvex problems

Yangyang Xu*, Yibo Xu*, Yonggui Yan*, and Jie Chen†

Abstract.
Stochastic gradient methods (SGMs) are predominant approaches for solving stochastic optimization. On smooth nonconvex problems, a few acceleration techniques have been applied to improve the convergence rate of SGMs. However, little exploration has been made on applying a certain acceleration technique to a stochastic subgradient method (SsGM) for nonsmooth nonconvex problems. In addition, few efforts have been made to analyze an (accelerated) SsGM with delayed derivatives. The information delay naturally happens in a distributed system, where computing workers do not coordinate with each other.

In this paper, we propose an inertial proximal SsGM for solving nonsmooth nonconvex stochastic optimization problems. The proposed method can have guaranteed convergence even with delayed derivative information in a distributed environment. Convergence rate results are established to three classes of nonconvex problems: weakly-convex nonsmooth problems with a convex regularizer, composite nonconvex problems with a nonsmooth convex regularizer, and smooth nonconvex problems. For each problem class, the convergence rate is $O(1/K^{\frac{1}{2}})$ in the expected value of the gradient norm square, for K iterations. In a distributed environment, the convergence rate of the proposed method will be slowed down by the information delay. Nevertheless, the slow-down effect will decay with the number of iterations for the latter two problem classes. We test the proposed method on three applications. The numerical results clearly demonstrate the advantages of using the inertial-based acceleration. Furthermore, we observe higher parallelization speed-up in asynchronous updates over the synchronous counterpart, though the former uses delayed derivatives. Our source code is released at https://github.com/RPI-OPT/Inertial-SsGM

Key words. stochastic (sub)gradient method, inertial acceleration, distributed parallelization, delayed (sub)gradient

AMS subject classifications. 90C15, 65Y05, 68W15, 65K05

1. Introduction.

The stochastic approximation method is one popular approach for solving stochastic problems. It can date back to [52] for solving root-finding problems. Nowadays, its first-order versions, such as the stochastic gradient method (SGM), have been extensively used to solve stochastic problems or deterministic problems that involve a huge amount of data (e.g., see [42,56]). A standard (or vanilla) SGM often converges slowly. Several acceleration techniques have been used to improve its theoretical and/or empirical convergence speed (e.g., [3, 15, 24, 62, 65]) for solving convex or smooth nonconvex problems. However, *for nonsmooth nonconvex problems, it appears that it is still unknown whether a proximal SGM or a stochastic subgradient method (SsGM) can still have guaranteed convergence if a certain acceleration technique is applied.* In this paper, we give a positive answer to this open question by using an inertial-type acceleration technique, even if the derivative information can be delayed in a distributed environment.

Our study focuses on stochastic optimization problems in the form of

$$\phi^* = \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\ \phi(\mathbf{x}) := F(\mathbf{x}) + r(\mathbf{x}),\ \text{with } F(\mathbf{x}) := \mathbb{E}_\xi[f(\mathbf{x};\xi)]. \tag{1.1}$$

Here, ξ is a random variable that can represent a stochastic scenario or a data point, F is often called a loss function or a data-fitting term, and r can include a hard constraint and/or a soft regularization term. We will study a few problem classes, where F is nonconvex and can be

*Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY (xuy21@rpi.edu)
†MIT-IBM Watson AI Lab, IBM Research

smooth or nonsmooth but r is convex and nondifferentiable if it exists. As a special case, when ξ is distributed on a finite (but possibly very large-scale) dataset, F will reduce to a finite-sum structured function that appears in any application involving a pre-collected dataset.

Applications in the form of (1.1) include the robust phase retrieval that has been used in imaging and speech processing [16,17], the blind deconvolution in astronomy and computer vision [8,27], the robust principal component analysis in image deconvolution [7,9], the online nonnegative matrix factorization in image processing and pattern recognition [21], and the sparsity-regularized deep learning [53]. Specific formulations of some applications are given in section 6.

1.1. Proposed algorithm. We propose to solve (1.1) in a distributed environment. Suppose there are multiple agents. One agent is designated as the *master* and all the others as *worker*s. The master performs update to $\mathbf{x}$ while the workers compute sample (sub)gradients; see Fig. 1 for an illustration. The master-worker architecture has been adopted in many works. It can naturally happen, either because data are collected from local devices and then sent to a central server for processing such as in a sensor network application [38], or because the pre-collected dataset is too large to fit on a single machine and must be distributed over multiple machines.

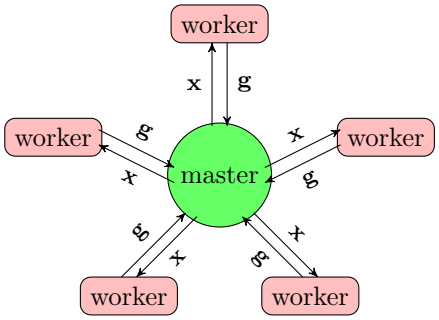

Figure 1: A master-worker architecture. The *master* performs update to $\mathbf{x}$; *worker*s compute sample (sub)gradients.

We assume that each worker can acquire samples of ξ and compute the (sub)gradient of each sampled function $f(\cdot\,;\xi)$. Each worker sends its computed sample (sub)gradient $\mathbf{g}$ to the master, and the latter updates $\mathbf{x}$ by using its received sample (sub)gradients and then sends the updated $\mathbf{x}$ to workers. Our scheme is described in Alg. 1, which is from the master's point of view.

Algorithm 1: A distributed stochastic inertial subgradient method for (1.1)

Initialization: choose $\mathbf{x}^{(0)} \in \mathrm{dom}(r)$ and set $\mathbf{x}^{(1)} = \mathbf{x}^{(0)}$
for $k = 1, 2, \ldots$ **do**
  Let $\mathbf{g}^{(k)} = \tilde{\nabla} f(\mathbf{x}^{(k-\tau_k)}; \xi_k)$ computed by a worker, where ξ_k is a sample of ξ and τ_k measures the possible delay;
  Choose stepsize $\alpha_k > 0$ and inertial parameter $\beta_k \geq 0$;
  Update the variable $\mathbf{x}$ by

$$\mathbf{x}^{(k+1)} = \mathbf{prox}_{\alpha_k r}\left(\mathbf{x}^{(k)} - \alpha_k \mathbf{g}^{(k)} + \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\right). \tag{1.2}$$

Here, $\tilde{\nabla} h(\mathbf{x})$ denotes a subgradient of a function h at $\mathbf{x}$, and it reduces to gradient if h is differentiable at $\mathbf{x}$. In (1.2), the proximal mapping is defined as

$$\mathbf{prox}_{\alpha r}(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathbb{R}^n} \left\{ r(\mathbf{y}) + \tfrac{1}{2\alpha}\|\mathbf{y} - \mathbf{x}\|^2 \right\}. \tag{1.3}$$

We use k to count the number of updates performed by the master. Notice that the master will update $\mathbf{x}$ once it receives a sample (sub)gradient from one worker, and we do not enforce coordination between the workers. Hence, the $\mathbf{g}^{(k)}$ used in (1.2) may not be a sample (sub)gradient

computed at $\mathbf{x}^{(k)}$ but at an outdated iterate $\mathbf{x}^{(k-\tau_k)}$. This setup with delayed information is the same as that in [1]. Also, instead of using a single sample, we can take multiple samples to compute $\mathbf{g}^{(k)}$ as the average of the multiple sample (sub)gradients.

Consider a special case, where $f(\cdot;\xi)$ is differentiable for each ξ and $r(\cdot)\equiv 0$. Then the update in (1.2) becomes $\mathbf{x}^{(k+1)}=\mathbf{x}^{(k)}-\alpha_k\mathbf{g}^{(k)}+\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})$. Let $\beta_k=\frac{\alpha_k}{\alpha_{k-1}}\beta$ for all $k\geq 1$ and for some $\beta\in(0,1)$. Define a recursive sequence by

$$\mathbf{m}^{(k)}=\beta\mathbf{m}^{(k-1)}+(1-\beta)\mathbf{g}^{(k)},\forall\, k\geq 1,\ \text{with } \mathbf{m}^{(0)}=\mathbf{0}. \tag{1.4}$$

Then the $\mathbf{x}$-update can be rewritten to

$$\mathbf{x}^{(k+1)}=\mathbf{x}^{(k)}-\tfrac{\alpha_k}{1-\beta}\mathbf{m}^{(k)}, \tag{1.5}$$

which is often referred as a momentum SGM in the literature (e.g., [20, 67])

Why use inertial force or momentum? Different from a standard proximal SsGM, we introduce an inertial force (or heavy-ball momentum term) $\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})$ in the update (1.2). If $\beta_k=0$, the update reduces to the standard proximal SsGM step. The heavy-ball momentum acceleration technique was first used in [48]. With the inertial force, a heavy-ball gradient method can mitigate the zigzagging behavior of a standard gradient descent method and potentially achieve faster convergence. For unconstrained strongly-convex quadratic optimization, it has been shown (cf. [50]) that the heavy-ball gradient method can achieve an optimal convergence rate. The advantage of using inertia has also been studied for deterministic composite nonconvex problems and stochastic smooth nonconvex problems. For example, the work [20] studies a more general momentum-based method, called Quasi-Hyperbolic Momentum (QHM), which includes the heavy-ball momentum as a special case. For unconstrained smooth problems, [20] gives a local linear convergence result that suggests the advantage of adding a heavy-ball momentum term in the update of a standard SGM. In addition, it provides supporting experiments to demonstrate that the optimal inertial parameter has a positive correlation with the condition number of the underlying problem. Although a heavy-ball momentum SGM has been extensively used in practice, a theoretical convergence guarantee is not yet achieved in the literature for nonconvex nonsmooth stochastic problems. We will provide a novel guideline of parameter setting for the inertial SGM or SsGM along with convergence guarantee, even if each $\mathbf{g}^{(k)}$ is computed at an outdated iterate. It is worth mentioning that for unconstrained smooth problems, a heavy-ball momentum SGM and Nesterov's Accelerated Gradient (NAG) are different special cases of QHM [20]. Though beyond the scope of this paper, our work may shed light on the acceleration effect of general momentum-based methods for nonsmooth nonconvex problems, such as QHM and NAG.

1.2. Related works.

Our method has a few key ingredients, including "stochastic subgradient", "inertia", "nonsmooth nonconvex", and "distributed delayed", which differentiate our method from existing ones. Below we review prior methods that share some ingredients with ours. We list a few closely-related methods with corresponding ingredients in Table 1.

Heavy-ball and inertial methods. Early advances based on the heavy-ball or inertial momentum acceleration technique can date back to [43, 48]. For decades, researchers have been designing heavy-ball or inertial methods for deterministic optimization [18, 30, 44–46, 68], structured stochastic optimization [19, 31, 32, 49, 61, 63], and even in the framework of maximal monotone operators [4, 5, 37]. Convergence analysis has been conducted to convex problems and also nonconvex

Table 1: A comparison of ingredients amongst several algorithms for solving problems in the form of (1.1). In the second column, "property of F" is to reflect the underlying assumption of F: "w.c." for weak convexity, "smooth" for Lipschitz continuous gradient, and "cvx" for convexity. In the third column, "inertia" is to reflect whether the algorithm introduces inertia. In the fourth column, "composite model" is to reflect the existence of r in (1.1): "proj." indicates a simple convex constraint, and "prox." indicates a proximable regularizer. In the fifth column, "distributed delayed" is to reflect whether the algorithm can handle a distributed setting with delayed (sub)gradient information. In the last column, convergence rate results for nonconvex models are listed: τ for the upper bound on the delay and K for the total number of iterations.

Method	property of F	inertia	composite model	distributed delayed	convergence rate
Mirror Descent [1]	smooth & cvx	no	no	yes	—
AdaptiveRevision [35]	smooth & cvx	no	no	yes	—
Random Incremental Subgrad. [41]	cvx	no	proj.	yes	—
AdaDelay [57]	smooth & cvx	no	proj.	yes	—
AsySG-con [28]	smooth	no	no	yes	$(1+\tau/\sqrt{K})/\sqrt{K}$
APAM [66]	smooth & cvx	yes	proj.	yes	—
	smooth	yes	no	yes	$(1+\tau/K^{1/4}+\tau^2/\sqrt{K})/\sqrt{K}$
SHB [33]	w.c.	yes	proj.	no	$1/\sqrt{K}$
This paper	w.c.	yes	proj. & prox.	yes	$(1+\tau/\sqrt{K}+\tau)/\sqrt{K}$
	smooth	yes	proj. & prox.	yes	$(1+\tau^2/\sqrt{K})/\sqrt{K}$
	smooth	yes	no	yes	$(1+\tau/\sqrt{K})/\sqrt{K}$

problems. For a convex deterministic model, [59, 60] provide last-iterate convergence for inertial methods. For a convex stochastic model, [40] proposes an inertial mirror descent method and establishes an $O(1/\sqrt{K})$ convergence rate result. Under a bounded-gradient assumption, [67] provides a unified convergence analysis of stochastic momentum methods for unconstrained smooth nonconvex stochastic optimization. [19] incorporates momentum acceleration in SGM and achieves an optimal oracle complexity result for (1.1) when F is smooth. The work [58] studies how heavy-ball technique can help SGM escape saddle points.

Distributed/parallel stochastic methods with delayed (sub)gradient information. There have been quite a few works about distributed delayed or asynchronous (async) parallel SGMs for convex or nonconvex problems and SsGMs for convex problems.

Similar to our method, [1] also adopts a master-worker setup. It analyzes a distributed delayed SGM for convex problems and establishes a convergence rate of $O(\frac{1+\tau^2/\sqrt{K}}{\sqrt{K}})$, where τ denotes the maximum delay of stochastic gradient and K is the total number of updates. Under a shared-memory setting, [51] proposes an async-parallel SGM for strongly-convex problems with a special sparsity structure and establishes a convergence rate of $O(\frac{1+\tau^2/\sqrt{n}}{K}\ln K)$, where n is the number of coordinates. [35] gives delay-tolerant algorithms for async distributed convex online learning problems. Its algorithms can achieve a regret of $O(\sqrt{(1+\tau)K})$ if a uniform upper bound τ on the delay is known and $O((1+\tau)\sqrt{K})$ otherwise. For smooth convex stochastic problems, [6, 57] adapt the stepsize of an async-parallel SGM to the staleness of stochastic gradient. More precisely, let τ_k denote the actual delay at iteration k. The stepsize of the methods in [6, 57] depends on τ_k. [57] analyzes its projected stochastic gradient scheme under the assumption that the delay has a bounded expectation $\mathbb{E}[\tau_k]=\bar{\tau}<\infty$ and a bounded second moment $\mathbb{E}[\tau_k^2]=\Omega(\bar{\tau}^2)$. The convergence rate is $O(\frac{\sqrt{1+\bar{\tau}+\bar{\tau}^4/\sqrt{K}}}{\sqrt{K}})$ if $\bar{\tau}$ is known and $O(\frac{1+\bar{\tau}+\bar{\tau}^4/\sqrt{K}}{\sqrt{K}})$ otherwise. Under the assumption $\mathbb{E}[\tau_k]=\bar{\tau}$, [6] achieves a rate of $O(\frac{1+\bar{\tau}^2/K}{K}\ln K)$ for unconstrained strongly convex

problems.

Async-parallel SGMs have also been studied for smooth non-convex problems. For example, [28] analyzes an async-parallel SGM for unconstrained stochastic problems and obtains a convergence rate of $O(\frac{1+\tau/\sqrt{K}}{\sqrt{K}})$ in terms of the expected value of gradient norm square; [23] analyzes an async-parallel variance-reduced SGM for a finite-sum structured problem and shows a sublinear convergence when $\tau = O(1)$; [66] focuses on async distributed and parallel adaptive (i.e., quasi-Newton-type) SGM for unconstrained stochastic problems and gives a convergence rate of $O(\frac{1+\tau/K^{1/4}+\tau^2/\sqrt{K}}{\sqrt{K}})$. The studies on delayed SsGMs are still limited and only for convex problems. For example, [41] proposes an async projected SsGM and shows an almost-sure subsequence convergence result but with no convergence rate result.

The distributed/parallel methods mentioned above either adopt a master-worker setup (i.e., centralized) or assume a shared-memory setting. Many other works about SGMs or SsGMs are built on a decentralized setting, where multiple agents are distributed on a connected network and can only communicate with their neighbors but not a central master agent. Extending our discussions to the decentralized setting is beyond the scope of this paper. The interested readers can refer to [13, 29, 34, 64] and the references therein.

Most closely-related works. The methods in [10, 33] are perhaps the most closely related to ours. [10] gives a decentralized projected deterministic subgradient method for weakly-convex optimization. It establishes a sublinear convergence result for the deterministic method. A stochastic variant is also given in [10] with subsequence convergence but no convergence rate. In comparison to [10], we incorporate the inertial-force acceleration in a proximal SsGM to achieve empirically faster convergence, and in addition, we allow for delayed subgradient and can still achieve sublinear convergence. [33] proposes a projected inertial SsGM for weakly-convex stochastic optimization. The method appears similar to Alg. 1. However, its analysis is completely different from ours, and it does not consider the delayed case. More importantly, its theoretical result is not established on the inertial-generated sequence. This is explained as follows. The update of the method in [33] is

$$\mathbf{x}^{(k+1)} = \mathrm{Proj}_X\left(\mathbf{x}^{(k)} - \alpha\beta\mathbf{g}^{(k)} + (1-\beta)(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\right), \tag{1.6}$$

where Proj_X denotes the projection onto a closed convex set X. Its analysis is only on the choice of $\alpha\beta = \Theta(\frac{1}{K})$ for a given maximum number K of updates and $1-\beta = 1-\frac{1}{\sqrt{K}}$. The sequence generated from (1.6) is similar to that we generate from (1.2), i.e., inertial-generated sequence. However, the theoretical result in [33] is not about $\{\mathbf{x}^{(k)}\}$ but on the extrapolated sequence $\{\bar{\mathbf{x}}^{(k)} := \mathbf{x}^{(k)} + \frac{1-\beta}{\beta}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\}$. There are two potential issues on analyzing the property of $\{\bar{\mathbf{x}}^{(k)}\}$. First, if $X \neq \mathbb{R}^n$, the sequence may not be in X. In fact, $\bar{\mathbf{x}}^{(k)}$ can be far away from X if $\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \neq \mathbf{0}$ as $\frac{1-\beta}{\beta} = \sqrt{K}-1$ is big. Second, if $X = \mathbb{R}^n$, it holds $\bar{\mathbf{x}}^{(k+1)} = \bar{\mathbf{x}}^{(k)} - \alpha\mathbf{g}^{(k)}$, and in this case, $\{\bar{\mathbf{x}}^{(k)}\}$ is more like a non-inertial sequence, as compared to the sequence generated by the momentum SGM in (1.5). In contrast, our analysis will be on the inertial-generated sequence.

1.3. Contributions.

- We propose a proximal inertial stochastic subgradient method in Alg. 1 for solving non-convex stochastic Problem (1.1). The method can tolerate a delay of derivative information in a distributed environment. To the best of our knowledge, it is the first method that applies the inertial-acceleration technique in a proximal stochastic subgradient method for non-convex problems.

- We provide convergence rate analysis of the proposed method for three problem classes in the form of (1.1). For each problem class, the method, with an appropriate setting of parameters, enjoys an $O(\frac{1}{\sqrt{K}})$ convergence rate in terms of the expected value of a gradient norm square, where K is the number of total iterations. First, when F is weakly-convex (see Def. 2.1 below) and possibly nondifferentiable and r is convex, we establish the $O(\frac{1}{\sqrt{K}})$ convergence rate by choosing $\alpha_k = \Theta(\frac{1}{\sqrt{K}})$ and $\beta_k = \Theta(\frac{1}{K^{1/4}}), \forall\, k \leq K$, provided that the delay τ_k follows a static distribution and is bounded by $\tau = O(1)$. Second, when F is smooth but possibly non-convex and r is convex, we obtain the $O(\frac{1}{\sqrt{K}})$ convergence rate by the same choice of α_k and β_k as in the first case, under a relaxed condition on τ_k, i.e., $\tau_k = O(K^{1/4})$ for all k. Third, for the case of a smooth F and $r \equiv 0$, we obtain the $O(\frac{1}{\sqrt{K}})$ convergence rate with the choice of $\alpha_k = \Theta(\frac{1}{\sqrt{K}})$ and $\beta_k = \beta \in (0,1), \forall\, k \leq K$, provided that $\tau_k = O(\sqrt{K})$ for all k. Hence, the proposed method can tolerate a larger delay if the problem has a nicer structure.
- We conduct numerical experiments of the proposed method on three applications to demonstrate the effect of the inertial acceleration and also to demonstrate the higher parallelization speed-up by the asynchronous implementation over a synchronous counterpart.

1.4. Notation and organization. We use lower-case bold letters $\mathbf{x}, \mathbf{y}, \ldots$ for vectors. A superscript $^{(k)}$ is used to specify the iterate, i.e., $\mathbf{x}^{(k)}$ denotes the k-th iterate. We use $\|\cdot\|$ to denote the Euclidean norm of a vector and also the spectral norm of a matrix. We use the big-O notation with the standard meaning to compare two quantities that can both approach to infinity or zero. The randomness of Alg. 1 comes from the samples $\{\xi_k\}_{k\geq 1}$. In our analysis, we use $\mathbb{E}_k$ for the conditional expectation with the history until the k-th iteration, i.e., $\mathbb{E}_k[\,\cdot\,] = \mathbb{E}\left[\,\cdot \,|\, \{\xi_j\}_{j=1}^{k-1}\right]$.

The rest of the paper is organized as follows. In section 2, we give some basic concepts and preliminary results. The detailed analysis and convergence rate results are shown in section 3-5 for three different problem classes. Numerical results are given in section 6. Finally, section 7 concludes the paper.

2. Preliminaries. In this section, we give some basic concepts and preliminary results that will be used in our analysis. For a function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, we let $\partial\phi(\mathbf{x})$ denote its subdifferential at $\mathbf{x}$, i.e., the set of subgradients, which consists of all vectors $\mathbf{v}$ satisfying

$$\phi(\mathbf{y}) \geq \phi(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x}\rangle + o\left(\|\mathbf{y} - \mathbf{x}\|\right) \quad \text{as } \mathbf{y} \to \mathbf{x}.$$

The definition and results below can be found in [11, 14].

Definition 2.1. *A function ϕ is ρ-weakly convex if $\phi(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex for some $\rho > 0$.*

Lemma 2.2. *If ϕ is ρ-weakly convex, then*

$$\phi(\mathbf{y}) \geq \phi(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x}\rangle - \tfrac{\rho}{2}\|\mathbf{y} - \mathbf{x}\|^2, \forall\, \mathbf{x},\, \mathbf{y} \in \mathrm{dom}(\phi),\, \forall\, \mathbf{v} \in \partial\phi(\mathbf{x}), \tag{2.1}$$

and

$$\langle \mathbf{v} - \mathbf{w}, \mathbf{x} - \mathbf{y}\rangle \geq -\rho\|\mathbf{y} - \mathbf{x}\|^2, \forall\, \mathbf{x},\, \mathbf{y} \in \mathrm{dom}(\phi),\, \forall\, \mathbf{v} \in \partial\phi(\mathbf{x}), \mathbf{w} \in \partial\phi(\mathbf{y}). \tag{2.2}$$

The class of weakly-convex functions is rather big. It includes all convex functions and all smooth functions. In addition, the composition function $h(c(\mathbf{x}))$ is also weakly-convex, if $h : \mathbb{R}^m \to$

$\mathbb{R}$ is convex and Lipschitz continuous and $c : \mathbb{R}^n \to \mathbb{R}^m$ is smooth. Specific applications that have weakly-convex objectives include nonlinear least squares, phase retrieval, robust PCA, robust low rank matrix recovery, optimization of the Conditional Value-at-Risk, and graph synchronization. More examples can be found in [14].

A key tool used in recent works (e.g., [2, 10, 11, 33, 39]) about stochastic weakly-convex minimization is the Moreau envelope [36], which is defined as follows.

Definition 2.3. *For a ρ-weakly convex function ϕ and $\lambda \in (0, 1/\rho)$, the Moreau envelope $\phi_\lambda(\cdot)$ is defined as*

$$\phi_\lambda(\mathbf{x}) = \min_{\mathbf{y}} \left\{ \phi(\mathbf{y}) + \tfrac{1}{2\lambda} \|\mathbf{y} - \mathbf{x}\|^2 \right\}. \tag{2.3}$$

The Moreau envelope is useful to characterize near-stationarity of a point $\mathbf{x}$ because of the results in the following lemma. From (2.4), we notice that if $\|\nabla \phi_\lambda(\mathbf{x})\|$ is small, then $\widetilde{\mathbf{x}} := \mathbf{prox}_{\lambda\phi}(\mathbf{x})$ will be a near-stationary point of ϕ and $\mathbf{x}$ is close to $\widetilde{\mathbf{x}}$.

Lemma 2.4. *Let ϕ be ρ-weakly convex, then for any $\lambda \in (0, 1/\rho)$, the Moreau envelope ϕ_λ is smooth with gradient given by*

$$\nabla \phi_\lambda(\mathbf{x}) = \lambda^{-1}\big(\mathbf{x} - \widetilde{\mathbf{x}}\big),$$

where $\widetilde{\mathbf{x}} := \mathbf{prox}_{\lambda\phi}(\mathbf{x})$. Moreover,

$$\|\mathbf{x} - \widetilde{\mathbf{x}}\| = \lambda \|\nabla \phi_\lambda(\mathbf{x})\|, \ \phi(\widetilde{\mathbf{x}}) \le \phi(\mathbf{x}), \quad \text{and} \quad \operatorname{dist}(\mathbf{0}, \partial\phi(\widetilde{\mathbf{x}})) \le \|\nabla \phi_\lambda(\mathbf{x})\|. \tag{2.4}$$

Besides the class of weakly-convex functions, we will also consider smooth functions in our analysis, for which we are able to obtain stronger theoretical results. By slightly abusing the notation, we also use ρ to denote the Lipschitz constant of a smooth function, as a ρ-smooth function must be ρ-weakly convex.

Definition 2.5. *A function ϕ is ρ-smooth, if it is differentiable, and*

$$\|\nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{y})\| \le \rho \|\mathbf{x} - \mathbf{y}\|, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

If ϕ is ρ-smooth, then

$$|\phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla \phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle| \le \tfrac{\rho}{2} \|\mathbf{x} - \mathbf{y}\|^2, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{2.5}$$

3. Convergence analysis for nonsmooth weakly-convex problems.

In this section, we analyze Alg. 1 for problems in the form of (1.1), where F is possibly nondifferentiable. Throughout this section, we make the following assumptions.

Assumption 1 (weak convexity). F is ρ-weakly convex with $\rho > 0$.

Assumption 2 (unbiased subgradient). $\mathbf{g}^{(k)}$ is an unbiased stochastic subgradient of F at $\mathbf{x}^{(k-\tau_k)}$ for each k, i.e., $\mathbb{E}_{\xi_k}[\mathbf{g}^{(k)}] \in \partial F(\mathbf{x}^{(k-\tau_k)})$.

Assumption 3 (bounded subgradient). There is a real number $M \ge 0$ such that $\mathbb{E}_\xi \|\tilde{\nabla} f(\mathbf{x}; \xi)\|^2 \le M^2$ for all $\mathbf{x} \in \operatorname{dom}(r)$ and all subgradient $\tilde{\nabla} f(\mathbf{x}; \xi) \in \partial f(\mathbf{x}; \xi)$.

3.1. Preparatory lemmas. For a fixed $\overline{\rho} > \rho$, we denote

$$\mathbf{v}^{(k)} = \mathbb{E}_{\xi_k}\left[\mathbf{g}^{(k)}\right] \in \partial F(\mathbf{x}^{(k-\tau_k)}), \quad \widetilde{\mathbf{x}}^{(k)} = \mathbf{prox}_{\phi/\overline{\rho}}(\mathbf{x}^{(k)}), \tag{3.1a}$$

and choose

$$\widetilde{\mathbf{v}}^{(k)} \in \partial F(\widetilde{\mathbf{x}}^{(k)}) \text{ such that } \overline{\rho}(\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}) \in \partial r(\widetilde{\mathbf{x}}^{(k)}) + \widetilde{\mathbf{v}}^{(k)}. \tag{3.1b}$$

Note that the existence of $\widetilde{\mathbf{v}}^{(k)}$ is guaranteed from the definition of $\widetilde{\mathbf{x}}^{(k)}$. By Assumption 3, it holds that

$$\mathbb{E}_{\xi_k}\|\mathbf{g}^{(k)}\|^2 \le M^2, \quad \|\mathbf{v}^{(k)}\|^2 \le M^2, \text{ and } \|\widetilde{\mathbf{v}}^{(k)}\|^2 \le M^2. \tag{3.2}$$

The next result is from [11, Lemma 3.2]. Its proof only relies on the definition of $\widetilde{\mathbf{x}}^{(k)}$ and the choice of $\widetilde{\mathbf{v}}^{(k)}$. Hence, the result still holds for our case, though the algorithm in [11, Lemma 3.2] does not have an inertial term in its update.

Lemma 3.1. *Let $\widetilde{\mathbf{x}}^{(k)}$ and $\widetilde{\mathbf{v}}^{(k)}$ be defined as in (3.1a) and (3.1b). Then*

$$\widetilde{\mathbf{x}}^{(k)} = \mathbf{prox}_{\alpha_k r}\big(\alpha_k\overline{\rho}\mathbf{x}^{(k)} - \alpha_k\widetilde{\mathbf{v}}^{(k)} + (1-\alpha_k\overline{\rho})\widetilde{\mathbf{x}}^{(k)}\big). \tag{3.3}$$

The next lemma extends the hypomonotonicity property of a weakly-convex function, in order to deal with the case with delayed subgradients.

Lemma 3.2. *Let $\widetilde{\mathbf{x}}^{(k)}$, $\mathbf{v}^{(k)}$ and $\widetilde{\mathbf{v}}^{(k)}$ be defined as in (3.1). Then under Assumption 1, it holds*

$$\begin{aligned} &-\left\langle \mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}, \mathbf{v}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\right\rangle \\ &\le F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-\tau_k)}) + \tfrac{\rho}{2}\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + \tfrac{\rho}{2}\|\mathbf{x}^{(k-\tau_k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 - \left\langle \mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}, \mathbf{v}^{(k)}\right\rangle. \end{aligned} \tag{3.4}$$

Proof. From the ρ-weak convexity of F, it follows that

$$\left\langle \mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}, \widetilde{\mathbf{v}}^{(k)}\right\rangle \le F(\mathbf{x}^{(k)}) - F(\widetilde{\mathbf{x}}^{(k)}) + \tfrac{\rho}{2}\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2, \tag{3.5}$$

and

$$-\left\langle \mathbf{x}^{(k-\tau_k)} - \widetilde{\mathbf{x}}^{(k)}, \mathbf{v}^{(k)}\right\rangle \le F(\widetilde{\mathbf{x}}^{(k)}) - F(\mathbf{x}^{(k-\tau_k)}) + \tfrac{\rho}{2}\|\mathbf{x}^{(k-\tau_k)} - \widetilde{\mathbf{x}}^{(k)}\|^2. \tag{3.6}$$

Hence, we obtain the desired result by adding the two inequalities in (3.5) and (3.6), and also noticing

$$-\left\langle \mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}, \mathbf{v}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\right\rangle = \left\langle \mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}, \widetilde{\mathbf{v}}^{(k)}\right\rangle - \left\langle \mathbf{x}^{(k-\tau_k)} - \widetilde{\mathbf{x}}^{(k)}, \mathbf{v}^{(k)}\right\rangle - \left\langle \mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}, \mathbf{v}^{(k)}\right\rangle.$$

This completes the proof. □

The result in the next lemma establishes a descent property of the iterate sequence from Alg. 1 by relating it to the virtual sequence $\{\widetilde{\mathbf{x}}^{(k)}\}$. It extends the result in [11, Lemma 3.3].

Lemma 3.3. *Let $\overline{\rho} \in (\rho, 2\rho]$ and $\alpha_k \in (0, 1/\overline{\rho}]$ for all k. Under Assumptions 1–3, the iterate sequence $\{\mathbf{x}^{(k)}\}$ from Alg. 1 with stepsize sequence $\{\alpha_k\}$ and inertial parameter $\{\beta_k\}$ satisfies*

$$\begin{aligned} \mathbb{E}_{\xi_k}\|\mathbf{x}^{(k+1)} - \widetilde{\mathbf{x}}^{(k)}\|^2 \le{}& \big(1 - 2\alpha_k(\overline{\rho} - \rho) + c_k\big)\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + (2 + \tfrac{1}{c_k})\beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 \\ &+ 8\alpha_k^2 M^2 + 2\alpha_k(1 - \alpha_k\overline{\rho})\widehat{\mathcal{E}}_k, \end{aligned} \tag{3.7}$$

where $\widetilde{\mathbf{x}}^{(k)}$ is defined in (3.1a), c_k is any positive number, and

$$\widehat{\mathcal{E}}_k := F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-\tau_k)}) - \tfrac{\rho}{2}\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + \tfrac{\rho}{2}\|\mathbf{x}^{(k-\tau_k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 - \left\langle \mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}, \mathbf{v}^{(k)}\right\rangle. \tag{3.8}$$

The next lemma will be used to bound $\sum_{k=1}^{K}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2$ for any given integer K.

Lemma 3.4. *Let $\{\mathbf{x}^{(k)}\}$ be generated from Alg. 1. Under Assumptions 1 and 3, it holds for any $\gamma>0$ that*

$$\big(1-\gamma-\tfrac{\alpha_k\rho}{2}-\tfrac{\beta_k}{2}\big)\mathbb{E}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\le\alpha_k\mathbb{E}\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big)+\tfrac{\beta_k}{2}\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2+\tfrac{\alpha_k^2M^2}{\gamma}. \tag{3.9}$$

Proof. By the convexity of r, we have $\big\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)},\tilde{\nabla}r(\mathbf{x}^{(k+1)})\big\rangle\le r(\mathbf{x}^{(k)})-r(\mathbf{x}^{(k+1)})$. In addition, it follows from (1.2) that $\mathbf{0}\in\alpha_k\partial r(\mathbf{x}^{(k+1)})+\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}+\alpha_k\mathbf{g}^{(k)}-\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})$. Hence,

$$\big\langle\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)},\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}+\alpha_k\mathbf{g}^{(k)}-\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})\big\rangle\le\alpha_k\big(r(\mathbf{x}^{(k)})-r(\mathbf{x}^{(k+1)})\big). \tag{3.10}$$

By the ρ-weak convexity of F, it holds

$$\big\langle\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)},\tilde{\nabla}F(\mathbf{x}^{(k+1)})\big\rangle\ge F(\mathbf{x}^{(k+1)})-F(\mathbf{x}^{(k)})-\tfrac{\rho}{2}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2,$$

and thus

$$\begin{aligned}\big\langle\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)},\alpha_k\mathbf{g}^{(k)}\big\rangle\ge{}&\alpha_k\big\langle\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)},\mathbf{g}^{(k)}-\tilde{\nabla}F(\mathbf{x}^{(k+1)})\big\rangle\\&+\alpha_k\big(F(\mathbf{x}^{(k+1)})-F(\mathbf{x}^{(k)})-\frac{\rho}{2}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\big).\end{aligned} \tag{3.11}$$

Plugging (3.11) into (3.10) and rearranging terms give

$$\begin{aligned}\big(1-\tfrac{\alpha_k\rho}{2}\big)\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\le{}&\alpha_k\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big)+\beta_k\big\langle\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)},\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\big\rangle\\&-\alpha_k\big\langle\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)},\mathbf{g}^{(k)}-\tilde{\nabla}F(\mathbf{x}^{(k+1)})\big\rangle.\end{aligned} \tag{3.12}$$

Now using Assumption 3 and the Young's inequality, we have

$$\begin{aligned}\big(1-\tfrac{\alpha_k\rho}{2}\big)\mathbb{E}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\le{}&\alpha_k\mathbb{E}\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big)+\tfrac{\beta_k}{2}\mathbb{E}\big(\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2+\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2\big)\\&+\gamma\mathbb{E}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2+\tfrac{\alpha_k^2M^2}{\gamma}.\end{aligned}$$

Rearranging terms in the above inequality gives the desired result. □

3.2. Convergence rate results. In this subsection, we establish the convergence rate results of Alg. 1 for nonsmooth weakly-convex problems by using the lemmas in the previous subsection. We first give a generic result as follows.

Theorem 3.5. *Given a positive integer K, let $\{\mathbf{x}^{(k)}\}_{k=1}^{K}$ be generated from Alg. 1 with a stepsize sequence $\{\alpha_k\}$ and inertial parameter sequence $\{\beta_k\}$. Under Assumptions 1–3, let $\overline{\rho}\in(\rho,2\rho]$ and assume $\alpha_k\in(0,1/\overline{\rho}]$ for all k. Then*

$$\begin{aligned}\mathbb{E}\big\|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(T)})\big\|^2\le{}&\tfrac{2\overline{\rho}}{(\overline{\rho}-\rho)\sum_{k=k_0}^{K}\alpha_k}\Big[\mathbb{E}\big[\phi_{1/\overline{\rho}}(\mathbf{x}^{(k_0)})-\phi^*\big]+\tfrac{\overline{\rho}}{2}\textstyle\sum_{k=k_0}^{K}(2+\frac{2}{\alpha_k(\overline{\rho}-\rho)})\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2\\&+4\overline{\rho}M^2\textstyle\sum_{k=k_0}^{K}\alpha_k^2+\sum_{k=k_0}^{K}\alpha_k\overline{\rho}(1-\alpha_k\overline{\rho})\mathbb{E}[\mathcal{E}_k]\Big],\end{aligned} \tag{3.13}$$

where $k_0\ge1$ is an integer, T is randomly selected from $\{k_0,\dots,K\}$ by the distribution

$$\mathrm{Prob}(T=k)=\tfrac{\alpha_k}{\sum_{j=k_0}^{K}\alpha_j},\ \forall\,k=k_0,\dots,K, \tag{3.14}$$

and

$$\mathcal{E}_k:=F(\mathbf{x}^{(k)})-F(\mathbf{x}^{(k-\tau_k)})+\big(\tfrac{\rho}{2}+\tfrac{\rho^2}{\overline{\rho}-\rho}\big)\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2-\big\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k-\tau_k)},\mathbf{v}^{(k)}\big\rangle. \tag{3.15}$$

Proof. By the definition of ϕ_λ in (2.3) and Lemma 3.3, we have

$$\begin{aligned}
&\mathbb{E}_{\xi_k}\left[\phi_{1/\overline{\rho}}(\mathbf{x}^{(k+1)})\right] \\
&\le \mathbb{E}_{\xi_k}\left[\phi(\widetilde{\mathbf{x}}^{(k)}) + \tfrac{\overline{\rho}}{2}\|\mathbf{x}^{(k+1)} - \widetilde{\mathbf{x}}^{(k)}\|^2\right] \\
&\le \phi(\widetilde{\mathbf{x}}^{(k)}) + \tfrac{\overline{\rho}}{2}\left[\left(1 - 2\alpha_k(\overline{\rho} - \rho) + c_k\right)\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + (2 + \tfrac{1}{c_k})\beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + 8\alpha_k^2 M^2\right] \\
&\quad + \alpha_k\overline{\rho}(1 - \alpha_k\overline{\rho})\widehat{\mathcal{E}}_k \\
&= \phi_{1/\overline{\rho}}(\mathbf{x}^{(k)}) - \tfrac{\overline{\rho}}{2}\left(2\alpha_k(\overline{\rho} - \rho) - c_k\right)\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + \tfrac{\overline{\rho}}{2}(2 + \tfrac{1}{c_k})\beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + 4\overline{\rho}\alpha_k^2 M^2 \\
&\quad + \alpha_k\overline{\rho}(1 - \alpha_k\overline{\rho})\widehat{\mathcal{E}}_k. \qquad (3.16)
\end{aligned}$$

where $\widehat{\mathcal{E}}_k$ is defined in (3.8). By the Young's inequality, we have

$$\begin{aligned}
-\tfrac{\rho}{2}\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + \tfrac{\rho}{2}\|\mathbf{x}^{(k-\tau_k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 &= \tfrac{\rho}{2}\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2 + \rho\langle\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}, \mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\rangle \\
&\le \left(\tfrac{\rho}{2} + \tfrac{\rho^2}{\overline{\rho} - \rho}\right)\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2 + \tfrac{\overline{\rho} - \rho}{4}\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2.
\end{aligned}$$

Using the definition of $\widehat{\mathcal{E}}_k$ in (3.8) and substituting the inequality above into (3.16), we have from $1 - \alpha_k\overline{\rho} \le 1$ and the definition of $\mathcal{E}_k$ in (3.15) that

$$\begin{aligned}
&\mathbb{E}_{\xi_k}\left[\phi_{1/\overline{\rho}}(\mathbf{x}^{(k+1)})\right] \\
&\le \phi_{1/\overline{\rho}}(\mathbf{x}^{(k)}) - \tfrac{\overline{\rho}}{2}\left(\tfrac{3}{2}\alpha_k(\overline{\rho} - \rho) - c_k\right)\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + \tfrac{\overline{\rho}}{2}(2 + \tfrac{1}{c_k})\beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + 4\overline{\rho}\alpha_k^2 M^2 \\
&\quad + \alpha_k\overline{\rho}(1 - \alpha_k\overline{\rho})\mathcal{E}_k. \qquad (3.17)
\end{aligned}$$

Taking full expectation and summing the inequality in (3.17) over $k = k_0, \ldots, K$, we have

$$\begin{aligned}
&\mathbb{E}\left[\phi_{1/\overline{\rho}}(\mathbf{x}^{(K+1)})\right] \\
&\le \mathbb{E}\left[\phi_{1/\overline{\rho}}(\mathbf{x}^{(k_0)})\right] - \tfrac{\overline{\rho}}{2}\textstyle\sum_{k=k_0}^{K}\left(\tfrac{3}{2}\alpha_k(\overline{\rho} - \rho) - c_k\right)\mathbb{E}\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 \\
&\quad + \tfrac{\overline{\rho}}{2}\textstyle\sum_{k=k_0}^{K}(2 + \frac{1}{c_k})\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + 4\overline{\rho}M^2\textstyle\sum_{k=k_0}^{K}\alpha_k^2 + \sum_{k=k_0}^{K}\alpha_k\overline{\rho}(1 - \alpha_k\overline{\rho})\mathcal{E}_k.
\end{aligned}$$

Choose $c_k = \frac{1}{2}\alpha_k(\overline{\rho} - \rho)$ for all $k \ge 1$ and rearrange the above inequality. We obtain

$$\begin{aligned}
\tfrac{\overline{\rho}(\overline{\rho} - \rho)}{2}\textstyle\sum_{k=k_0}^{K}\alpha_k\mathbb{E}\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 &\le \mathbb{E}\left[\phi_{1/\overline{\rho}}(\mathbf{x}^{(k_0)}) - \phi^*\right] + 4\overline{\rho}M^2\textstyle\sum_{k=k_0}^{K}\alpha_k^2 \\
&\quad + \tfrac{\overline{\rho}}{2}\textstyle\sum_{k=k_0}^{K}(2 + \frac{2}{\alpha_k(\overline{\rho} - \rho)})\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + \sum_{k=k_0}^{K}\alpha_k\overline{\rho}(1 - \alpha_k\overline{\rho})\mathbb{E}[\mathcal{E}_k], \qquad (3.18)
\end{aligned}$$

where we have used the fact $\phi_{1/\overline{\rho}}(\mathbf{x}) \ge \phi^*, \forall\, \mathbf{x} \in \mathrm{dom}(r)$. From Lemma 2.4, we have $\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 = \|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(k)})\|^2/\overline{\rho}^2$. Hence, plugging this equation into the left-hand side of (3.18) and using the choice of T in (3.14), we obtain the desired result. □

To show the convergence rate in (3.13), it suffices to bound the summation terms on $\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2$ and the delay term $\mathbb{E}[\mathcal{E}_k]$. If the delay is arbitrary, it is impossible to have convergence, and thus a certain condition on τ_k is needed. For nonsmooth problems, we make the following assumption.

Assumption 4 (stochastic delay). There is an integer τ such that the staleness τ_k follows the distribution

$$\mathrm{Prob}(\tau_k = j) = p_j, \quad \text{for } j = 0, 1, \ldots, \tau, \ \forall\, k.$$

If the computing environment does not change during all the iterations, the assumption will hold. In addition, one can track the delay at the master node and thus estimate the probability. However, we do not need to know the values of $\{p_j\}$ or τ in the computation and analysis, but we only require their existence. A similar assumption has been made in [22, 47, 57].

In the rest of this section, we show convergence rate results separately for the case with a fixed stepsize sequence and the one with a varying stepsize sequence.

3.2.1. Convergence rate with a fixed stepsize. In this subsubsection, we consider the case where $\alpha_k = \alpha_1$ and $\beta_k = \beta_1$ for all $k \geq 1$. In this case, it is easy to bound the summation term about $\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2$.

Lemma 3.6. *Given a positive integer K, let $\alpha_k = \frac{\alpha}{\sqrt{K}}, \forall\, k = 1, \ldots, K$ for some $\alpha > 0$. Also, let $\beta_k = \frac{\beta}{K^{1/4}}, \forall\, k$ for some nonnegative β such that $\frac{\beta}{K^{1/4}} < 1 - \frac{\alpha\rho}{2\sqrt{K}}$. Then under Assumptions 1 and 3, it holds*

$$\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \leq \frac{\alpha}{\gamma\sqrt{K}}\big(\phi(\mathbf{x}^{(1)}) - \phi^*\big) + \frac{\alpha^2 M^2}{\gamma^2}, \text{ where } \gamma = \tfrac{1}{2}\big(1 - \tfrac{\alpha\rho}{2\sqrt{K}} - \tfrac{\beta}{K^{1/4}}\big). \tag{3.19}$$

Proof. Let $\gamma = \frac{1}{2}\big(1 - \frac{\alpha\rho}{2\sqrt{K}} - \frac{\beta}{K^{1/4}}\big)$ in (3.9) and sum it up over k. We have

$$\begin{aligned}&\big(1 - \gamma - \tfrac{\alpha\rho}{2\sqrt{K}} - \tfrac{\beta}{2K^{1/4}}\big)\textstyle\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2\\ \leq\ & \tfrac{\alpha}{\sqrt{K}}\mathbb{E}\big(\phi(\mathbf{x}^{(1)}) - \phi(\mathbf{x}^{(K+1)})\big) + \tfrac{\beta}{2K^{1/4}}\textstyle\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + \tfrac{\alpha^2 M^2}{\gamma}.\end{aligned}$$

Since $\mathbf{x}^{(0)} = \mathbf{x}^{(1)}$ and $\phi(\mathbf{x}^{(K+1)}) \geq \phi^*$, the above inequality together with the choice of γ implies the desired result. We complete the proof. □

When a fixed stepsize sequence is used, we can bound $\sum_{k=1}^K \mathbb{E}[\mathcal{E}_k]$ as in the next lemma.

Lemma 3.7. *Let $\mathcal{E}_k$ be defined in (3.15). Given a positive integer K, let $\alpha_k = \frac{\alpha}{\sqrt{K}}, \forall\, k = 1, \ldots, K$ for some $\alpha > 0$. Also, let $\beta_k = \frac{\beta}{K^{1/4}}, \forall\, k$ for some nonnegative β such that $\frac{\beta}{K^{1/4}} < 1 - \frac{\alpha\rho}{2\sqrt{K}}$. Suppose that $F(\mathbf{x})$ is upper bounded by C_F for all $\mathbf{x} \in \mathrm{dom}(r)$. Then under Assumptions 1, 3, and 4, we have*

$$\begin{aligned}\sum_{k=1}^K \mathbb{E}[\mathcal{E}_k] \leq\ & \tau\max\big\{0, -F(\mathbf{x}^{(1)})\big\} + \tau C_F + \tau^2\big(\tfrac{\rho}{2} + \tfrac{\rho^2}{\bar\rho - \rho}\big)\Big(\tfrac{\alpha}{\gamma\sqrt{K}}\big(\phi(\mathbf{x}^{(1)}) - \phi^*\big) + \tfrac{\alpha^2 M^2}{\gamma^2}\Big)\\ &+ M\tau\sqrt{K}\sqrt{\tfrac{\alpha}{\gamma\sqrt{K}}\big(\phi(\mathbf{x}^{(1)}) - \phi^*\big) + \tfrac{\alpha^2 M^2}{\gamma^2}},\end{aligned} \tag{3.20}$$

where $\gamma = \frac{1}{2}\big(1 - \frac{\alpha\rho}{2\sqrt{K}} - \frac{\beta}{K^{1/4}}\big)$.

Now from Theorem 3.5 and Lemmas 3.6 and 3.7, we can easily show the following convergence rate result.

Theorem 3.8 (convergence rate with fixed stepsize). *Under Assumptions 1–3 and 4, let $\bar\rho \in (\rho, 2\rho]$ and K be the maximum number of iterations. Let $\{\mathbf{x}^{(k)}\}$ be the sequence from Alg. 1 with $\alpha_k = \frac{\alpha}{\sqrt{K}}$ and $\beta_k = \frac{\beta}{K^{1/4}}, \forall\, k = 1, \ldots, K$ for some $\alpha > 0$ and nonnegative β such that $\frac{\alpha}{\sqrt{K}} \in (0, 1/\bar\rho]$*

and $\frac{\beta}{K^{1/4}} < 1 - \frac{\alpha\rho}{2\sqrt{K}}$. *Suppose that* $F(\mathbf{x})$ *is upper bounded by* C_F *for all* $\mathbf{x} \in \mathrm{dom}(r)$. *Then*

$$\begin{aligned}\mathbb{E}\left\|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(T)})\right\|^2 \le \frac{2\overline{\rho}}{(\overline{\rho}-\rho)\alpha\sqrt{K}}\Bigg[&\frac{\overline{\rho}}{2}\Big(2+\frac{2\sqrt{K}}{\alpha(\overline{\rho}-\rho)}\Big)\frac{\beta^2}{\sqrt{K}}\Big(\frac{\alpha}{\gamma\sqrt{K}}\big(\phi(\mathbf{x}^{(1)})-\phi^*\big)+\frac{\alpha^2M^2}{\gamma^2}\Big)+4\overline{\rho}M^2\alpha^2\\ &+\frac{\alpha\overline{\rho}\tau}{\sqrt{K}}\Big(\max\big\{0,-F(\mathbf{x}^{(1)})\big\}+C_F+\tau\big(\frac{\rho}{2}+\frac{\rho^2}{\overline{\rho}-\rho}\big)\big(\frac{\alpha}{\gamma\sqrt{K}}\big(\phi(\mathbf{x}^{(1)})-\phi^*\big)+\frac{\alpha^2M^2}{\gamma^2}\big)\Big)\\ &+\phi_{1/\overline{\rho}}(\mathbf{x}^{(1)})-\phi^*+M\alpha\overline{\rho}\tau\sqrt{\frac{\alpha}{\gamma\sqrt{K}}\big(\phi(\mathbf{x}^{(1)})-\phi^*\big)+\frac{\alpha^2M^2}{\gamma^2}}\Bigg],\end{aligned} \tag{3.21}$$

where $\gamma = \frac{1}{2}\big(1 - \frac{\alpha\rho}{2\sqrt{K}} - \frac{\beta}{K^{1/4}}\big)$ *and* T *is randomly selected from* $\{1,\dots,K\}$ *by* (3.14).

Proof. Notice $\sum_{k=1}^K \alpha_k = \alpha\sqrt{K}$ and $\sum_{k=1}^K \alpha_k^2 = \alpha^2$. Then the inequality in (3.21) directly follows by substituting (3.19) and (3.20) into (3.13) with $k_0 = 1$, and also noticing $1 - \frac{\alpha\overline{\rho}}{\sqrt{K}} \le 1$. □

Remark 3.9. The result in (3.21) indicates a convergence rate of $O(1/\sqrt{K})$. For the no-delay case (i.e., $\tau = 0$), the assumption $F(\mathbf{x}) \le C_F, \forall\, \mathbf{x} \in \mathrm{dom}(r)$ is not needed. The delay case has the same-order convergence as the no-delay case. However, their constants are different. Compared to the no-delay case, the delay one has a few additional terms dependent on τ. The term dependent on τ in the second line on the right-hand side of (3.21) is negligible if K is a large number, but the term in the third line will not vanish as $K \to \infty$. In other words, the delay always has a non-negligible effect on the convergence rate. To take a clearer look at the effect, let $\overline{\rho} = 2\rho$, $\beta = 0$, and $K \to \infty$. Then $\gamma \to \frac{1}{2}$, and the terms enclosed in the big square brackets of (3.21) roughly equal $\phi_{1/\overline{\rho}}(\mathbf{x}^{(1)}) - \phi^* + 8\rho\alpha^2M^2 + 4\rho\alpha^2M^2\tau$. Hence, the delay can slow down the convergence rate by $\frac{\tau}{\tau+2+(\phi_{1/\overline{\rho}}(\mathbf{x}^{(1)})-\phi^*)/(4\rho\alpha^2M^2)}$. This indicates that the delay will have a smaller effect if ρ is smaller (i.e., F is closer to convexity) or if α is smaller (i.e., a smaller learning rate is used).

3.2.2. Convergence rate with varying stepsizes. When α_k varies with k, $\sum_k \alpha_k\big(\phi(\mathbf{x}^{(k)}) - \phi(\mathbf{x}^{(k+1)})\big)$ may not be a telescoping series any more, so we cannot directly obtain a bound as in (3.19) by summing up (3.9). Below we make an additional assumption and show a bound on $\sum_{k=1}^K \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2$ when $\alpha_k = \alpha/\sqrt{k}$ for all $k \ge 1$.

Assumption 5. At least one of the following conditions holds.

1. ϕ is bounded on $\mathrm{dom}(r)$, i.e., there is C_ϕ such that $|\phi(\mathbf{x})| \le C_\phi, \forall\, \mathbf{x} \in \mathrm{dom}(r)$.
2. The function r has the form of $r = r_1 + r_2$, where r_1 is the indicator function of a closed convex set $X \subseteq \mathbb{R}^n$, and r_2 is convex. In addition, there is $M_r \ge 0$ such that $\|\mathbf{v}\| \le M_r$ for all $\mathbf{x} \in X$ and all $\mathbf{v} \in \partial r_2(\mathbf{x})$.

In condition 1 of Assumption 5, the boundedness of ϕ can be guaranteed if ϕ is continuous and $\mathrm{dom}(r)$ is compact. The second condition trivially holds if $r_2 \equiv 0$, and it also holds if $X = \mathbb{R}^n$ and r_2 is a Lipschitz continuous function such as a certain norm.

Lemma 3.10. *Under Assumptions 1 and 3, let* $\{\alpha_k\}$ *be a positive nonincreasing sequence and* $\alpha_1 < \frac{2}{\rho}$. *Also, let* $\beta_k \le \tilde{\beta}, \forall\, k \ge 1$ *for some* $\tilde{\beta}$ *such that* $0 \le \tilde{\beta} < 1 - \frac{\alpha_1\rho}{2}$. *Then if the first condition in Assumption 5 holds, we have for any positive integer* K,

$$\textstyle\sum_{k=1}^K \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \le \frac{2\alpha_1 C_\phi}{\gamma} + \sum_{k=1}^K \frac{\alpha_k^2 M^2}{\gamma^2}, \ \text{where } \gamma = \tfrac{1}{2}\big(1 - \tfrac{\alpha_1\rho}{2} - \tilde{\beta}\big). \tag{3.22}$$

Proof. When condition 1 of Assumption 5 holds, i.e., $|\phi(\mathbf{x})| \le C_\phi, \forall\, \mathbf{x} \in \mathrm{dom}(r)$, we have from the

nonincreasing monotonicity of α_k that

$$\begin{aligned}\sum_{k=1}^K \alpha_k\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big) &= \alpha_1\phi(\mathbf{x}^{(1)})+\sum_{k=2}^K(\alpha_k-\alpha_{k-1})\phi(\mathbf{x}^{(k)})-\alpha_K\phi(\mathbf{x}^{(K+1)})\\ &\le \alpha_1 C_\phi-\sum_{k=2}^K(\alpha_k-\alpha_{k-1})C_\phi+\alpha_K C_\phi = 2\alpha_1 C_\phi.\end{aligned}\tag{3.23}$$

Hence, let $\gamma=\frac{1}{2}\big(1-\frac{\alpha_1\rho}{2}-\tilde\beta\big)$ in (3.9) and sum it up over k. We have by $\gamma\le 1-\gamma-\frac{\alpha_k\rho}{2}-\frac{\beta_k+\beta_{k+1}}{2},\forall\, k\ge 1$ that

$$\gamma\sum_{k=1}^K\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\le 2\alpha_1 C_\phi+\sum_{k=1}^K\frac{\alpha_k^2M^2}{\gamma},$$

which apparently implies the desired result. □

Lemma 3.11. *Suppose that Assumption 3 and condition 2 of Assumption 5 hold. Let $\{\mathbf{x}^{(k)}\}$ be the sequence from Alg. 1 with a stepsize sequence $\{\alpha_k\}$ and inertial parameter $\{\beta_k\}$ such that $\beta_k\le\tilde\beta<1$. Then for any positive integer K,*

$$\sum_{k=1}^K\mathbb{E}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\le(M_r^2+M^2)\tfrac{4(1+\tilde\beta^2)}{(1-\tilde\beta^2)^2}\sum_{k=1}^K\alpha_k^2.\tag{3.24}$$

We still need to bound $\sum_{k=k_0}^K\alpha_k\overline{\rho}(1-\alpha_k\overline{\rho})\mathbb{E}[\mathcal{E}_k]$ in (3.13).

Lemma 3.12. *Under Assumptions 1–5, let $\overline{\rho}\in(\rho,2\rho]$ and $\alpha_k=\frac{\alpha}{\sqrt{k}},\forall\, k\ge1$ for some $0<\alpha\le 1/\overline{\rho}$. Also, let $\beta_k=\min\big\{\tilde\beta,\frac{\beta}{k^{1/4}}\big\},\forall\, k$, for some $\tilde\beta$ such that $0\le\tilde\beta<1-\frac{\alpha\rho}{2}$. Furthermore, assume $|F(\mathbf{x})|\le C_F,\forall\,\mathbf{x}\in\mathrm{dom}(r)$. Then for any integer K and $1\le k_0\le K$, it holds*

$$\begin{aligned}\sum_{k=k_0}^K\alpha_k\overline{\rho}(1-\alpha_k\overline{\rho})\mathbb{E}[\mathcal{E}_k]\le{}& 2\alpha_{k_0}\overline{\rho}\tau C_F+\alpha_{k_0}\tau^2\overline{\rho}\big(\tfrac{\rho}{2}+\tfrac{\rho^2}{\overline{\rho}-\rho}\big)\big(C_1+C_2\alpha^2(1+\ln K)\big)\\ &+M\tau\overline{\rho}\sqrt{\textstyle\sum_{k=k_0}^K\alpha_k^2}\sqrt{C_1+C_2\alpha^2(1+\ln K)},\end{aligned}\tag{3.25}$$

where $\mathcal{E}_k$ is defined in (3.15), and C_1 and C_2 are given in (3.27) below.

Now we are ready to show the convergence rate result for the case with varying stepsize.

Theorem 3.13 (convergence rate with varying stepsize). *Under the same assumptions of Lemma 3.12, let $\{\mathbf{x}^{(k)}\}$ be the sequence from Alg. 1. We have*

$$\begin{aligned}\mathbb{E}\big\|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(T)})\big\|^2\le{}&\tfrac{\overline{\rho}}{(\overline{\rho}-\rho)\alpha(\sqrt{K+1}-\sqrt{k_0})}\Big[\mathbb{E}\big[\phi_{1/\overline{\rho}}(\mathbf{x}^{(k_0)})-\phi^*\big]+4\overline{\rho}M^2\alpha^2(1+\ln K-\ln k_0)\\ &+\tfrac{\overline{\rho}}{2}\big(2\tilde\beta^2+\tfrac{2\beta^2}{\alpha(\overline{\rho}-\rho)}\big)\big(C_1+C_2\alpha^2(1+\ln K)\big)\\ &+2\tfrac{\alpha}{\sqrt{k_0}}\overline{\rho}\tau C_F+\tfrac{\alpha}{\sqrt{k_0}}\tau^2\overline{\rho}\big(\tfrac{\rho}{2}+\tfrac{\rho^2}{\overline{\rho}-\rho}\big)\big(C_1+C_2\alpha^2(1+\ln K)\big)\\ &+\alpha M\tau\overline{\rho}\sqrt{1+\ln K-\ln k_0}\sqrt{C_1+C_2\alpha^2(1+\ln K)}\Big],\end{aligned}\tag{3.26}$$

where T is randomly selected from $\{k_0,\dots,K\}$ by (3.14) and

$$C_1=\tfrac{4\alpha C_\phi}{1-\frac{\alpha\rho}{2}-\tilde\beta},\quad C_2=\tfrac{4M^2}{(1-\frac{\alpha\rho}{2}-\tilde\beta)^2},\ \textit{if condition 1 of Assumption 5 holds; or,}\tag{3.27a}$$

$$C_1=0,\quad C_2=(M_r^2+M^2)\tfrac{4(1+\tilde\beta^2)}{(1-\tilde\beta^2)^2},\ \textit{if condition 2 of Assumption 5 holds.}\tag{3.27b}$$

Proof. By the choice of $\{\alpha_k\}$ and $\{\beta_k\}$, we have

$$\textstyle\sum_{k=k_0}^{K}(2+\frac{2}{\alpha_k(\overline{\rho}-\rho)})\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2 \le (2\tilde{\beta}^2+\frac{2\beta^2}{\alpha(\overline{\rho}-\rho)})\textstyle\sum_{k=k_0}^{K}\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2,$$

which, together with (3.22) and (3.24) and also Lemma A.2 with $a=1$, gives

$$\begin{aligned}\tfrac{\overline{\rho}}{2}\textstyle\sum_{k=k_0}^{K}(2+\frac{2}{\alpha_k(\overline{\rho}-\rho)})\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2 &\le \tfrac{\overline{\rho}}{2}(2\tilde{\beta}^2+\tfrac{2\beta^2}{\alpha(\overline{\rho}-\rho)})\big(C_1+C_2\textstyle\sum_{k=1}^{K}\alpha_k^2\big)\\ &\le \tfrac{\overline{\rho}}{2}(2\tilde{\beta}^2+\tfrac{2\beta^2}{\alpha(\overline{\rho}-\rho)})\big(C_1+C_2\alpha^2(1+\ln K)\big),\end{aligned} \tag{3.28}$$

with C_1 and C_2 defined in (3.27). In addition, $\sum_{k=k_0}^{K}\alpha_k \ge \alpha\int_{k_0}^{K+1}\frac{1}{\sqrt{x}}dx = 2\alpha(\sqrt{K+1}-\sqrt{k_0})$ and $\sum_{k=k_0}^{K}\alpha_k^2 \le \alpha^2+\alpha^2\int_{k_0}^{K}\frac{1}{x}dx = \alpha^2(1+\ln K-\ln k_0)$. Hence, substituting (3.25) and (3.28) into (3.13) gives the desired result. □

Remark 3.14. For the no-delay case (i.e., $\tau=0$), we can set $k_0=1$ in Theorem 3.13; then the assumption $|F(\mathbf{x})|\le C_F, \forall\,\mathbf{x}\in\mathrm{dom}(r)$ is not needed anymore. When $\tau>0$, the negative effect by the delay will not vanish as $K\to\infty$, similar to what we observe for the result in Theorem 3.8. Suppose that we have an estimate on τ and $K\gg\tau^4$. We can set $k_0=\Omega(\tau^4)$. Then the terms caused by the delay will near-linearly depend on τ.

4. Convergence analysis for nonconvex composite problems. In this section, we analyze Alg. 1 for problems in the form of (1.1), where F is smooth and r is a possibly nonsmooth convex function. Instead of the ρ-weak convexity, we assume the ρ-smoothness condition on F. Here, we abuse the notation of ρ, which is used as the weak-convexity constant in the previous section. Nevertheless, if F is ρ-smooth, it is also ρ-weakly convex. The stronger assumption will enable us to obtain better convergence result in terms of the effect caused by the staleness of the gradient.

Assumption 6 (ρ-smoothness). $F(\mathbf{x})$ is ρ-smooth in $\mathrm{dom}(r)$, i.e.,

$$\|\nabla F(\mathbf{x})-\nabla F(\mathbf{y})\|\le\rho\|\mathbf{x}-\mathbf{y}\|, \forall\,\mathbf{x},\mathbf{y}\in\mathrm{dom}(r).$$

When F is smooth, it is standard to replace Assumption 3 by the one below.

Assumption 7 (bounded variance). There is $\sigma\ge 0$ such that $\mathbb{E}_\xi\|\nabla f(\mathbf{x};\xi)-\nabla F(\mathbf{x})\|^2\le\sigma^2$ for all $\mathbf{x}\in\mathrm{dom}\,r$.

In addition, when F is smooth, we only need a boundedness condition on the staleness but not a static distribution anymore.

Assumption 8 (bounded staleness). There is a finite integer τ such that $\tau_k\le\tau$ for all $k\ge 1$.

We can track the delay and ensure the boundedness of delay by discarding too outdated sample gradients.

Lemma 4.1. *Under Assumptions 2, 6, and 7, the iterates $\{\mathbf{x}^{(k)}\}$ from Algorithm 1 satisfy*

$$\mathbb{E}_{\xi_k}\|\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k)})\|^2\le\sigma^2+\rho^2\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2.$$

Proof. When F is differentiable, the condition in Assumption 2 becomes $\mathbb{E}_{\xi_k}[\mathbf{g}^{(k)}]=\nabla F(\mathbf{x}^{(k-\tau_k)})$. Hence,

$$\begin{aligned}\mathbb{E}_{\xi_k}\|\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k)})\|^2 &= \mathbb{E}_{\xi_k}\|\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k-\tau_k)})+\nabla F(\mathbf{x}^{(k-\tau_k)})-\nabla F(\mathbf{x}^{(k)})\|^2\\ &= \mathbb{E}_{\xi_k}\|\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k-\tau_k)})\|^2+\|\nabla F(\mathbf{x}^{(k-\tau_k)})-\nabla F(\mathbf{x}^{(k)})\|^2\\ &\le \sigma^2+\rho^2\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2,\end{aligned}$$

where the second equality follows from $\mathbb{E}_{\xi_k}\langle \mathbf{g}^{(k)} - \nabla F(\mathbf{x}^{(k-\tau_k)}), \nabla F(\mathbf{x}^{(k-\tau_k)}) - \nabla F(\mathbf{x}^{(k)})\rangle = 0$, and the inequality holds by using Assumptions 6 and 7. This completes the proof. □

Lemma 4.2. *Under Assumptions 2, 6 and 7, let $\overline{\rho} > \rho$ and $\alpha_k \in (0, 1/\overline{\rho}]$ for all k. Then the iterates $\{\mathbf{x}^{(k)}\}$ from Algorithm 1 with a stepsize sequence $\{\alpha_k\}$ satisfies*

$$\begin{aligned}\mathbb{E}_{\xi_k}\|\mathbf{x}^{(k+1)} - \widetilde{\mathbf{x}}^{(k)}\|^2 \le \|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 - \big(\tfrac{1}{2}\alpha_k(\overline{\rho}-\rho) - c_k\big)\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 \\ + (2 + \tfrac{1}{c_k})\beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + \alpha_k^2\sigma^2 + 2\big(\alpha_k^2 + \tfrac{\alpha_k}{\overline{\rho}-\rho}\big)\rho^2\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2,\end{aligned} \tag{4.1}$$

where c_k is any positive number, and $\widetilde{\mathbf{x}}^{(k)}$ is defined in (3.1a).

Using the previous two lemmas, we show a convergence result below for generic parameters.

Theorem 4.3. *Under Assumptions 2, 6 and 7, let $\overline{\rho} > \rho$ and $\alpha_k \in (0, 1/\overline{\rho}]$ for all $k \ge 1$. Given a positive integer K, let $\{\mathbf{x}^{(k)}\}_{k=1}^K$ be the sequence generated from Algorithm 1 with a stepsize sequence $\{\alpha_k\}_{k=1}^K$ and inertial parameter $\{\beta_k\}$. Then*

$$\begin{aligned}\mathbb{E}\|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(T)})\|^2 \le \tfrac{8\overline{\rho}}{(\overline{\rho}-\rho)\sum_{k=1}^K \alpha_k}\Big[\tfrac{\overline{\rho}}{2}\textstyle\sum_{k=1}^K \big(2 + \tfrac{4}{\alpha_k(\overline{\rho}-\rho)}\big)\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 \\ \phi_{1/\overline{\rho}}(\mathbf{x}^{(1)}) - \phi^* + \tfrac{\sigma^2\overline{\rho}}{2}\textstyle\sum_{k=1}^K \alpha_k^2 + \overline{\rho}\rho^2\sum_{k=1}^K \big(\alpha_k^2 + \tfrac{\alpha_k}{\overline{\rho}-\rho}\big)\mathbb{E}\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2\Big],\end{aligned} \tag{4.2}$$

where T is randomly selected from $\{1, \ldots, K\}$ by (3.14).

Proof. By the definition of ϕ_λ in (2.3) and Lemma 4.2, we have

$$\begin{aligned}&\mathbb{E}_{\xi_k}\big[\phi_{1/\overline{\rho}}(\mathbf{x}^{(k+1)})\big] \\ &\le \mathbb{E}_{\xi_k}\big[\phi(\widetilde{\mathbf{x}}^{(k)}) + \tfrac{\overline{\rho}}{2}\|\mathbf{x}^{(k+1)} - \widetilde{\mathbf{x}}^{(k)}\|^2\big] \\ &\le \phi(\widetilde{\mathbf{x}}^{(k)}) + \tfrac{\overline{\rho}}{2}\Big[\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 - (\tfrac{1}{2}\alpha_k(\overline{\rho}-\rho) - c_k)\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 \\ &\quad + \big(2 + \tfrac{1}{c_k}\big)\beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + \alpha_k^2\sigma^2 + 2\big(\alpha_k^2 + \tfrac{\alpha_k}{\overline{\rho}-\rho}\big)\rho^2\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2\Big] \\ &= \phi_{1/\overline{\rho}}(\mathbf{x}^{(k)}) - \tfrac{\overline{\rho}}{2}\big(\tfrac{1}{2}\alpha_k(\overline{\rho}-\rho) - c_k\big)\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + \tfrac{\overline{\rho}}{2}\big(2 + \tfrac{1}{c_k}\big)\beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 \\ &\quad + \tfrac{\overline{\rho}\alpha_k^2\sigma^2}{2} + 2\big(\alpha_k^2 + \tfrac{\alpha_k}{\overline{\rho}-\rho}\big)\tfrac{\overline{\rho}\rho^2}{2}\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2.\end{aligned} \tag{4.3}$$

Take full expectation on both sides of (4.3) and sum up it over $k = 1, \ldots, K$. Then we have

$$\begin{aligned}&\mathbb{E}\big[\phi_{1/\overline{\rho}}(\mathbf{x}^{(K+1)})\big] \\ &\le \mathbb{E}\big[\phi_{1/\overline{\rho}}(\mathbf{x}^{(1)})\big] - \tfrac{\overline{\rho}}{2}\textstyle\sum_{k=1}^K \big(\tfrac{1}{2}\alpha_k(\overline{\rho}-\rho) - c_k\big)\mathbb{E}\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + \tfrac{\overline{\rho}}{2}\textstyle\sum_{k=1}^K \big(2 + \tfrac{1}{c_k}\big)\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 \\ &\quad + \tfrac{\sigma^2\overline{\rho}}{2}\textstyle\sum_{k=1}^K \alpha_k^2 + \overline{\rho}\rho^2\sum_{k=1}^K \big(\alpha_k^2 + \tfrac{\alpha_k}{\overline{\rho}-\rho}\big)\mathbb{E}\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2.\end{aligned} \tag{4.4}$$

Choose $c_k = \frac{1}{4}\alpha_k(\overline{\rho}-\rho)$ for all k and replace $\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2$ by $\frac{1}{\overline{\rho}^2}\|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(k)})\|^2$ from Lemma 2.4. We have from (4.4) that

$$\begin{aligned}&\mathbb{E}\big[\phi_{1/\overline{\rho}}(\mathbf{x}^{(K+1)})\big] \\ &\le \mathbb{E}\big[\phi_{1/\overline{\rho}}(\mathbf{x}^{(1)})\big] - \tfrac{1}{8\overline{\rho}}\textstyle\sum_{k=1}^K \alpha_k(\overline{\rho}-\rho)\mathbb{E}\|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(k)})\|^2 + \tfrac{\overline{\rho}}{2}\textstyle\sum_{k=1}^K \big(2 + \tfrac{4}{\alpha_k(\overline{\rho}-\rho)}\big)\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 \\ &\quad + \tfrac{\sigma^2\overline{\rho}}{2}\textstyle\sum_{k=1}^K \alpha_k^2 + \overline{\rho}\rho^2\sum_{k=1}^K \big(\alpha_k^2 + \tfrac{\alpha_k}{\overline{\rho}-\rho}\big)\mathbb{E}\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2.\end{aligned} \tag{4.5}$$

Rearrange terms in (4.5) and notice $\phi_{1/\overline{\rho}}(\mathbf{x}^{(K+1)}) \geq \phi^*$, we obtain the desired result by the definition of T. □

To show the convergence rate, we still need the following result to bound $\sum_{k\geq 1}\mathbb{E}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2$.

Lemma 4.4. *Let $\{\mathbf{x}^{(k)}\}$ be generated from Alg. 1. Under Assumptions 6 and 7, it holds for any $\gamma > 0$,*

$$(1-\gamma-\tfrac{\alpha_k\rho}{2}-\tfrac{\beta_k}{2})\mathbb{E}_{\xi_k}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2 \leq \alpha_k\mathbb{E}_{\xi_k}\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big)+\tfrac{\beta_k}{2}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2 + \tfrac{\alpha_k^2}{2\gamma}\left(\rho^2\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-\tau_k)}\|^2+\sigma^2\right). \tag{4.6}$$

Proof. By the ρ-smoothness of F and $\alpha_k > 0$, it holds

$$\alpha_k\big(F(\mathbf{x}^{(k+1)})-F(\mathbf{x}^{(k)})\big) \leq \alpha_k\left(\langle \mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}, \nabla F(\mathbf{x}^{(k)})\rangle + \tfrac{\rho}{2}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\right). \tag{4.7}$$

Also notice that (3.10) still holds. Hence, we obtain, by adding (3.10) and (4.7) and rearranging terms, that

$$(1-\tfrac{\alpha_k\rho}{2})\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2 \leq \alpha_k\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big)+\alpha_k\big\langle \mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}, \nabla F(\mathbf{x}^{(k)})-\mathbf{g}^{(k)}\big\rangle + \beta_k\big\langle \mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}, \mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\big\rangle. \tag{4.8}$$

By the Young's inequality, we have for any $\gamma > 0$,

$$\alpha_k\big\langle \mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}, \nabla F(\mathbf{x}^{(k)})-\mathbf{g}^{(k)}\big\rangle \leq \gamma\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2+\tfrac{\alpha_k^2}{4\gamma}\|\nabla F(\mathbf{x}^{(k)})-\mathbf{g}^{(k)}\|^2, \tag{4.9}$$

and

$$\beta_k\big\langle \mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}, \mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\big\rangle \leq \tfrac{\beta_k}{2}\left(\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2+\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2\right). \tag{4.10}$$

Plugging (4.9) and (4.10) into (4.8) gives

$$(1-\tfrac{\alpha_k\rho}{2})\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2 \leq \alpha_k\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big)+\gamma\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2 + \tfrac{\alpha_k^2}{4\gamma}\|\nabla F(\mathbf{x}^{(k)})-\mathbf{g}^{(k)}\|^2+\tfrac{\beta_k}{2}\left(\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2+\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2\right). \tag{4.11}$$

Now notice $\mathbb{E}_{\xi_k}\|\nabla F(\mathbf{x}^{(k)})-\mathbf{g}^{(k)}\|^2 \leq 2\|\nabla F(\mathbf{x}^{(k)})-\nabla F(\mathbf{x}^{(k-\tau_k)})\|^2+2\mathbb{E}_{\xi_k}\|\nabla F(\mathbf{x}^{(k-\tau_k)})-\mathbf{g}^{(k)}\|^2$ and use Assumptions 6 and 7. We obtain the desired result by taking a conditional expectation about ξ_k over both sides of (4.11) and rearranging terms. □

Now we are ready to show the convergence rate result.

Theorem 4.5 (convergence rate with fixed stepsize). *Under Assumptions 2, 6, 7 and 8, let $\overline{\rho} > \rho$ and K be the maximum number of iterations. Choose $\alpha_k = \frac{\alpha}{\sqrt{K}}$ and $\beta_k = \frac{\beta}{K^{1/4}}$ for some $\alpha > 0$ and $\beta \geq 0$ such that $\tilde{\gamma} := \frac{1}{2}-\frac{\alpha\rho}{2\sqrt{K}}-\frac{\tau\alpha^2\rho^2}{K}-\frac{\beta}{K^{1/4}} > 0$. Let $\{\mathbf{x}^{(k)}\}$ be the sequence from Alg. 1. Then*

$$\mathbb{E}\|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(T)})\|^2 \leq \tfrac{8\overline{\rho}}{(\overline{\rho}-\rho)\alpha\sqrt{K}}\Bigg[\phi_{1/\overline{\rho}}(\mathbf{x}^{(1)})-\phi^*+\tfrac{\sigma^2\overline{\rho}\alpha^2}{2} + \tfrac{1}{\tilde{\gamma}}\left(\overline{\rho}\big(1+\tfrac{2\sqrt{K}}{\alpha(\overline{\rho}-\rho)}\big)\tfrac{\beta^2}{\sqrt{K}}+\tau^2\overline{\rho}\rho^2\big(\tfrac{\alpha^2}{K}+\tfrac{\alpha}{(\overline{\rho}-\rho)\sqrt{K}}\big)\right)\big(\tfrac{\alpha}{\sqrt{K}}\mathbb{E}(\phi(\mathbf{x}^{(1)})-\phi^*)+\alpha^2\sigma^2\big)\Bigg], \tag{4.12}$$

where T is randomly selected from $\{1,\ldots,K\}$ by (3.14) with $k_0 = 1$.

Proof. With $\alpha_k = \frac{\alpha}{\sqrt{K}}$ and $\beta_k = \frac{\beta}{K^{1/4}}$, we take full expectation over (4.6) with $\gamma = \frac{1}{2}$ and sum it up over $k = 1$ through K to have

$$\begin{aligned}(\tfrac{1}{2} - \tfrac{\alpha\rho}{2\sqrt{K}} - \tfrac{\beta}{2K^{1/4}}) \textstyle\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \le \tfrac{\alpha}{\sqrt{K}}\mathbb{E}\big(\phi(\mathbf{x}^{(1)}) - \phi(\mathbf{x}^{(K+1)})\big) \\ + \tfrac{\beta}{2K^{1/4}} \textstyle\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + \tfrac{\alpha^2}{K} \textstyle\sum_{k=1}^K \big(\rho^2 \mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}\|^2 + \sigma^2\big).\end{aligned} \tag{4.13}$$

Notice that $\mathbf{x}^{(0)} = \mathbf{x}^{(1)}$ and by Assumption 7, it holds

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}\|^2 \le \tau \textstyle\sum_{j=1}^{\tau} \|\mathbf{x}^{(k+1-j)} - \mathbf{x}^{(k-j)}\|^2. \tag{4.14}$$

Hence, we have from (4.13) by rearranging terms and using $\phi(\mathbf{x}) \ge \phi^*, \forall\, \mathbf{x} \in \mathrm{dom}(r)$ that

$$(\tfrac{1}{2} - \tfrac{\alpha\rho}{2\sqrt{K}} - \tfrac{\tau^2\alpha^2\rho^2}{K} - \tfrac{\beta}{K^{1/4}}) \textstyle\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \le \tfrac{\alpha}{\sqrt{K}}\mathbb{E}\big(\phi(\mathbf{x}^{(1)}) - \phi^*\big) + \alpha^2\sigma^2. \tag{4.15}$$

Therefore,

$$\begin{aligned}&\tfrac{\overline{\rho}}{2} \textstyle\sum_{k=1}^K \big(2 + \tfrac{4}{\alpha_k(\overline{\rho}-\rho)}\big)\beta_k^2 \mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + \tau^2\overline{\rho}\rho^2 \textstyle\sum_{k=1}^K \big(\alpha_k^2 + \tfrac{\alpha_k}{\overline{\rho}-\rho}\big)\mathbb{E}\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2 \\ \le\ &\Big(\tfrac{\overline{\rho}}{2}\big(2 + \tfrac{4\sqrt{K}}{\alpha(\overline{\rho}-\rho)}\big)\tfrac{\beta^2}{\sqrt{K}} + \tau^2\overline{\rho}\rho^2\big(\tfrac{\alpha^2}{K} + \tfrac{\alpha}{(\overline{\rho}-\rho)\sqrt{K}}\big)\Big) \textstyle\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 \\ \le\ &\tfrac{1}{\tilde{\gamma}}\Big(\tfrac{\overline{\rho}}{2}\big(2 + \tfrac{4\sqrt{K}}{\alpha(\overline{\rho}-\rho)}\big)\tfrac{\beta^2}{\sqrt{K}} + \tau^2\overline{\rho}\rho^2\big(\tfrac{\alpha^2}{K} + \tfrac{\alpha}{(\overline{\rho}-\rho)\sqrt{K}}\big)\Big)\big(\tfrac{\alpha}{\sqrt{K}}\mathbb{E}\big(\phi(\mathbf{x}^{(1)}) - \phi^*\big) + \alpha^2\sigma^2\big),\end{aligned} \tag{4.16}$$

where the first inequality follows from (4.14), and the second inequality is from (4.15) and the definition of $\tilde{\gamma}$. Now plug (4.16) and the choice of $\{\alpha_k\}$ into (4.2) to obtain the desired result. □

Remark 4.6. We make a few remarks here about Theorem 4.5. First, in the proof, we take $\gamma = \frac{1}{2}$ for simplicity while using (4.6). The analysis goes through for any $\gamma > 0$ such that $1 - \gamma - \frac{\alpha\rho}{2\sqrt{K}} - \frac{\tau^2\alpha^2\rho^2}{2\gamma K} - \frac{\beta}{K^{1/4}} > 0$. Second, we see from (4.12) that a positive τ will slow down the convergence but its effect will be reduced in an order of $K^{-\frac{1}{4}}$. Hence, if K is big enough such that $K^{1/4} \gg \tau$, then the effect caused by the staleness is negligible.

The $O(\frac{1}{\sqrt{K}})$ convergence above is established by using a fixed stepsize sequence. We can show a similar result for the choice of $\alpha_k = \Theta(\frac{1}{\sqrt{k}})$ by assuming condition 1 of Assumption 5. The proof is given in Appendix B.

Theorem 4.7 (convergence rate with varying stepsize). *Suppose Assumptions 2, 6, 7 and 8, and also condition 1 of Assumption 5 hold. Let $\overline{\rho} > \rho$, $\alpha_k = \frac{\alpha}{\sqrt{k+a-1}}$ and $\beta_k = \min\big\{\tilde{\beta}, \frac{\beta}{(k+a-1)^{1/4}}\big\}$, for all $k \ge 1$, for some $\alpha > 0$, $\beta \ge 0$, $\tilde{\beta} \ge 0$, and $a \ge 1$ such that*

$$\tilde{\gamma} := \tfrac{1}{2}\big(1 - \tfrac{\alpha\rho}{\sqrt{a}} - \tilde{\beta}^2 - \tfrac{2\tau^2\rho^2\alpha^2}{a}\big) > 0. \tag{4.17}$$

Let $\{\mathbf{x}^{(k)}\}$ be the sequence from Alg. 1. Then,

$$\begin{aligned}\mathbb{E}\|\nabla\phi_{1/\overline{\rho}}(\mathbf{x}^{(T)})\|^2 \le \tfrac{4\overline{\rho}}{(\overline{\rho}-\rho)\alpha(\sqrt{K+a}-\sqrt{a})}\Bigg[\phi_{1/\overline{\rho}}(\mathbf{x}^{(1)}) - \phi^* + \tfrac{\sigma^2\overline{\rho}\alpha^2}{2}(1 + \ln\tfrac{a+K-1}{a}) \\ + \Big(\overline{\rho}\big(\tilde{\beta}^2 + \tfrac{2\beta^2}{\alpha(\overline{\rho}-\rho)}\big) + \tau^2\overline{\rho}\rho^2\big(\tfrac{\alpha^2}{a} + \tfrac{\alpha}{\sqrt{a}(\overline{\rho}-\rho)}\big)\Big)\tfrac{2}{\tilde{\gamma}}\big(\alpha_1 C_\phi + \sigma^2\alpha^2(1 + \ln\tfrac{a+K-1}{a})\big)\Bigg],\end{aligned} \tag{4.18}$$

where T is randomly selected from $\{1, \ldots, K\}$ by (3.14) with $k_0 = 1$.

Remark 4.8. When there is no delay, i.e., $\tau = 0$, we can choose $a = 1$ and obtain a convergence rate of $\tilde{\Theta}(\frac{1}{\sqrt{K}})$. When there is delay, i.e., $\tau \geq 1$, (4.18) with $a = \Theta(\tau^4)$, which can ensure (4.17), gives a rate of $\widetilde{\Theta}(\frac{1}{\sqrt{K+\tau^4}-\sqrt{\tau^4}}) = \widetilde{\Theta}\Big(\frac{1}{\sqrt{K}}(\sqrt{1+\frac{\tau^4}{K}}+\sqrt{\frac{\tau^4}{K}})\Big)$. In this case, the delay will have a negligible effect on the convergence speed if $\tau = o(K^{\frac{1}{4}})$.

5. Convergence analysis for smooth nonconvex problems. In this section, we consider the case where $r = 0$, i.e., a non-regularized smooth problem. For this special case, we are able to show a stronger result under the same assumptions as we used in section 4, in the sense that the delay has a weaker effect on the convergence speed. However, the analysis is significantly different from those in the previous two sections. Throughout this section, we let

$$\beta_k = \tfrac{\alpha_k}{\alpha_{k-1}}\beta, \text{ for all } k \geq 1 \text{ and for some } \beta \in (0,1). \tag{5.1}$$

Then the update in (1.2) reduces to (1.5) with $\mathbf{m}$-vectors defined in (1.4). We declare the following notation, as they appear extensively in this section:

$$\mathbf{u}^{(k)} = \nabla F(\mathbf{x}^{(k-\tau_k)}) \text{ and } u_k = \mathbb{E}\|\mathbf{u}^{(k)}\|^2 \text{ for all } k \geq 1. \tag{5.2}$$

With the setting in (5.1), we define the following quantities that are critical for bounding the staleness:

$$\theta_{k,j} = \textstyle\sum_{l=0}^{\min\{\tau_k-1,k-j-1\}} \alpha_{k-l-1}\beta^{k-j-l-1}, \text{ and } \pi_{k,j}(t) = \sum_{l=0}^{\min\{\tau_k-1,k-j-1\}} t^{k-j-l-1}. \tag{5.3}$$

Lemma 5.1. *Let $t \in (0,1)$, we have the following results:*

$$\pi_{k,j}(t) = \begin{cases} \frac{1-t^{k-j}}{1-t} & \text{if } j \geq k-\tau_k+1, \\ \frac{1-t^{\tau_k}}{1-t}t^{k-\tau_k-j} & \text{if } j \leq k-\tau_k; \end{cases} \quad \textstyle\sum_{j=1}^{k-1}\pi_{k,j}(t) \leq \frac{\tau}{1-t}; \quad \sum_{j=1}^{k-1}\pi_{k,j}^2(t) \leq \frac{\tau}{(1-t)^2}. \tag{5.4}$$

Lemma 5.2. *Let $\{\mathbf{x}^{(k)}\}_{k\geq 1}$ and $\{\mathbf{m}^{(k)}\}_{k\geq 1}$ be generated from (1.5) and (1.4). Under Assumptions 2 and 7, it holds for $k \geq 1$,*

$$\mathbb{E}\|\mathbf{m}^{(k)}\|^2 \leq (1-\beta)\textstyle\sum_{j=1}^{k}\beta^{k-j}u_j + (1-\beta)^2\sum_{j=1}^{k}\beta^{2(k-j)}u_j + \frac{(1-\beta)^2}{1-\beta^2}\sigma^2, \tag{5.5}$$

$$\mathbb{E}\|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2 \leq \textstyle\sum_{l=1}^{k-1}\theta_{k,l}\sum_{j=1}^{k-1}\theta_{k,j}u_j + \sum_{j=1}^{k-1}\theta_{k,j}^2u_j + \sigma^2\sum_{j=1}^{k-1}\theta_{k,j}^2. \tag{5.6}$$

In the remaining analysis, we follow the analytical framework of [67]. We define an auxiliary sequence $\mathbf{z}^{(k)}$ as follows:

$$\mathbf{z}^{(k)} = \mathbf{x}^{(k)} + \tfrac{\beta}{1-\beta}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = \tfrac{1}{1-\beta}\mathbf{x}^{(k)} - \tfrac{\beta}{1-\beta}\mathbf{x}^{(k-1)}, \forall\, k \geq 1. \tag{5.7}$$

Recall $\mathbf{x}^{(0)} = \mathbf{x}^{(1)}$, so clearly, $\mathbf{z}^{(1)} = \mathbf{x}^{(1)}$.

Lemma 5.3. *Let $\mathbf{z}^{(k)}$ be defined as in (5.7) and $\alpha_0 = \alpha_1$. We have for $k \geq 1$,*

$$\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)} = \tfrac{\beta}{1-\beta}(1-\alpha_k/\alpha_{k-1})(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}) - \tfrac{\alpha_k}{1-\beta}\mathbf{g}^{(k)}, \tag{5.8}$$

and

$$\|\nabla F(\mathbf{z}^{(k)}) - \nabla F(\mathbf{x}^{(k)})\| \leq \tfrac{\rho\beta}{1-\beta}\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\|. \tag{5.9}$$

Now we are ready to show the main result. We first show the convergence by imposing general conditions on $\{\alpha_k\}$ and then specify the choice of the parameters that satisfies the imposed conditions.

Theorem 5.4. *Given a maximum number K of iterations, let $\{\mathbf{x}^{(k)}\}_{k=1}^K$ be generated from Alg. 1 with a non-increasing positive sequence $\{\alpha_k\}_{k=1}^K$. Let $\bar{\mathbf{x}}^{(K)}$ be drawn from $\{\mathbf{x}^{(k)}\}_{k=1}^K$ with probability*

$$\text{Prob}(\bar{\mathbf{x}}^{(K)} = \mathbf{x}^{(k)}) = \frac{\alpha_k}{\sum_{j=1}^K \alpha_j}, \ \forall k = 1, \dots, K. \tag{5.10}$$

Under Assumptions 2 and 6–8, if for all $k \geq 2$,

$$(1 - \alpha_k/\alpha_{k-1})^2 \leq \frac{\alpha_k}{2(1-\beta)}, \tag{5.11}$$

and for all $j \geq 1$,

$$\frac{3\rho\alpha_j}{1-\beta} + \rho^2 \left[\frac{\tau(\tau-1)\alpha_1\alpha_j}{(1-\beta)^2} + \frac{(\tau-1)\alpha_j^2}{(1-\beta)^2} + \frac{\tau\alpha_j^2}{(1-\beta)^3} + \frac{\alpha_j^2}{(1-\beta)^2(1-\beta^2)} \right] + \frac{2(1+5\rho)\beta^2}{(1-\beta)^2(1-\beta^2)}\alpha_j \leq 1, \tag{5.12}$$

then it holds

$$\begin{aligned} \mathbb{E}\|\nabla F(\bar{\mathbf{x}}^{(K)})\|^2 \leq \frac{4\sigma^2}{(1-\beta)\sum_{k=1}^K \alpha_k} \Bigg[& \frac{\rho^2\tau}{2(1-\beta)} \sum_{k=1}^K \alpha_k \alpha^2_{\max\{k-\tau_k,1\}} + \frac{(1+5\rho)\beta^2}{2(1-\beta^2)} \sum_{k=2}^K \alpha_{k-1}^2 \\ & + \rho \sum_{k=1}^K \alpha_k^2 \Bigg] + \frac{4(1-\beta)\left[F(\mathbf{x}^{(1)}) - \inf_{\mathbf{x}} F(\mathbf{x})\right]}{\sum_{k=1}^K \alpha_k}. \end{aligned} \tag{5.13}$$

Proof. By the ρ-smoothness of F, it follows from (2.5) that

$$\begin{aligned} 0 &\leq F(\mathbf{z}^{(k)}) - F(\mathbf{z}^{(k+1)}) + \nabla F(\mathbf{z}^{(k)})^\top (\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}) + \tfrac{\rho}{2}\|\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}\|^2 \\ &= F(\mathbf{z}^{(k)}) - F(\mathbf{z}^{(k+1)}) + \nabla F(\mathbf{x}^{(k)})^\top (\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}) \\ &\quad + (\nabla F(\mathbf{z}^{(k)}) - \nabla F(\mathbf{x}^{(k)}))^\top (\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}) + \tfrac{\rho}{2}\|\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}\|^2. \end{aligned} \tag{5.14}$$

Taking the conditional expectation and using (5.8) and Assumption 2, we have from (5.14) that

$$\begin{aligned} 0 \leq\ & \mathbb{E}_k[F(\mathbf{z}^{(k)}) - F(\mathbf{z}^{(k+1)})] + \nabla F(\mathbf{x}^{(k)})^\top \left(\tfrac{\beta}{1-\beta}(1 - \alpha_k/\alpha_{k-1})(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}) - \tfrac{\alpha_k}{1-\beta}\mathbf{u}^{(k)} \right) \\ & + (\nabla F(\mathbf{z}^{(k)}) - \nabla F(\mathbf{x}^{(k)}))^\top \left(\tfrac{\beta}{1-\beta}(1 - \alpha_k/\alpha_{k-1})(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}) - \tfrac{\alpha_k}{1-\beta}\mathbf{u}^{(k)} \right) \\ & + \tfrac{\rho}{2}\mathbb{E}_k \left\| \tfrac{\beta}{1-\beta}(1 - \alpha_k/\alpha_{k-1})(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}) - \tfrac{\alpha_k}{1-\beta}\mathbf{g}^{(k)} \right\|^2. \end{aligned} \tag{5.15}$$

We bound the right-hand side of (5.15) as follows:

- in the first line of (5.15), applying the Cauchy-Schwarz inequality gives

$$\nabla F(\mathbf{x}^{(k)})^\top \tfrac{\beta}{1-\beta}(1-\alpha_k/\alpha_{k-1})(\mathbf{x}^{(k-1)}-\mathbf{x}^{(k)}) \leq \tfrac{1}{2}(1-\tfrac{\alpha_k}{\alpha_{k-1}})^2\|\nabla F(\mathbf{x}^{(k)})\|^2 + \tfrac{\beta^2}{2(1-\beta)^2}\|\mathbf{x}^{(k-1)}-\mathbf{x}^{(k)}\|^2;$$

- in the second line of (5.15), it follows from (5.9) and $0 \leq 1 - \frac{\alpha_k}{\alpha_{k-1}} \leq 1$ that

$$(\nabla F(\mathbf{z}^{(k)}) - \nabla F(\mathbf{x}^{(k)}))^\top \tfrac{\beta}{1-\beta}(1 - \alpha_k/\alpha_{k-1})(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}) \leq \tfrac{\rho\beta^2}{(1-\beta)^2}\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\|^2$$

and in addition, by the Cauchy-Schwarz inequality,

$$\begin{aligned} (\nabla F(\mathbf{z}^{(k)}) - \nabla F(\mathbf{x}^{(k)}))^\top (-\tfrac{\alpha_k}{1-\beta}\mathbf{u}^{(k)}) &\leq \rho \cdot \tfrac{\beta}{1-\beta}\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\| \cdot \tfrac{\alpha_k}{1-\beta}\|\mathbf{u}^{(k)}\| \\ &\leq \tfrac{\rho\beta^2}{2(1-\beta)^2}\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\|^2 + \tfrac{\rho\alpha_k^2}{2(1-\beta)^2}\|\mathbf{u}^{(k)}\|^2; \end{aligned}$$

- in the last line of (5.15), using the Young's inequality gives

$$\tfrac{\rho}{2}\mathbb{E}_k\|\tfrac{\beta}{1-\beta}(1-\alpha_k/\alpha_{k-1})(\mathbf{x}^{(k-1)}-\mathbf{x}^{(k)})-\tfrac{\alpha_k}{1-\beta}\mathbf{g}^{(k)}\|^2 \le \tfrac{\rho\beta^2}{(1-\beta)^2}\|\mathbf{x}^{(k-1)}-\mathbf{x}^{(k)}\|^2+\tfrac{\rho\alpha_k^2}{(1-\beta)^2}\mathbb{E}_k\|\mathbf{g}^{(k)}\|^2;$$

- furthermore, by Assumption 7, we have

$$\mathbb{E}_k\|\mathbf{g}^{(k)}\|^2 = \mathbb{E}_k\|\nabla f(\mathbf{x}^{(k-\tau_k)};\xi_k)-\mathbf{u}^{(k)}\|^2+\|\mathbf{u}^{(k)}\|^2 \le \sigma^2+\|\mathbf{u}^{(k)}\|^2.$$

Substitute the above four items into (5.15), combine like terms, and take total expectation. We have

$$\begin{aligned}
0 &\le \mathbb{E}[F(\mathbf{z}^{(k)})-F(\mathbf{z}^{(k+1)})]+\tfrac{1}{2}(1-\tfrac{\alpha_k}{\alpha_{k-1}})^2\mathbb{E}\|\nabla F(\mathbf{x}^{(k)})\|^2-\tfrac{\alpha_k}{1-\beta}\mathbb{E}[\nabla F(\mathbf{x}^{(k)})^\top\mathbf{u}^{(k)}]\\
&\quad+\tfrac{(1+5\rho)\beta^2}{2(1-\beta)^2}\mathbb{E}\|\mathbf{x}^{(k-1)}-\mathbf{x}^{(k)}\|^2+\tfrac{\rho\alpha_k^2\sigma^2}{(1-\beta)^2}+\tfrac{3\rho\alpha_k^2}{2(1-\beta)^2}u_k\\
&= \mathbb{E}[F(\mathbf{z}^{(k)})-F(\mathbf{z}^{(k+1)})]+\frac{1}{2}(1-\frac{\alpha_k}{\alpha_{k-1}})^2\mathbb{E}\|\nabla F(\mathbf{x}^{(k)})\|^2\\
&\quad-\tfrac{\alpha_k}{2(1-\beta)}\left[\mathbb{E}\|\nabla F(\mathbf{x}^{(k)})\|^2+u_k-\mathbb{E}\|\nabla F(\mathbf{x}^{(k-\tau_k)})-\nabla F(\mathbf{x}^{(k)})\|^2\right]\\
&\quad+\tfrac{(1+5\rho)\beta^2}{2(1-\beta)^2}\mathbb{E}\|\mathbf{x}^{(k-1)}-\mathbf{x}^{(k)}\|^2+\tfrac{\rho\alpha_k^2\sigma^2}{(1-\beta)^2}+\tfrac{3\rho\alpha_k^2}{2(1-\beta)^2}u_k,
\end{aligned}\tag{5.16}$$

where the equality is due to $\mathbf{a}^\top\mathbf{b}=\frac{1}{2}[\|\mathbf{a}\|^2+\|\mathbf{b}\|^2-\|\mathbf{a}-\mathbf{b}\|^2]$, for any two vectors $\mathbf{a}$ and $\mathbf{b}$.

Using (1.5) and the smoothness of F and then substituting (5.5) and (5.6) to (5.16), we have

$$\begin{aligned}
0 &\le \mathbb{E}[F(\mathbf{z}^{(k)})-F(\mathbf{z}^{(k+1)})]+\tfrac{1}{2}\left[(1-\tfrac{\alpha_k}{\alpha_{k-1}})^2-\tfrac{\alpha_k}{1-\beta}\right]\mathbb{E}\|\nabla F(\mathbf{x}^{(k)})\|^2\\
&\quad+\tfrac{\alpha_k\rho^2}{2(1-\beta)}\left[\textstyle\sum_{l=1}^{k-1}\theta_{k,l}\sum_{j=1}^{k-1}\theta_{k,j}u_j+\sum_{j=1}^{k-1}\theta_{k,j}^2u_j+\sigma^2\sum_{j=1}^{k-1}\theta_{k,j}^2\right]+\tfrac{1}{2}\big(\tfrac{3\rho\alpha_k^2}{(1-\beta)^2}-\tfrac{\alpha_k}{1-\beta}\big)u_k\\
&\quad+\tfrac{(1+5\rho)\beta^2\alpha_{k-1}^2}{2(1-\beta)^2}\left[\textstyle\sum_{j=1}^{k-1}\big(\tfrac{\beta^{k-j-1}}{1-\beta}+\beta^{2(k-j-1)}\big)u_j+\tfrac{\sigma^2}{1-\beta^2}\right]+\tfrac{\rho\alpha_k^2\sigma^2}{(1-\beta)^2}.
\end{aligned}\tag{5.17}$$

Summing the above inequality over $k=1,\dots,K$ and utilizing (5.11) lead to

$$\begin{aligned}
0 &\le F(\mathbf{x}^{(1)})-\mathbb{E}[F(\mathbf{z}^{(K+1)})]-\tfrac{1}{4(1-\beta)}\textstyle\sum_{k=1}^K\alpha_k\mathbb{E}\|\nabla F(\mathbf{x}^{(k)})\|^2+\frac{1}{2}\sum_{k=1}^K\big(\tfrac{3\rho\alpha_k^2}{(1-\beta)^2}-\tfrac{\alpha_k}{1-\beta}\big)u_k\\
&\quad+\tfrac{\rho^2}{2(1-\beta)}\textstyle\sum_{k=1}^K\alpha_k\left[\sum_{l=1}^{k-1}\theta_{k,l}\sum_{j=1}^{k-1}\theta_{k,j}u_j+\sum_{j=1}^{k-1}\theta_{k,j}^2u_j\right]\\
&\quad+\tfrac{(1+5\rho)\beta^2}{2(1-\beta)^2}\textstyle\sum_{k=1}^K\alpha_{k-1}^2\sum_{j=1}^{k-1}\big(\frac{\beta^{k-j-1}}{1-\beta}+\beta^{2(k-j-1)}\big)u_j\\
&\quad+\left[\tfrac{\rho^2}{2(1-\beta)}\textstyle\sum_{k=1}^K\alpha_k\sum_{j=1}^{k-1}\theta_{k,j}^2+\frac{(1+5\rho)\beta^2}{2(1-\beta)^2(1-\beta^2)}\sum_{k=1}^K\alpha_{k-1}^2+\frac{\rho}{(1-\beta)^2}\sum_{k=1}^K\alpha_k^2\right]\sigma^2.
\end{aligned}\tag{5.18}$$

Since $\{\alpha_k\}$ is non-increasing, it holds from (5.3) that

$$\theta_{k,j}\le\alpha_{\max\{k-\tau_k,j\}}\pi_{k,j}(\beta),\tag{5.19}$$

which together with the two inequalities in (5.4) gives

$$\textstyle\sum_{j=1}^{k-1}\theta_{k,j}\le\alpha_{\max\{k-\tau_k,1\}}\frac{\tau}{1-\beta},\ \text{and}\ \sum_{j=1}^{k-1}\theta_{k,j}^2\le\alpha_{\max\{k-\tau_k,1\}}^2\frac{\tau}{(1-\beta)^2}.\tag{5.20}$$

Plugging the latter inequality of (5.20) into the fourth line of (5.18), and also interchanging the summations in the second and third lines of (5.18) yield

$$\begin{aligned} 0 \le\ & F(\mathbf{x}^{(1)}) - \inf_{\mathbf{x}} F(\mathbf{x}) - \tfrac{1}{4(1-\beta)} \textstyle\sum_{k=1}^{K} \alpha_k \mathbb{E}\|\nabla F(\mathbf{x}^{(k)})\|^2 + \tfrac{1}{2} \sum_{k=1}^{K} \big(\tfrac{3\rho\alpha_k^2}{(1-\beta)^2} - \tfrac{\alpha_k}{1-\beta}\big) u_k \\ & + \tfrac{\rho^2}{2(1-\beta)} \textstyle\sum_{j=1}^{K-1} u_j \sum_{k=j+1}^{K} \alpha_k \theta_{k,j} \big(\theta_{k,j} + \sum_{l=1}^{k-1} \theta_{k,l}\big) + \tfrac{(1+5\rho)\beta^2}{(1-\beta)^3(1-\beta^2)} \sum_{j=1}^{K-1} u_j \alpha_j^2 \\ & + \Big[\tfrac{\rho^2\tau}{2(1-\beta)} \textstyle\sum_{k=1}^{K} \alpha_k \alpha_{\max\{k-\tau_k,1\}}^2 + \tfrac{(1+5\rho)\beta^2}{2(1-\beta^2)} \sum_{k=1}^{K} \alpha_{k-1}^2 + \rho \sum_{k=1}^{K} \alpha_k^2\Big] \tfrac{\sigma^2}{(1-\beta)^2}, \end{aligned} \tag{5.21}$$

where the last summation in the second line is simplified by utilizing the following summation bound,

$$\textstyle\sum_{k=1}^{K} \alpha_{k-1}^2 \sum_{j=1}^{k-1} t^{k-j-1} u_j = \sum_{j=1}^{K-1} u_j \sum_{k=j+1}^{K} \alpha_{k-1}^2 t^{k-j-1} \le \sum_{j=1}^{K-1} u_j \alpha_j^2/(1-t).$$

Furthermore,

$$\begin{aligned} & \textstyle\sum_{k=j+1}^{K} \alpha_k \theta_{k,j} \big(\theta_{k,j} + \sum_{l=1}^{k-1} \theta_{k,l}\big) \\ \le\ & \alpha_j \textstyle\sum_{k=j+1}^{K} \alpha_{\max\{k-\tau_k,j\}} \pi_{k,j}(\beta) \big(\alpha_{\max\{k-\tau_k,j\}} \pi_{k,j}(\beta) + \tfrac{\alpha_{\max\{k-\tau_k,1\}}\tau}{1-\beta}\big) \\ \le\ & \alpha_j \textstyle\sum_{k:\ j+1\le k\le K,\, k\le j+\tau-1} \alpha_j \tfrac{1}{1-\beta} \big(\alpha_j \tfrac{1}{1-\beta} + \tfrac{\alpha_1\tau}{1-\beta}\big) \\ & + \alpha_j \textstyle\sum_{k:\ j+1\le k\le K,\, k\ge j+\tau} \alpha_j \tfrac{\beta^{k-\tau-j}}{1-\beta} \big(\alpha_j \tfrac{\beta^{k-\tau-j}}{1-\beta} + \tfrac{\alpha_j\tau}{1-\beta}\big) \\ \le\ & \tfrac{\tau(\tau-1)\alpha_1\alpha_j^2}{(1-\beta)^2} + \tfrac{(\tau-1)\alpha_j^3}{(1-\beta)^2} + \tfrac{\tau\alpha_j^3}{(1-\beta)^3} + \tfrac{\alpha_j^3}{(1-\beta)^2(1-\beta^2)}. \end{aligned} \tag{5.22}$$

In the above, the first inequality follows from $\alpha_k \le \alpha_j$ for all $k \ge j$, (5.19) and (5.20); the second inequality breaks the summation on k into two parts: in the first part $k \le j+\tau-1$, we used $\pi_{k,j}(\beta) \le \frac{1}{1-\beta}$ by (5.3) and also $\alpha_{\max\{k-\tau_k,j\}} \le \alpha_j$ and $\alpha_{\max\{k-\tau_k,1\}} \le \alpha_1$; and in the second part $k \ge j+\tau$, since $k \ge j+\tau_k$, we have $\pi_{k,j}(\beta) \le \frac{\beta^{k-\tau-j}}{1-\beta}$ from the second case in the equality of (5.4) and also, $\alpha_{\max\{k-\tau_k,j\}} = \alpha_{\max\{k-\tau_k,1\}} \le \alpha_j$.

Now substitute (5.22) into (5.21), use the assumption in (5.12) to drop the non-positive terms about u_j, also use the definition of $\bar{\mathbf{x}}^{(K)}$ in (5.10), and then rearrange terms to obtain the desired result in (5.13). □

Below we specify the setting of $\{\alpha_k\}$ and show the sublinear convergence.

Corollary 5.5. *Given a maximum number K of iterations, let $\alpha_k = \alpha/\sqrt{K}$ for all $k = 1, \dots, K$, and for some $\alpha > 0$. If $\alpha > 0$ and $\beta > 0$ are chosen such that*

$$\tau^2 + \tfrac{\tau}{1-\beta} + \tfrac{\beta^2}{1-\beta^2} \le \tfrac{(1-\beta)^2 K}{2\alpha^2\rho^2}, \ \text{and}\ 3\rho + \tfrac{2(1+5\rho)\beta^2}{(1-\beta)(1-\beta^2)} \le \tfrac{(1-\beta)\sqrt{K}}{2\alpha}, \tag{5.23}$$

then under Assumptions 2 and 6–8, the iterate $\bar{\mathbf{x}}^{(K)}$ given in (5.10) satisfies

$$\mathbb{E}\|\nabla F(\bar{\mathbf{x}}^{(K)})\|^2 \le \Big(\tfrac{\rho^2\alpha\tau}{2(1-\beta)\sqrt{K}} + \tfrac{(1+5\rho)\beta^2}{2(1-\beta^2)} + \rho\Big) \tfrac{4\alpha\sigma^2}{(1-\beta)\sqrt{K}} + \tfrac{4(1-\beta)\big[F(\mathbf{x}^{(1)}) - \inf_{\mathbf{x}} F(\mathbf{x})\big]}{\alpha\sqrt{K}}. \tag{5.24}$$

Proof. When $\alpha_k \equiv \alpha/\sqrt{K}$, (5.11) is trivially true, and in addition, when (5.23) hold, it is not hard to verify

$$\rho^2 \left[\tfrac{\tau(\tau-1)\alpha_1\alpha_j}{(1-\beta)^2} + \tfrac{(\tau-1)\alpha_j^2}{(1-\beta)^2} + \tfrac{\tau\alpha_j^2}{(1-\beta)^3} + \tfrac{\alpha_j^2}{(1-\beta)^2(1-\beta^2)}\right] \le \tfrac{1}{2}, \ \text{and}\ \left[\tfrac{3\rho}{1-\beta} + \tfrac{2(1+5\rho)\beta^2}{(1-\beta)^2(1-\beta^2)}\right] \alpha_j \le \tfrac{1}{2}, \tag{5.25}$$

which implies (5.12). Finally, (5.13) simplifies to (5.24). □

Remark 5.6. From (5.24), we see that the delay can reduce the convergence speed of Alg. 1 by roughly $O(\frac{\tau}{\sqrt{K}})$. When $\tau = o(\sqrt{K})$, the slow-down effect is negligible.

Corollary 5.7. *Given a maximum number K of iterations, let $\alpha_k = \alpha/\sqrt{a+k-1}$ for all $k = 1,\dots,K$, and for some $a \geq 2\tau$ such that $a\sqrt{a+1} \geq \frac{1-\beta}{2\alpha}$. If*

$$\tau^2 + \tfrac{\tau}{1-\beta} + \tfrac{\beta^2}{1-\beta^2} \leq \tfrac{(1-\beta)^2 a}{2\alpha^2\rho^2}, \text{ and } 3\rho + \tfrac{2(1+5\rho)\beta^2}{(1-\beta)(1-\beta^2)} \leq \tfrac{(1-\beta)\sqrt{a}}{2\alpha}, \tag{5.26}$$

then under Assumptions 2 and 6–8, the output of Alg. 1 satisfies

$$\begin{aligned}\mathbb{E}\|\nabla F(\bar{\mathbf{x}}^{(K)})\|^2 \leq\ & \tfrac{2(1-\beta)\left[F(\mathbf{x}^{(1)}) - \inf_{\mathbf{x}} F(\mathbf{x})\right]}{\alpha(\sqrt{a+K}-\sqrt{a})} \\ & + \left[\tfrac{\rho^2\alpha(1+2a)\tau}{(1-\beta)a\sqrt{a}} + \tfrac{(1+5\rho)\beta^2}{2(1-\beta^2)}\left(2 + \ln\tfrac{a+K-2}{a}\right) + \rho\left(1 + \ln\tfrac{a+K-1}{a}\right)\right] \cdot \tfrac{2\alpha\sigma^2}{(1-\beta)(\sqrt{a+K}-\sqrt{a})}.\end{aligned} \tag{5.27}$$

Remark 5.8. Note that the logarithmic terms in (5.27) dominate the τ-related term if $\tau \leq \frac{\sqrt{a-1}}{\alpha\rho}$, which matches the condition in (5.26). When there is no delay, i.e., $\tau = 0$, a convergence rate of $\widetilde{\Theta}(\frac{1}{\sqrt{K}})$ can be achieved with $a = 1$; when there is a delay, i.e., $\tau > 0$, (5.27) with $a = \Theta(\tau^2)$ gives a rate of $\widetilde{\Theta}(\frac{1}{\sqrt{K+\tau^2}-\sqrt{\tau^2}}) = \widetilde{\Theta}\Big(\frac{1}{\sqrt{K}}(\sqrt{1+\frac{\tau^2}{K}} + \sqrt{\frac{\tau^2}{K}})\Big)$. In this case, the delay will have a negligible effect on the convergence speed if $\tau = o(\sqrt{K})$.

6. Numerical experiments.

In this section, we test Alg. 1 by numerical experiments on three examples: phase retrieval problem, neural network training, and sparse bilinear logistic regression. For each example, we test the effect of the inertial force with different β_k. Also, we demonstrate the advantage of the asynchronous implementation over the synchronous version (i.e., $\tau_k = 0, \forall\, k$) of Alg. 1. In all the tests, we compare the performance of Alg. 1 with different settings of $\{\alpha_k\}$ and $\{\beta_k\}$, which are fixed to constants for all iterations k or decrease with respect to the number of epochs.

6.1. Phase retrieval problem.

The phase retrieval problem aims to recover a signal $\mathbf{x}^* \in \mathbb{R}^d$ from m measuring vectors[1] $\{\mathbf{a}_i\}_{i=1}^m$ and the correspondingly obtained magnitudes $\{b_i\}_{i=1}^m$. It can be formulated into the following non-smooth minimization problem [12, 16, 17]:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m \left|\,|\langle \mathbf{a}_i, \mathbf{x}\rangle|^2 - b_i^2\right|, \tag{6.1}$$

which is in the form of (1.1) with $F(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m \left|\,|\langle \mathbf{a}_i, \mathbf{x}\rangle|^2 - b_i^2\right|$ and $r(\mathbf{x}) \equiv 0$. In the test, the vector $\mathbf{a}_i$ followed the standard multivariate Gaussian distribution, i.e., $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and we let $b_i = |\langle \mathbf{a}_i, \mathbf{x}^*\rangle|, \forall\, i$, for a ground truth $\mathbf{x}^*$. Hence, the optimal objective value is *zero*.

Synthetic data. We first solved (6.1) with $\mathbf{x}^*$ generated from a uniform distribution on the d-dimensional unit sphere. Fig. 2 shows the results for $m = 50,000$ and $d = 20,000$. We tested the algorithm for several pairs of (m, d) and observed similar results. In the test, we computed a stochastic subgradient by using 100 data points, i.e., the minibatch size was set to 100. The parameters either followed a constant scheme with $\alpha_k = \alpha, \beta_k = \beta, \forall\, k$ where $\alpha = 5 \times 10^{-5}$ and

[1] In general, the signal $\mathbf{x}$ and the measuring vectors $\{\mathbf{a}_i\}$ can be complex-valued. For simplicity, we focus on the real field.

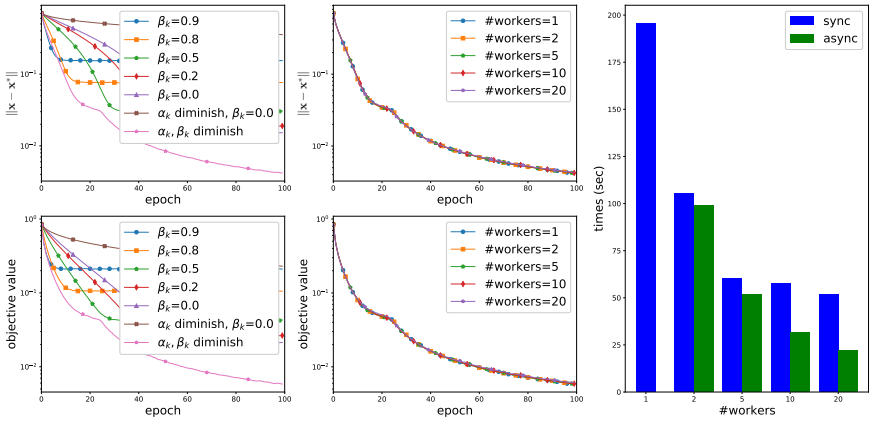

Figure 2: Results by Alg. 1 on solving instances of the Phase Retrieval Problem (6.1) with randomly generated $\mathbf{x}^*$, $m = 50,000$ and $d = 20,000$. Left: non-parallel implementation of Alg. 1 with different choices of $\{\alpha_k\}$ and $\{\beta_k\}$; Middle: async-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and $\{\beta_k\}$, and with different numbers of workers; Right: running time (in second) of the sync-parallel and async-parallel implementation of Alg. 1 with different numbers of workers.

$\beta \in \{0, 0.2, 0.5, 0.8, 0.9\}$; or diminished with $\alpha_k = \frac{5\times10^{-5}}{\sqrt{e_k+1}}$ and $\beta_k = \min\{0.9, \frac{2}{(e_k+1)^{1/4}}\}, \forall\, k$, or $\beta_k = 0, \forall\, k$. Here, e_k denotes the epoch number at the k-th iteration. During the test, we also experimented with different values of the constant α. We found that for a smaller α, the algorithm converged more slowly but could reach a lower objective value. The choice $\alpha = 5 \times 10^{-5}$ resulted in a good trade-off between the convergence speed and the final objective value.

From the left subfigure in Fig. 2, we see that the algorithm with a bigger β converged faster but achieved a higher objective value. The convergence of the algorithm with a diminishing $\{\alpha_k\}$ and constant $\beta_k = 0$ is the slowest. The best results were obtained by the choice of diminishing $\{\alpha_k\}$ and $\{\beta_k\}$. Comparing the curve with diminishing $\{\alpha_k\}$ and $\{\beta_k\}$ to that with $\beta_k = 0.9, \forall\, k$, we notice that the two curves are almost the same within the first 5 epochs, i.e., before the latter one becomes flat. However, the former can decrease the objective to a significantly smaller value. Thus both the choices of $\{\alpha_k\}$ and $\{\beta_k\}$ contribute to the best results. With the diminishing $\{\alpha_k\}$ and $\{\beta_k\}$ that yield the best results for the non-parallel case, we then compared the sync-parallel and async-parallel implementations of Alg. 1. The middle subfigure in Fig. 2 shows the results for the async-parallel version with different numbers of workers. The right subfigure shows the running time of both versions. The results show that the convergence speed (in terms of epoch number) of the async-parallel method is almost never affected by the asynchrony (or information delay). In addition, we see that the async-parallel implementation yielded significantly higher parallelization speed-up over the sync-parallel one, according to the right subfigure in Fig. 2.

Image data. We also solved (6.1) with $\mathbf{x}^*$ flattened from an image. We tested with two images: a CT scan image[2] of size 94×138 after downsampling and the cameraman image[3] of size 196×196 after cropping. Fig. 3 shows the ground-truth images, Fig. 4 and Fig. 6 show convergence curves and computing times, and Fig. 5 and Fig. 7 show recovered images. In the test, for the CT scan

[2]https://aimi.stanford.edu/radiopaedia-list-ai-imaging-datasets
[3]https://github.com/antimatter15/cameraman

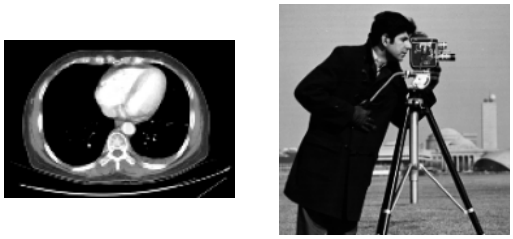

Figure 3: Ground-truth images. Left: a CT scan image. Right: the cameraman image.

image, $d = 12,972$, and we selected $m = 40,000$, computed each stochastic subgradient by using 100 randomly sampled data points, and set $\alpha_k = \frac{10^{-4}}{\sqrt{e_k+1}}$; for the cameraman image, $d = 38,416$, and we selected $m = 60,000$, computed each stochastic subgradient by using 60 randomly sampled data points and set $\alpha_k = \frac{5\times10^{-5}}{\sqrt{e_k+1}}$. We first tested the non-parallel version of Alg. 1 with $\beta_k = \beta, \forall k$, where $\beta \in \{0, 0.2, 0.5, 0.8, 0.9\}$, and then tested the parallel version by different numbers of workers.

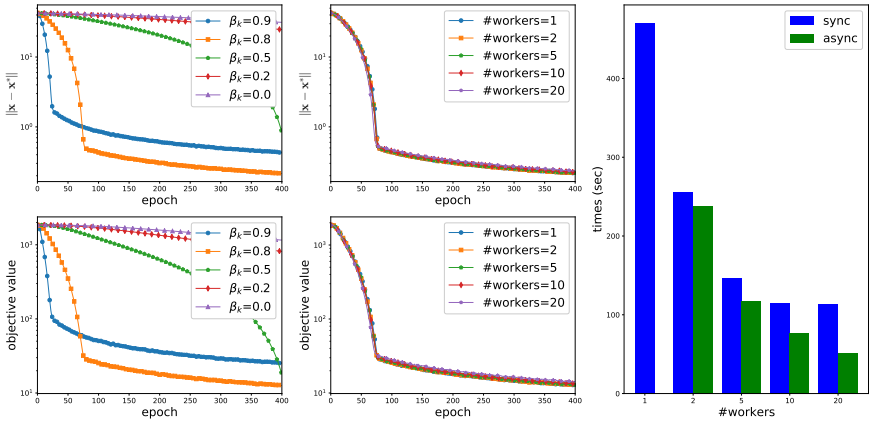

Figure 4: Results by Alg. 1 on solving instances of the Phase Retrieval Problem (6.1) with a CT scan image as $\mathbf{x}^*$ and $m = 40,000$. Left: non-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and different choices of $\{\beta_k\}$; Middle: async-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and $\beta_k = 0.8$, and with different numbers of workers; Right: running time (in second) of the sync-parallel and async-parallel implementation of Alg. 1 with different numbers of workers.

From the left subfigures in Fig. 4 and Fig. 6, we see that the algorithm with a bigger β converged faster. After 400 epochs, the algorithm achieved the lowest objective value and the smallest distance from $\mathbf{x}^*$ with $\beta_k = 0.8, \forall\, k$ for the CT scan image, and with $\beta_k = 0.9, \forall\, k$ for the cameraman image. Alg. 1 recovered the image clearly for the CT scan image with $\beta_k \equiv \beta \in \{0.9, 0.8, 0.5\}$ shown in the top subfigures in Fig. 5 and for the cameraman image with $\beta_k = 0.9, \forall\, k$ in the top subfigures in Fig. 7. The recovered images became clearer as the β value increases. The middle subfigures in Fig. 4 and Fig. 6 show the results for the async-parallel version of Alg. 1 with different numbers of workers, and the bottom subfigures in Fig. 5 and Fig. 7 show the corresponding recovered images. The right subfigures in Fig. 4 and Fig. 6 show the running time of both sync-parallel and async-

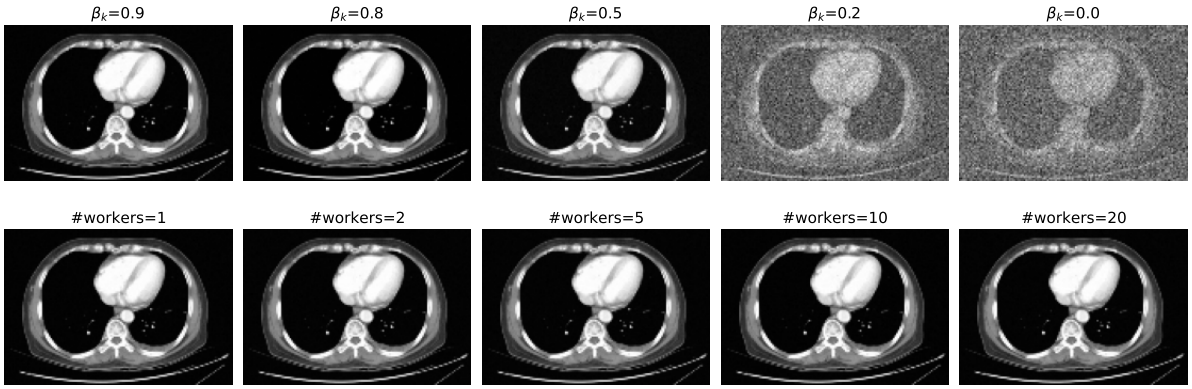

Figure 5: Recovered images by Alg. 1 on solving instances of the Phase Retrieval Problem (6.1) with a CT scan image as $\mathbf{x}^*$ and $m = 40,000$. Top: non-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and different choices of $\{\beta_k\}$; Bottom: async-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and $\beta_k = 0.8$, and with different numbers of workers.

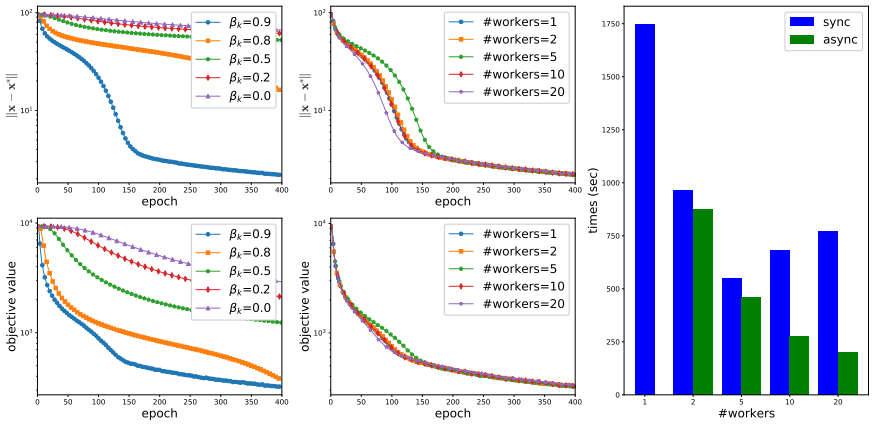

Figure 6: Results by Alg. 1 on solving instances of the Phase Retrieval Problem (6.1) with the crameman image as $\mathbf{x}^*$ and $m = 60,000$. Left: non-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and different choices of $\{\beta_k\}$; Middle: async-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and $\beta_k = 0.9$, and with different numbers of workers; Right: running time (in second) of the sync-parallel and async-parallel implementation of Alg. 1 with different numbers of workers.

parallel versions of Alg. 1. The results show that the convergence speed (in terms of epoch number) of the async-parallel method is rarely affected by the asynchrony (or information delay). In addition, we see that the async-parallel implementation yielded significantly higher parallelization speed-up over the sync-parallel one.

6.2. Neural network models training. In this subsection, we trained two neural network models by Alg. 1. One is LeNet5 on the MNIST dataset [26] and the other AllCNN [55] on the Cifar10 dataset [25]. LeNet5 has 2 convolutional, 2 max-pooling, and 3 fully-connected layers. AllCNN has 9

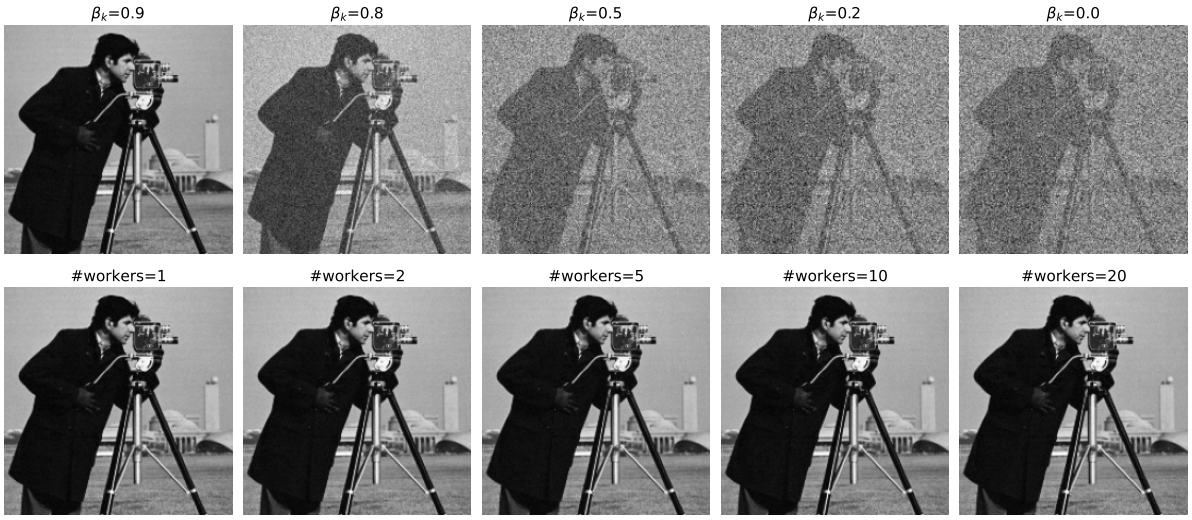

Figure 7: Recovered images by Alg. 1 on solving instances of the Phase Retrieval Problem (6.1) with the crameman image as $\mathbf{x}^*$ and $m = 60,000$. Top: non-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and different choices of $\{\beta_k\}$; Bottom: async-parallel implementation of Alg. 1 with diminishing $\{\alpha_k\}$ and $\beta_k = 0.9$, and with different numbers of workers.

convolutional and 1 avg-pooling layers. The outputs of the two models are re-scaled as probabilities in all classes for each data sample by the softmax function. The estimated probabilities and the true class labels are fed to the negative log likelihood loss function to get the losses. The objective is to minimize the mean loss over all data samples, which is in the form of (1.1) with the model weights as $\mathbf{x}$, the mean loss as $F(\mathbf{x})$ and $r(\mathbf{x}) \equiv 0$. For both trainings, we set $\alpha_k = \alpha, \forall\, k$ and selected the best α from $\{0.01, 0.005, 0.001, 0.0005, 0.0001\}$. For training LeNet5, we used $\alpha_k = 0.001$, and for training Cifar10, we used $\alpha_k = 0.005, \forall\, k$.

The results of training LeNet5 on the MNIST dataset are shown in Fig. 8. In the test, we computed a stochastic subgradient by using 40 data samples, i.e., the minibatch size was set to 40. We first tested Alg. 1 with $\beta_k = \beta, \forall k$, where $\beta \in \{0, 0.2, 0.5, 0.8, 0.9\}$, or $\beta_k = \min\left\{0.9, \frac{2}{(e_k+1)^{1/4}}\right\}, \forall\, k$. The first column of Fig. 8 shows that the algorithm with a bigger β gave better results. Notice that the algorithm with $\beta_k = 0.9$ or $\beta_k = \min\left\{0.9, \frac{2}{(e_k+1)^{1/4}}\right\}, \forall\, k$ give the highest testing accuracy. For these two choices, we ran the async-parallel version of Alg. 1 with different numbers of workers. From the results in the second and third columns of Fig. 8, we see that the asynchrony had negative effect on the behavior of the algorithm, especially when more workers were used. Nevertheless, the final training loss for all different number of workers is almost the same, and the final testing accuracy by using 10 or 20 workers is slightly lower than that produced by using fewer workers. The fourth column compares the running time of the sync-parallel and async-parallel implementations of Alg. 1 with $\beta_k = \min\left\{0.9, \frac{2}{(e_k+1)^{1/4}}\right\}, \forall\, k$. Again, the bars show significantly higher parallelization speed-up by the async-parallel implementation over the sync-parallel one.

The results of training AllCNN on the Cifar10 dataset are shown in Fig. 9. In the test, we set the minibatch size to 100 and $\beta_k = \beta, \forall k$, where $\beta \in \{0, 0.2, 0.5, 0.8, 0.9\}$. The left column of Fig. 9 shows that the algorithm with a bigger β gave better results. The choice of $\beta_k = 0.9, \forall\, k$ yielded the

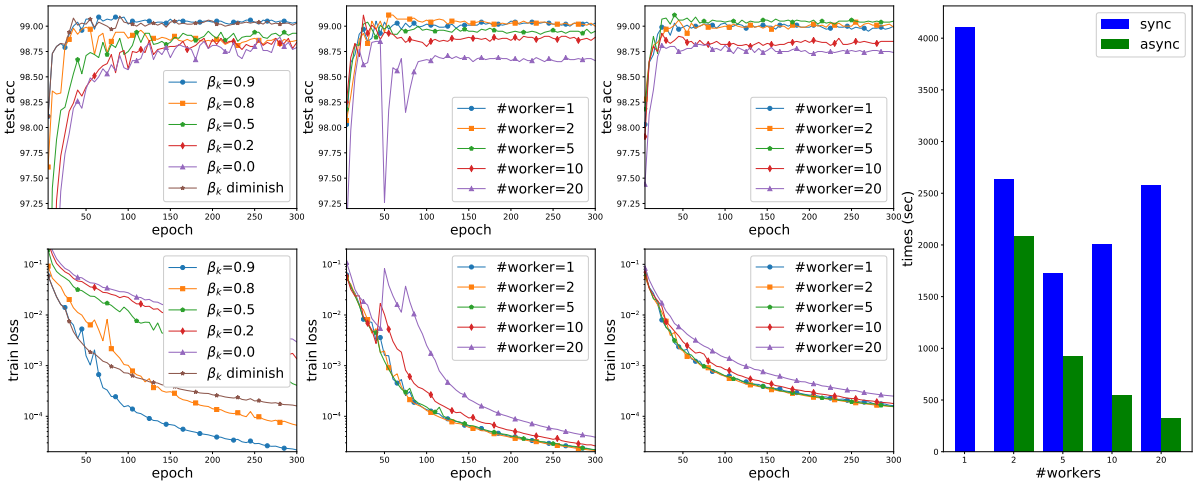

Figure 8: Results by Alg. 1 on training LeNet5 on the MNIST dataset. First column: non-parallel implementation of Alg. 1 with $\alpha_k = 0.001, \forall\, k$ and different choices of $\{\beta_k\}$; Second column: async-parallel implementation of Alg. 1 with $\alpha_k = 0.001$ and $\beta_k = 0.9, \forall\, k$; Third column: async-parallel implementation of Alg. 1 with $\alpha_k = 0.001$ and $\beta_k = \min\left\{0.9, \frac{2}{(e_k+1)^{1/4}}\right\}, \forall\, k$; Fourth column: running time (in second) of the sync-parallel and async-parallel implementations of Alg. 1 with different numbers of workers.

best results. With this choice, we compared the sync-parallel and async-parallel implementations of Alg. 1. The middle column in Fig. 9 shows the results for the async-parallel version with different numbers of workers. The right column shows the running time of both versions. From the results, we see that the convergence speed (in terms of epoch number) of the async-parallel method is almost not affected by the asynchrony. In addition, we see again that the async-parallel implementation yielded higher parallelization speed-up over the sync-parallel one.

6.3. Sparse bilinear logistic regression. In this subsection, we test Alg. 1 on solving the sparse bilinear logistic regression (BLR) built in [54]. Let $\{(X_i, y_i)\}_{i=1}^m$ be the training data set with each data sample $X_i \in \mathbb{R}^{s\times t}$ and label $y_i \in \{1, 2, ..., C\}$ for $i = 1, 2, ..., m$, where C is the number of classes. The sparse BLR is modeled as

$$\min_{\mathcal{U},\mathcal{V},\mathbf{b}} -\frac{1}{m}\sum_{i=1}^{m} \log\left(\frac{\exp[\mathrm{tr}(U_{y_i} X_i V_{y_i}) + b_{y_i}]}{\sum_{j=1}^{C} \exp[\mathrm{tr}(U_j X_i V_j) + b_j]}\right) + \lambda(\|\mathcal{U}\|_1 + \|\mathcal{V}\|_1 + \|\mathbf{b}\|_1), \tag{6.2}$$

where $\mathcal{U} = (U_1, U_2, ..., U_C), \mathcal{V} = (V_1, V_2, ..., V_C), \mathbf{b} = (b_1, b_2, ..., b_C)$ with $U_j \in \mathbb{R}^{p\times s}, V_j \in \mathbb{R}^{t\times p}, b_j \in \mathbb{R}$ for $j = 1, 2, ..., C$, $\|\mathcal{U}\|_1 := \sum_{j=1}^{C}\sum_{i=1}^{p}\sum_{l=1}^{s} |(U_j)_{i,l}|$, $\lambda \geq 0$ is the weight for the sparse regularizer, and $\mathrm{tr}(S) := \sum_{i=1}^{p} S_{i,i}$ for any matrix $S \in \mathbb{R}^{p\times p}$. To solve (6.2), we apply Alg. 1 with $\mathbf{x} = (\mathcal{U}, \mathcal{V}, \mathbf{b})$, $F(\mathbf{x})$ being the first term in (6.2), and $r(\mathbf{x}) = \lambda(\|\mathcal{U}\|_1 + \|\mathcal{V}\|_1 + \|\mathbf{b}\|_1)$.

In this test, we used the MNIST dataset [26] and set the minibatch to 100 while computing a stochastic gradient of F. To obtain a relatively high accuracy and also relatively cheap computation, we chose $p = 5$ and $\lambda = 10^{-3}$. The learning rate was set to $\alpha_k = \alpha, \forall k$ with α tuned from $\{0.01, 0.005, 0.001, 0.0005, 0.0001\}$. To ensure convergence and also satisfactory final testing

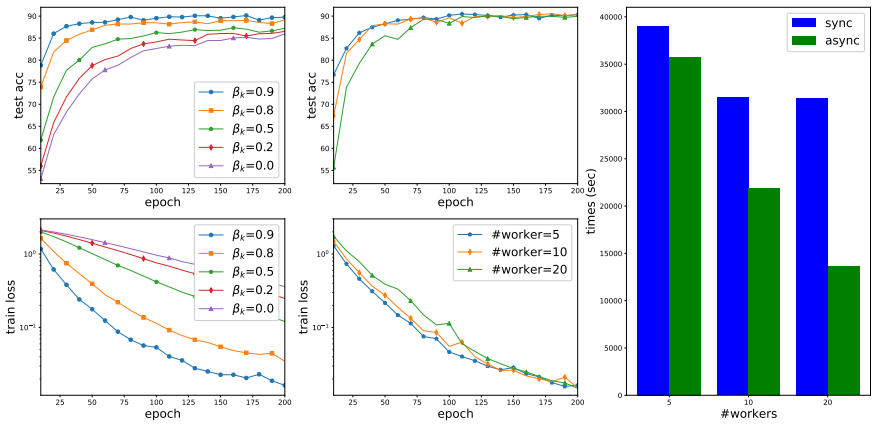

Figure 9: Results by Alg. 1 on training AllCNN on the Cifar10 dataset. Left:non-parallel implementation of Alg. 1 with $\alpha_k = 0.005$ and different $\{\beta_k\}$; Middle: async-parallel implementation of Alg. 1 with $\alpha_k = 0.005, \beta_k = 0.9$, and with different numbers of workers; Right: running time (in second) of the sync-parallel and async-parallel implementations of Alg. 1 with different numbers of workers.

accuracy for both async-parallel and sync-parallel implementations of Alg. 1, we set $\alpha = 0.0005$. Note that the sync-parallel version could converge faster in the beginning with a larger α but the final testing accuracy and training loss were similar to those produced by using $\alpha = 0.0005$.

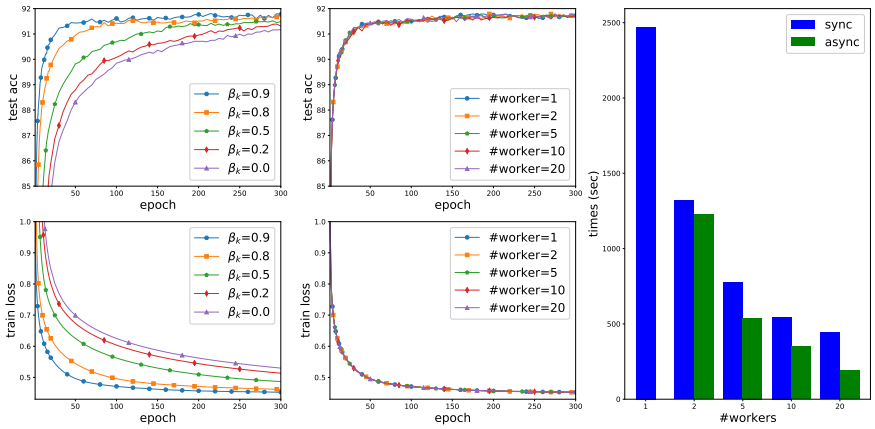

Figure 10: Results by Alg. 1 to solve the sparse bilinear logistic regression (6.2) on the MNIST dataset with $p = 5$ and $\lambda = 0.001$. Left: non-parallel implementation of Alg. 1 with $\alpha_k = 0.0005$ and different $\{\beta_k\}$; Middle: async-parallel implementation of Alg. 1 with $\alpha_k = 0.0005, \beta_k = 0.9, \forall\, k$, and with different numbers of workers; Right: running time (in second) of the sync-parallel and async-parallel implementations of Alg. 1 with different numbers of workers.

The left column of Fig. 10 shows the results by Alg. 1 with $\beta_k = \beta \in \{0, 0.2, 0.5, 0.8, 0.9\}, \forall k$. We see that the algorithm with a bigger β converges faster. The middle column in Fig. 10 shows the results by the async-parallel implementation of Alg. 1 with $\beta_k = 0.9, \forall k$ and with different numbers

of workers. The right column shows the running time of both sync-parallel and async-parallel implementations. The results show that the convergence speed (in terms of epoch number) of the async-parallel method is almost not affected by the asynchrony. In addition, we see that the async-parallel implementation yielded significantly higher parallelization speed-up over the sync-parallel one.

7. Conclusions. We have proposed an inertial-accelerated proximal stochastic subgradient method for solving non-convex stochastic optimization. An $O(1/K^{\frac{1}{2}})$ convergence rate result is established for three different problem classes, by the measure of the expected value of the gradient norm square of the objective function or its Moreau envelope, where K is the number of total iterations. The same-order convergence rate can be shown even if the derivative information is outdated in an asynchronous distributed computing environment, provided that the delay (or staleness) of the derivative is in a tolerable range. Numerical experiments on phase retrieval, neural network training, and sparse bilinear logistic regression demonstrate faster convergence by using the inertial-acceleration technique and also the higher parallelization speed-up of the asynchronous computing over the synchronous counterpart.

Acknowledgements. The authors would like to thank two anonymous referees for their valuable comments and also the associate editor for the suggestion to add tests with images, which help greatly improve the paper. The work of Y. Xu and Y. Yan is partly supported by NSF grant DMS-2053493 and the Rensselaer-IBM AI Research Collaboration, part of the IBM AI Horizons Network. J. Chen is supported in part by DOE Award DE-OE0000910.

Appendix A. Remaining proofs. In this section, we provide proofs of the lemmas that are used in our analysis.

Proof of Lemma 3.3. For ease of notation, we denote $\delta = 1 - \alpha_k\overline{\rho}$ in this proof. We have

$$
\begin{aligned}
&\|\mathbf{x}^{(k+1)} - \widetilde{\mathbf{x}}^{(k)}\|^2 \\
&= \left\|\mathbf{prox}_{\alpha_k r}\big(\mathbf{x}^{(k)} - \alpha_k\mathbf{g}^{(k)} + \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\big) - \mathbf{prox}_{\alpha_k r}\big(\alpha_k\overline{\rho}\mathbf{x}^{(k)} - \alpha_k\widetilde{\mathbf{v}}^{(k)} + (1-\alpha_k\overline{\rho})\widetilde{\mathbf{x}}^{(k)}\big)\right\|^2 \\
&\le \left\|\delta(\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}) - \alpha_k(\mathbf{g}^{(k)} - \widetilde{\mathbf{v}}^{(k)}) + \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\right\|^2 \qquad \text{(A.1)} \\
&= \left\|\delta(\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}) + \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\right\|^2 + \alpha_k^2\left\|\mathbf{g}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\right\|^2 \qquad \text{(A.2)} \\
&\quad - 2\alpha_k\big\langle\delta(\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}) + \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \mathbf{g}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\big\rangle,
\end{aligned}
$$

where the first equality is from (1.2) and (3.3), and the inequality follows from the nonexpansiveness of the proximal mapping. Taking conditional expectation on ξ_k over the equation in (A) gives

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}\|\mathbf{x}^{(k+1)} - \widetilde{\mathbf{x}}^{(k)}\|^2 \\
&\le \left\|\delta(\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}) + \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\right\|^2 + \alpha_k^2\mathbb{E}_{\xi_k}\left\|\mathbf{g}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\right\|^2 \\
&\quad - 2\alpha_k\big\langle\delta(\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}) + \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \mathbf{v}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\big\rangle \\
&\le \big(\delta\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\| + \beta_k\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|\big)^2 + 4\alpha_k^2M^2 \\
&\quad - 2\alpha_k\delta\big\langle\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}, \mathbf{v}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\big\rangle - 2\alpha_k\beta_k\big\langle\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}, \mathbf{v}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\big\rangle \\
&\le \delta^2(1+c_k)\|\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}\|^2 + (1+\tfrac{1}{c_k})\beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + 4\alpha_k^2M^2 \\
&\quad - 2\alpha_k\delta\big\langle\mathbf{x}^{(k)} - \widetilde{\mathbf{x}}^{(k)}, \mathbf{v}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\big\rangle + \beta_k^2\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 + \alpha_k^2\|\mathbf{v}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\|^2,
\end{aligned}
$$

where the second inequality holds by (3.2), and the third inequality follows from the Young's inequality along with a scalar $c_k > 0$. Now we obtain the desired result by plugging (3.4) into the above inequality,

bounding $\|\mathbf{v}^{(k)} - \widetilde{\mathbf{v}}^{(k)}\|^2 \le 4M^2$, and noticing

$$\delta^2(1+c_k) + 2\alpha_k\delta\rho = 1 - 2\alpha_k(\overline{\rho}-\rho) - \alpha_k^2\overline{\rho}(2\rho-\overline{\rho}) + c_k\delta^2 \le 1 - 2\alpha_k(\overline{\rho}-\rho) + c_k,$$

where the equality holds because $\delta = 1-\alpha_k\overline{\rho}$, and the inequality follows from $\delta<1$, $c_k>0$, and $\overline{\rho}\le 2\rho$. □

Proof of Lemma 3.7. Taking conditional expectation on τ_k, we have $\mathbb{E}_{\tau_k}[F(\mathbf{x}^{(k-\tau_k)})] = \sum_{j=0}^{\tau} p_j F(\mathbf{x}^{(k-j)})$, where we let $\mathbf{x}^{(k)} = \mathbf{x}^{(1)}, \forall\, k \le 0$. Hence,

$$\begin{aligned}\textstyle\sum_{k=1}^K \mathbb{E}\big[F(\mathbf{x}^{(k-\tau_k)})\big] &= \textstyle\sum_{k=1}^K \sum_{j=0}^{\tau} p_j \mathbb{E}\big[F(\mathbf{x}^{(k-j)})\big] \\ &= \textstyle\sum_{k=1}^K \sum_{t=k-\tau}^{k} p_{k-t}\mathbb{E}\big[F(\mathbf{x}^{(t)})\big] = \sum_{t=1-\tau}^{K}\sum_{k=\max\{1,t\}}^{\min\{K,t+\tau\}} p_{k-t}\mathbb{E}\big[F(\mathbf{x}^{(t)})\big],\end{aligned}$$

and

$$\begin{aligned}\sum_{k=1}^K \mathbb{E}\big[F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-\tau_k)})\big] &= \sum_{k=1}^K \mathbb{E}\big[F(\mathbf{x}^{(k)})\big] - \sum_{k=1-\tau}^{K}\sum_{t=\max\{1,k\}}^{\min\{K,k+\tau\}} p_{t-k}\mathbb{E}\big[F(\mathbf{x}^{(k)})\big] \\ &= \left(1 - \sum_{k=1-\tau}^{1}\sum_{t=1}^{k+\tau} p_{t-k}\right) F(\mathbf{x}^{(1)}) + \sum_{k=K-\tau+1}^{K}\left(1-\sum_{t=k}^{K} p_{t-k}\right)\mathbb{E}\big[F(\mathbf{x}^{(k)})\big] \\ &\le \tau\max\big\{0, -F(\mathbf{x}^{(1)})\big\} + \tau C_F. \qquad \text{(A.3)}\end{aligned}$$

In addition, because $\tau_k \le \tau, \forall\, k$, it holds $\sum_{k=1}^K \|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2 \le \tau^2 \sum_{k=1}^K \|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\|^2$, which together with (3.19) gives

$$\textstyle\sum_{k=1}^K \|\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)}\|^2 \le \tau^2\left(\frac{\alpha}{\gamma\sqrt{K}}\big(\phi(\mathbf{x}^{(1)}) - \phi^*\big) + \frac{\alpha^2M^2}{\gamma^2}\right). \qquad \text{(A.4)}$$

For the last term in $\mathcal{E}_k$, we use (3.2) and Assumption 4 to bound it as follows

$$-\big\langle \mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}, \mathbf{v}^{(k)}\big\rangle \le M\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}\| \le M\textstyle\sum_{j=1}^{\tau}\|\mathbf{x}^{(k+1-j)} - \mathbf{x}^{(k-j)}\|,$$

and thus

$$\textstyle\sum_{k=1}^K -\big\langle \mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}, \mathbf{v}^{(k)}\big\rangle \le M\tau\sum_{k=1}^K \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|. \qquad \text{(A.5)}$$

By the Cauchy-Schwarz inequality and Jensen's inequality, we have

$$\textstyle\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \le \sqrt{K}\sqrt{\sum_{k=1}^K \big(\mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|\big)^2} \le \sqrt{K}\sqrt{\sum_{k=1}^K \mathbb{E}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2},$$

which together with (3.19) and (A.5) gives

$$\textstyle\sum_{k=1}^K \mathbb{E}\left[-\big\langle \mathbf{x}^{(k)} - \mathbf{x}^{(k-\tau_k)}, \mathbf{v}^{(k)}\big\rangle\right] \le M\tau\sqrt{K}\sqrt{\frac{\alpha}{\gamma\sqrt{K}}\big(\phi(\mathbf{x}^{(1)}) - \phi^*\big) + \frac{\alpha^2M^2}{\gamma^2}}. \qquad \text{(A.6)}$$

Now we obtain the desired result from (A.3), (A.4), and (A.6). □

Proof of Lemma 3.11. When condition 2 of Assumption 5 holds, the update in (1.2) indicates that there exists a subgradient $\tilde{\nabla} r_2(\mathbf{x}^{(k+1)})$ such that

$$\big\langle \mathbf{y} - \mathbf{x}^{(k+1)}, \alpha_k\tilde{\nabla} r_2(\mathbf{x}^{(k+1)}) + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} + \alpha_k\mathbf{g}^{(k)} - \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\big\rangle \ge 0, \text{ for all } \mathbf{y}\in X.$$

Letting $\mathbf{y} = \mathbf{x}^{(k)}$ and rearranging terms in the above inequality, we have

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \le \big\langle \mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}, \alpha_k(\tilde{\nabla} r_2(\mathbf{x}^{(k+1)}) + \mathbf{g}^{(k)}) - \beta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})\big\rangle, \qquad \text{(A.7)}$$

which together with the Cauchy-Schwarz inequality gives

$$\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\| \le \|\alpha_k(\tilde{\nabla} r_2(\mathbf{x}^{(k+1)})+\mathbf{g}^{(k)})-\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})\|.$$

Hence, by the triangle inequality and the Young's inequality, we have for any $c>0$,

$$\begin{aligned}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2 &\le (\alpha_k\|\tilde{\nabla} r_2(\mathbf{x}^{(k+1)})+\mathbf{g}^{(k)}\|+\beta_k\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|)^2\\ &\le \alpha_k^2(1+\tfrac{1}{c})\|\tilde{\nabla} r_2(\mathbf{x}^{(k+1)})+\mathbf{g}^{(k)}\|^2+\beta_k^2(1+c)\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2\\ &\le 2\alpha_k^2(1+\tfrac{1}{c})(\|\tilde{\nabla} r_2(\mathbf{x}^{(k+1)})\|^2+\|\mathbf{g}^{(k)}\|^2)+\beta_k^2(1+c)\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2.\end{aligned}$$

Take full expectation on both sides of the above inequality and use Assumption 3 and condition 2 of Assumption 5 to obtain

$$\mathbb{E}\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2 \le 2\alpha_k^2(1+\tfrac{1}{c})\left(M_r^2+M^2\right)+\beta_k^2(1+c)\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2.$$

Let $c=\frac{1}{2}(1/\tilde{\beta}^2-1)$ and sum up the above inequality over $k=1$ to K. We obtain (3.24) by rearranging terms and using $\mathbf{x}^{(0)}=\mathbf{x}^{(1)}$. □

Proof of Lemma 3.12. By similar arguments as in the proof to obtain (A.3), we have

$$\begin{aligned}&\sum_{k=k_0}^{K}\alpha_k\mathbb{E}\big[F(\mathbf{x}^{(k)})-F(\mathbf{x}^{(k-\tau_k)})\big]\\ =\ &\sum_{k=k_0}^{K}\alpha_k\mathbb{E}\big[F(\mathbf{x}^{(k)})-C_F\big]-\sum_{k=k_0-\tau}^{K}\sum_{t=\max\{k_0,k\}}^{\min\{K,k+\tau\}}\alpha_t p_{t-k}\mathbb{E}\big[F(\mathbf{x}^{(k)})-C_F\big]\\ =\ &-\sum_{k=k_0-\tau}^{k_0-1}\sum_{t=\max\{k_0,k\}}^{\min\{K,k+\tau\}}\alpha_t p_{t-k}\mathbb{E}\big[F(\mathbf{x}^{(k)})-C_F\big]+\sum_{k=k_0}^{K}\left(\alpha_k-\sum_{t=\max\{k_0,k\}}^{\min\{K,k+\tau\}}\alpha_t p_{t-k}\right)\mathbb{E}\big[F(\mathbf{x}^{(k)})-C_F\big]\\ \text{(A.8)}\ \le\ &2\alpha_{k_0}\tau C_F,\end{aligned}$$

where the inequality holds by the nonincreasing monotonicity of $\{\alpha_k\}$ and the fact $|F(\mathbf{x}^{(k)})|\le C_F, \forall\, k$.

In addition, from the nonincreasing monotonicity of $\{\alpha_k\}$ and $\tau_k\le\tau, \forall\, k$, it holds

$$\textstyle\sum_{k=k_0}^{K}\alpha_k\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2 \le \alpha_{k_0}\tau^2\sum_{k=k_0-\tau+1}^{K}\|\mathbf{x}^{(k-1)}-\mathbf{x}^{(k)}\|^2 \le \alpha_{k_0}\tau^2\sum_{k=2}^{K}\|\mathbf{x}^{(k-1)}-\mathbf{x}^{(k)}\|^2.$$

Hence, by (3.22) and (3.24), and the definitions of C_1 and C_2 in (3.27), we have from the above that

$$\text{(A.9)}\qquad \textstyle\sum_{k=k_0}^{K}\alpha_k\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2 \le \alpha_{k_0}\tau^2\big(C_1+C_2\sum_{k=1}^{K}\alpha_k^2\big)$$

Finally, similar to (A.5), we have

$$\text{(A.10)}\qquad \textstyle\sum_{k=k_0}^{K}\left[-\alpha_k\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k-\tau_k)},\mathbf{v}^{(k)}\rangle\right] \le M\tau\sum_{k=k_0}^{K}\alpha_k\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|.$$

By the Cauchy-Schwarz inequality and Jensen's inequality, it holds

$$\sum_{k=k_0}^{K}\alpha_k\mathbb{E}\big[\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|\big] \le \sqrt{\sum_{k=k_0}^{K}\alpha_k^2}\sqrt{\sum_{k=k_0}^{K}\mathbb{E}\big[\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2\big]} \le \sqrt{\sum_{k=k_0}^{K}\alpha_k^2}\sqrt{C_1+C_2\sum_{k=1}^{K}\alpha_k^2},$$

which together with (A.10) gives

$$\text{(A.11)}\qquad \textstyle\sum_{k=k_0}^{K}\mathbb{E}\left[-\alpha_k\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k-\tau_k)},\mathbf{v}^{(k)}\rangle\right] \le M\tau\sqrt{\sum_{k=k_0}^{K}\alpha_k^2}\sqrt{C_1+C_2\sum_{k=1}^{K}\alpha_k^2}.$$

Now (3.25) follows from (A.8), (A.9), and (A.11), and also $1-\alpha_k\overline{\rho}\le 1$ and $\sum_{k=1}^K \alpha_k^2 \le \alpha^2(1+\ln K)$. □

Proof of Lemma 4.2. As in the proof of Lemma 3.3, we denote $\delta = 1-\alpha_k\overline{\rho}$ and take conditional expectation about ξ_k over both sides of (A.1) to have

$$\begin{aligned}
\mathbb{E}_{\xi_k}\|\mathbf{x}^{(k+1)}-\widetilde{\mathbf{x}}^{(k)}\|^2 &\le \mathbb{E}_{\xi_k}\big\|\delta(\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)})-\alpha_k(\mathbf{g}^{(k)}-\widetilde{\mathbf{v}}^{(k)})+\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})\big\|^2\\
&= \mathbb{E}_{\xi_k}\big\|\delta(\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)})-\alpha_k(\nabla F(\mathbf{x}^{(k)})-\nabla F(\widetilde{\mathbf{x}}^{(k)}))+\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})-\alpha_k\mathbf{w}^{(k)}\big\|^2\\
&= \big\|\delta(\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)})-\alpha_k(\nabla F(\mathbf{x}^{(k)})-\nabla F(\widetilde{\mathbf{x}}^{(k)}))+\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})\big\|^2+\alpha_k^2\mathbb{E}_{\xi_k}\|\mathbf{w}^{(k)}\|^2\\
&\quad -2\alpha_k\mathbb{E}_{\xi_k}\big\langle\delta(\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)})-\alpha_k(\nabla F(\mathbf{x}^{(k)})-\nabla F(\widetilde{\mathbf{x}}^{(k)}))+\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}),\mathbf{w}^{(k)}\big\rangle\\
&\le \big((\delta+\alpha_k\rho)\|\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)}\|+\beta_k\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|\big)^2+\alpha_k^2\big(\sigma^2+\rho^2\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2\big)+\mathcal{E},
\end{aligned}\tag{A.12}$$

where we have used Assumption 6 and Lemma 4.1 to obtain the last inequality, and we denote $\mathbf{w}^{(k)} = \mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k)})$ and

$$\mathcal{E} := -2\alpha_k\big\langle\delta(\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)})-\alpha_k(\nabla F(\mathbf{x}^{(k)})-\nabla F(\widetilde{\mathbf{x}}^{(k)}))+\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}),\nabla F(\mathbf{x}^{(k-\tau_k)})-\nabla F(\mathbf{x}^{(k)})\big\rangle.$$

Now we apply the Young's inequality to bound the first square term in (A.12) to obtain

$$\begin{aligned}
\mathbb{E}_{\xi_k}\|\mathbf{x}^{(k+1)}-\widetilde{\mathbf{x}}^{(k)}\|^2 \le{}& (1+c_k)(\delta+\alpha_k\rho)^2\|\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)}\|^2\\
&+(1+\tfrac{1}{c_k})\beta_k^2\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2+\alpha_k^2\sigma^2+\alpha_k^2\rho^2\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2+\mathcal{E},
\end{aligned}\tag{A.13}$$

where c_k is any positive number. Recall $\delta = 1-\alpha_k\overline{\rho}$, and thus

$$(1+c_k)(\delta+\alpha_k\rho)^2 = (1+c_k)\left(1-\alpha_k(\overline{\rho}-\rho)(2-\alpha_k(\overline{\rho}-\rho))\right) \le (1+c_k)\left(1-\alpha_k(\overline{\rho}-\rho)\right) \le 1+c_k-\alpha_k(\overline{\rho}-\rho),$$

where the two inequalities follow from $0<\alpha_k(\overline{\rho}-\rho)<1$ and $c_k>0$. Hence, (A.13) implies

$$\begin{aligned}
\mathbb{E}_{\xi_k}\|\mathbf{x}^{(k+1)}-\widetilde{\mathbf{x}}^{(k)}\|^2 \le{}& \left(1+c_k-\alpha_k(\overline{\rho}-\rho)\right)\|\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)}\|^2\\
&+(1+\tfrac{1}{c_k})\beta_k^2\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2+\alpha_k^2\sigma^2+\alpha_k^2\rho^2\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2+\mathcal{E}.
\end{aligned}\tag{A.14}$$

Below we bound $\mathcal{E}$. We have by the triangle inequality and the ρ-smoothness of F that

$$\begin{aligned}
\mathcal{E} &\le 2\alpha_k\rho\big((\delta+\alpha_k\rho)\|\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)}\|+\beta_k\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|\big)\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|\\
&\le \tfrac{1}{2}\alpha_k(\overline{\rho}-\rho)\|\mathbf{x}^{(k)}-\widetilde{\mathbf{x}}^{(k)}\|^2+\tfrac{2\alpha_k\rho^2}{\overline{\rho}-\rho}\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2\\
&\quad+\beta_k^2\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2+\alpha_k^2\rho^2\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2,
\end{aligned}\tag{A.15}$$

where we have used $\delta+\alpha_k\rho = 1-\alpha_k(\overline{\rho}-\rho)<1$ and the Young's inequality twice to obtain the second inequality. Plug (A.15) into (A.14) and rearrange terms. We obtain (4.1) and complete the proof. □

Lemma A.1. *Let $\{\mathbf{x}^{(k)}\}_{k\ge1}$ and $\{\mathbf{g}^{(k)}\}_{k\ge1}$ be generated from Algorithm 1, and let $\{q_k\}_{k\ge1}$ be a sequence of constants. Under Assumptions 2 and 7, we have*

$$\mathbb{E}\big\|\textstyle\sum_{j=1}^k q_j\mathbf{g}^{(j)}\big\|^2 \le \sum_{l=1}^k q_l\sum_{j=1}^k q_j u_j+\sum_{j=1}^k q_j^2u_j+\sigma^2\sum_{j=1}^k q_j^2.\tag{A.16}$$

Proof of Lemma A.1. From Algorithm 1, we have

$$\textstyle\sum_{j=1}^k q_j\mathbf{g}^{(j)} = \sum_{j=1}^k q_j\nabla f(\mathbf{x}^{(j-\tau_j)};\xi_j).\tag{A.17}$$

Taking a total expectation and using Assumption 2 results in

$$\mathbb{E}\big[\textstyle\sum_{j=1}^k q_j\mathbf{g}^{(j)}\big] = \sum_{j=1}^k q_j\mathbb{E}[\nabla F(\mathbf{x}^{(j-\tau_j)})] = \sum_{j=1}^k q_j\mathbb{E}\mathbf{u}^{(j)},\tag{A.18}$$

which further implies that

$$\left\|\mathbb{E}\big[\textstyle\sum_{j=1}^k q_j \mathbf{g}^{(j)}\big]\right\|^2 \le \sum_{l=1}^k q_l \sum_{j=1}^k q_j \mathbb{E}\|\mathbf{u}^{(j)}\|^2 = \sum_{l=1}^k q_l \sum_{j=1}^k q_j u_j. \tag{A.19}$$

In (A.19), the inequality is obtained by using the triangle inequality, Cauchy-Schwarz inequality, and then Jensen's inequality. We further bound the variance as follows:

$$\begin{aligned}\mathbb{E}\left\|\textstyle\sum_{j=1}^k q_j \mathbf{g}^{(j)} - \mathbb{E}\big[\sum_{j=1}^k q_j \mathbf{g}^{(j)}\big]\right\|^2 &= \mathbb{E}\left\|\textstyle\sum_{j=1}^k q_j \big(\nabla f(\mathbf{x}^{(j-\tau_j)};\xi_j) - \mathbb{E}\mathbf{u}^{(j)}\big)\right\|^2 \\ &= \textstyle\sum_{j=1}^k q_j^2 \mathbb{E}\left\|\nabla f(\mathbf{x}^{(j-\tau_j)};\xi_j) - \mathbb{E}\mathbf{u}^{(j)}\right\|^2 \\ &= \textstyle\sum_{j=1}^k q_j^2 \big(\mathbb{E}\|\nabla f(\mathbf{x}^{(j-\tau_j)};\xi_j) - \mathbf{u}^{(j)}\|^2 + \mathbb{E}\|\mathbf{u}^{(j)} - \mathbb{E}\mathbf{u}^{(j)}\|^2\big) \\ &\le \textstyle\sum_{j=1}^k q_j^2 \big(\sigma^2 + u_j\big). \end{aligned} \tag{A.20}$$

Here, the second equality is because the expectations are null for all cross terms $\mathbb{E}(\mathbf{g}^{(j)}-\mathbb{E}\mathbf{g}^{(j)})^\top(\mathbf{g}^{(j')}-\mathbb{E}\mathbf{g}^{(j')})$ with $j > j'$, since each ξ_j is independent from $\{\mathbf{x}^{(j)},\dots,\mathbf{x}^{(1)}\}$ and $\xi_{j'}$; the third equality is because of Assumption 2; the inequality is by Assumption 7 and that the variance is upper-bounded by the second moment. Combine (A.19) and (A.20) gives (A.16). □

Proof of Lemma 5.1. By definition (5.3), we obtain the equality in (5.4). Then the first inequality in (5.4) follows from

$$\textstyle\sum_{j=1}^{k-1} \pi_{k,j}(t) = \sum_{j=k-\tau_k+1}^{k-1} \frac{1-t^{k-j}}{1-t} + \frac{1-t^{\tau_k}}{1-t}\sum_{j=1}^{k-\tau_k} t^{k-\tau_k-j} = \frac{\tau_k(1-t)-t^{k-\tau_k}(1-t^{\tau_k})}{(1-t)^2} \le \frac{\tau}{1-t},$$

and the second inequality follows from

$$\begin{aligned}\textstyle\sum_{j=1}^{k-1} \pi_{k,j}^2(t) &= \textstyle\sum_{j=k-\tau_k+1}^{k-1} \frac{1-2t^{k-j}+t^{2(k-j)}}{(1-t)^2} + \frac{(1-t^{\tau_k})^2}{(1-t)^2}\sum_{j=1}^{k-\tau_k} t^{2(k-\tau_k-j)} \\ &= \frac{\tau_k(1-t^2)-2(1-t^{\tau_k})(1+t)+(1-t^{2\tau_k})+(1-t^{\tau_k})^2(1-t^{2(k-\tau_k)})}{(1-t)^2(1-t^2)} \le \frac{\tau}{(1-t)^2}.\end{aligned}$$

Proof of Lemma 5.2. From (1.4) and Assumption 2, we have $\mathbf{m}^{(k)} = \sum_{j=1}^k \beta^{k-j}(1-\beta)\mathbf{g}^{(j)}$; apply Lemma A.1 with the choice of $q_j = \beta^{k-j}(1-\beta)$ for all $j \in [k]$ to obtain (5.5) from (A.16). Meanwhile,

$$\begin{aligned}\mathbf{x}^{(k-\tau_k)} - \mathbf{x}^{(k)} &= -\textstyle\sum_{l=0}^{\tau_k-1}\big(\mathbf{x}^{(k-l)} - \mathbf{x}^{(k-l-1)}\big) = \sum_{l=0}^{\tau_k-1} \frac{\alpha_{k-l-1}}{1-\beta}\mathbf{m}^{(k-l-1)} \\ &= \textstyle\sum_{l=0}^{\tau_k-1} \alpha_{k-l-1}\sum_{j=1}^{k-l-1}\beta^{k-l-j-1}\mathbf{g}^{(j)} = \sum_{j=1}^{k-1}\theta_{k,j}\mathbf{g}^{(j)}\end{aligned}$$

by (1.5) and (5.3). Apply Lemma A.1 with the choice of $q_j = \theta_{k,j}$ for all $j \in [k]$. We have (5.6) from (A.16). □

Proof of Lemma 5.3. From (5.7), we have that for $k \ge 1$,

$$\begin{aligned}\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)} &= \tfrac{1}{1-\beta}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) - \tfrac{\beta}{1-\beta}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) \\ &= -\tfrac{1}{(1-\beta)^2}\alpha_k\mathbf{m}^{(k)} + \tfrac{\beta}{(1-\beta)^2}\alpha_{k-1}\mathbf{m}^{(k-1)} \\ &= \tfrac{-1}{(1-\beta)^2}\alpha_k(\beta\mathbf{m}^{(k-1)} + (1-\beta)\mathbf{g}^{(k)}) + \tfrac{\beta}{(1-\beta)^2}\alpha_{k-1}\mathbf{m}^{(k-1)} \\ &= \tfrac{\beta}{(1-\beta)^2}(\alpha_{k-1}-\alpha_k)\mathbf{m}^{(k-1)} - \tfrac{\alpha_k}{1-\beta}\mathbf{g}^{(k)} \\ &= \tfrac{\beta}{1-\beta}(1-\alpha_k/\alpha_{k-1})\tfrac{\alpha_{k-1}}{1-\beta}\mathbf{m}^{(k-1)} - \tfrac{\alpha_k}{1-\beta}\mathbf{g}^{(k)}.\end{aligned}$$

The second equality is by (1.5); the third equality is by (1.4). The above equality together with (1.5) gives (5.8), and (5.9) trivially holds by the smoothness of F and (5.7). □

The inequalities in the lemma below are easy to show.

Lemma A.2. *Let a be a positive integer. Then*

$$\sum_{k=1}^{K}\frac{1}{\sqrt{a+k-1}}\geq\int_{a}^{a+K}\frac{1}{\sqrt{x}}dx=2(\sqrt{a+K}-\sqrt{a}),$$

$$\sum_{k=1}^{K}\frac{1}{a+k-1}\leq 1+\int_{a}^{a+K-1}\frac{1}{x}dx=1+\ln\frac{a+K-1}{a}.$$

Proof of Corollary 5.7. With $\alpha_k=\alpha/\sqrt{a+k-1},\forall\, k\geq 1$, (5.11) holds if and only if

$$\frac{\alpha}{2(1-\beta)\sqrt{a+k-1}}\geq(1-\sqrt{a+k-2}/\sqrt{a+k-1})^2=\frac{1}{((\sqrt{a+k-2}+\sqrt{a+k-1})\sqrt{a+k-1})^2}. \tag{A.21}$$

Notice $\frac{1}{(\sqrt{a+1}+\sqrt{a})^2}\leq\frac{1}{4a}$, and thus $a\sqrt{a+1}\geq\frac{1-\beta}{2\alpha}$ indicates $\alpha\geq\frac{2(1-\beta)}{\sqrt{a+1}(\sqrt{a+1}+\sqrt{a})^2}$, which further implies the inequality in (A.21) for all $k\geq 2$. Moreover, when (5.26) holds, it is not difficult to verify that the two inequalities in (5.25) are true, so we have (5.12) and thus (5.13) from Theorem 5.4.

Below we simplify the inequality in (5.13) for the setting of α_k. First,

$$\begin{aligned}\sum_{k=1}^{K}\alpha_k\alpha^2_{\max\{k-\tau_k,1\}}&\leq\sum_{k=1}^{K}\frac{\alpha^3}{\sqrt{a+k-1}(a+k-1-\tau)}\leq\sum_{k=1}^{K}\frac{2\alpha^3}{\sqrt{a+k-1}(a+k-1)}\\&\leq 2\alpha^3\left(\frac{1}{a\sqrt{a}}+\int_{a}^{a+K-1}\frac{1}{x\sqrt{x}}dx\right)\leq\frac{2\alpha^3(1+2a)}{a\sqrt{a}};\end{aligned}$$

second, by Lemma A.2,

$$\sum_{k=1}^{K}\alpha_k=\sum_{k=1}^{K}\frac{\alpha}{\sqrt{a+k-1}}\geq 2\alpha(\sqrt{a+K}-\sqrt{a}),\ \text{and}\ \sum_{k=1}^{K}\alpha_k^2=\sum_{k=1}^{K}\frac{\alpha^2}{a+k-1}\leq\alpha^2(1+\ln\tfrac{a+K-1}{a}).$$

Substituting the above three inequalities into (5.13) gives (5.27). □

Appendix B. Proof of Theorem 4.7.

The key of the proof is to bound $\sum_k\mathbb{E}[\|\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)}\|^2]$ while using Theorem 4.3. First, similar to (A.7), we have

$$\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\leq\big\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)},\alpha_k(\tilde{\nabla}r(\mathbf{x}^{(k+1)})+\mathbf{g}^{(k)})-\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})\big\rangle, \tag{B.1}$$

where $\tilde{\nabla}r(\mathbf{x}^{(k+1)})$ is a subgradient of r at $\mathbf{x}^{(k+1)}$. By the convexity of r, it holds

$$\big\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)},\tilde{\nabla}r(\mathbf{x}^{(k+1)})\big\rangle\leq r(\mathbf{x}^{(k)})-r(\mathbf{x}^{(k+1)}). \tag{B.2}$$

In addition, from the ρ-smoothness of F and the Young's inequality, we have

$$\begin{aligned}&\big\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)},\mathbf{g}^{(k)}\big\rangle=\big\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)},\nabla F(\mathbf{x}^{(k)})+\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k)})\big\rangle\\&\leq F(\mathbf{x}^{(k)})-F(\mathbf{x}^{(k+1)})+\tfrac{\rho}{2}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)}\|^2+\tfrac{1}{4\alpha_k}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)}\|^2+\alpha_k\|\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k)})\|^2,\end{aligned} \tag{B.3}$$

and

$$\big\langle\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)},-\beta_k(\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)})\big\rangle\leq\tfrac{1}{4}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k+1)}\|^2+\beta_k^2\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2 \tag{B.4}$$

Plugging (B.2), (B.3) and (B.4) into (B.1) and rearranging terms yield

$$\tfrac{1}{2}(1-\alpha_k\rho)\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\leq\alpha_k\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big)+\alpha_k^2\|\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k)})\|^2+\beta_k^2\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2. \tag{B.5}$$

Moreover, by Assumptions 7 and 8 and the ρ-smoothness of F, we have

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k)})\|^2\big]&\leq 2\mathbb{E}\big[\|\mathbf{g}^{(k)}-\nabla F(\mathbf{x}^{(k-\tau_k)})\|^2\big]+2\mathbb{E}\big[\|\nabla F(\mathbf{x}^{(k-\tau_k)})-\nabla F(\mathbf{x}^{(k)})\|^2\big]\\&\leq 2\sigma^2+2\rho^2\mathbb{E}\big[\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2\big]\leq 2\sigma^2+2\tau\rho^2\textstyle\sum_{j=1}^{\tau}\mathbb{E}\big[\|\mathbf{x}^{(k-j)}-\mathbf{x}^{(k-j+1)}\|^2\big].\end{aligned} \tag{B.6}$$

Now taking full expectation on (B.5), substituting (B.6) there, and summing over $k = 1$ to K, we obtain by rearranging terms that

$$\begin{aligned}(\text{B.7})\quad &\textstyle\sum_{k=1}^{K} \frac{1}{2}(1-\alpha_k\rho-\beta_{k+1}^2)\mathbb{E}\big[\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\big]\\ \le\ &\textstyle\sum_{k=1}^{K}\alpha_k\mathbb{E}\big(\phi(\mathbf{x}^{(k)})-\phi(\mathbf{x}^{(k+1)})\big)+2\sigma^2\sum_{k=1}^{K}\alpha_k^2+2\tau\rho^2\sum_{k=1}^{K}\alpha_k^2\sum_{j=1}^{\tau}\mathbb{E}\big[\|\mathbf{x}^{(k-j)}-\mathbf{x}^{(k-j+1)}\|^2\big],\end{aligned}$$

where we have used $\mathbf{x}^{(0)} = \mathbf{x}^{(1)}$. Since α_k is nonincreasing, we have

$$\textstyle\sum_{k=1}^{K}\alpha_k^2\sum_{j=1}^{\tau}\mathbb{E}\big[\|\mathbf{x}^{(k-j)}-\mathbf{x}^{(k-j+1)}\|^2\big]\le\tau\sum_{k=1}^{K}\alpha_k^2\mathbb{E}\big[\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2\big],$$

which substituted into (B.7) and together with (3.23) gives

$$(\text{B.8})\qquad \textstyle\sum_{k=1}^{K}\frac{1}{2}(1-\alpha_k\rho-\beta_{k+1}^2-2\tau^2\rho^2\alpha_{k+1}^2)\mathbb{E}\big[\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\big]\le 2\alpha_1 C_\phi+2\sigma^2\sum_{k=1}^{K}\alpha_k^2.$$

By the choice of parameters and the definition of $\tilde{\gamma}$ in (4.17), we have from (B.8) and Lemma A.2 that

$$(\text{B.9})\qquad \textstyle\sum_{k=1}^{K}\mathbb{E}\big[\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\|^2\big]\le\frac{2}{\tilde{\gamma}}\left(\alpha_1 C_\phi+\sigma^2\alpha^2(1+\ln\tfrac{a+K-1}{a})\right).$$

Notice $\big(2+\frac{4}{\alpha_k(\overline{\rho}-\rho)}\big)\beta_k^2\le 2\tilde{\beta}^2+\frac{4\beta^2}{\alpha(\overline{\rho}-\rho)}$ and $\alpha_k^2+\frac{\alpha_k}{\overline{\rho}-\rho}\le\frac{\alpha^2}{a}+\frac{\alpha}{\sqrt{a}(\overline{\rho}-\rho)}, \forall\, k\ge 1$. Therefore,

$$\begin{aligned}&\tfrac{\overline{\rho}}{2}\textstyle\sum_{k=1}^{K}\big(2+\frac{4}{\alpha_k(\overline{\rho}-\rho)}\big)\beta_k^2\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2+\overline{\rho}\rho^2\textstyle\sum_{k=1}^{K}\big(\alpha_k^2+\frac{\alpha_k}{\overline{\rho}-\rho}\big)\mathbb{E}\|\mathbf{x}^{(k-\tau_k)}-\mathbf{x}^{(k)}\|^2\\ \le\ &\Big(\tfrac{\overline{\rho}}{2}\big(2\tilde{\beta}^2+\tfrac{4\beta^2}{\alpha(\overline{\rho}-\rho)}\big)+\tau^2\overline{\rho}\rho^2\big(\tfrac{\alpha^2}{a}+\tfrac{\alpha}{\sqrt{a}(\overline{\rho}-\rho)}\big)\Big)\textstyle\sum_{k=1}^{K}\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|^2\\ \le\ &\Big(\tfrac{\overline{\rho}}{2}\big(2\tilde{\beta}^2+\tfrac{4\beta^2}{\alpha(\overline{\rho}-\rho)}\big)+\tau^2\overline{\rho}\rho^2\big(\tfrac{\alpha^2}{a}+\tfrac{\alpha}{\sqrt{a}(\overline{\rho}-\rho)}\big)\Big)\tfrac{2}{\tilde{\gamma}}\left(\alpha_1 C_\phi+\sigma^2\alpha^2(1+\ln\tfrac{a+K-1}{a})\right).\end{aligned}\qquad(\text{B.10})$$

Now plug (B.10) and the choice of $\{\alpha_k\}$ into (4.2) to obtain the desired result. □

REFERENCES

[1] A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

[2] A. Alacaoglu, Y. Malitsky, and V. Cevher. Convergence of adaptive algorithms for weakly convex constrained optimization. *arXiv preprint arXiv:2006.06650*, 2020.

[3] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

[4] F. Alvarez. Weak convergence of a relaxed and inertial hybrid projection-proximal point algorithm for maximal monotone operators in hilbert space. *SIAM Journal on Optimization*, 14(3):773–782, 2004.

[5] F. Alvarez and H. Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1):3–11, 2001.

[6] K. Bäckström, M. Papatriantafilou, and P. Tsigas. Mindthestep-asyncpsgd: Adaptive asynchronous parallel stochastic gradient descent. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 16–25. IEEE, 2019.

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[8] T. F. Chan and C.-K. Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3):370–375, 1998.

[9] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[10] S. Chen, A. Garcia, and S. Shahrampour. Distributed projected subgradient method for weakly convex optimization. *arXiv preprint arXiv:2004.13233*, 2020.

[11] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019.

[12] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020.
[13] T. T. Doan, C. L. Beck, and R. Srikant. Convergence rate of distributed subgradient methods under communication delays. In *2018 Annual American Control Conference (ACC)*, pages 5310–5315. IEEE, 2018.
[14] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.
[15] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
[16] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.
[17] Y. C. Eldar and S. Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.
[18] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.
[19] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.
[20] I. Gitman, H. Lang, P. Zhang, and L. Xiao. Understanding the role of momentum in stochastic gradient methods. *arXiv preprint arXiv:1910.13962*, 2019.
[21] N. Guan, D. Tao, Z. Luo, and B. Yuan. Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1087–1099, 2012.
[22] R. Hannah and W. Yin. On unbounded delays in asynchronous parallel fixed-point algorithms. *Journal of Scientific Computing*, 76(1):299–326, 2018.
[23] Z. Huo and H. Huang. Asynchronous mini-batch gradient descent with variance reduction for non-convex optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
[24] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.
[25] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.
[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
[27] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding blind deconvolution algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2354–2367, 2011.
[28] X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.
[29] X. Lian, W. Zhang, C. Zhang, and J. Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pages 3043–3052, 2018.
[30] J. Liang, J. Fadili, and G. Peyré. A multi-step inertial forward-backward splitting method for non-convex optimization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
[31] N. Loizou and P. Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *arXiv preprint arXiv:1710.10737*, 2017.
[32] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.
[33] V. Mai and M. Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning*, pages 6630–6639. PMLR, 2020.
[34] I. Masubuchi, J. Tsukamoto, T. Wada, R. Morita, T. Asai, Y. Ohta, and Y. Fujisaki. Distributed multi-agent optimization with local constraints via a subgradient method with delayed information of feasibility. In *Proceedings of the 21st International Symposium on Mathematical Theory of Networks and Systems*, pages 23–28, 2014.
[35] B. McMahan and M. Streeter. Delay-tolerant algorithms for asynchronous distributed online learning. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
[36] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.
[37] A. Moudafi and M. Oliny. Convergence of a splitting inertial proximal method for monotone operators. *Journal of Computational and Applied Mathematics*, 155(2):447–454, 2003.
[38] S. D. Muruganathan, D. C. Ma, R. I. Bhasin, and A. O. Fapojuwo. A centralized energy-efficient routing

protocol for wireless sensor networks. *IEEE Communications Magazine*, 43(3):S8–13, 2005.
[39] P. Nazari, D. A. Tarzanagh, and G. Michailidis. Adaptive first-and zeroth-order methods for weakly convex stochastic optimization problems. *arXiv preprint arXiv:2005.09261*, 2020.
[40] A. V. Nazin. Algorithms of inertial mirror descent in convex problems of stochastic optimization. *Automation and Remote Control*, 79(1):78–88, 2018.
[41] A. Nedić, D. P. Bertsekas, and V. S. Borkar. Distributed asynchronous incremental subgradient methods. *Studies in Computational Mathematics*, 8(C):381–407, 2001.
[42] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.
[43] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.
[44] P. Ochs. Local convergence of the heavy-ball method and ipiano for non-convex optimization. *Journal of Optimization Theory and Applications*, 177(1):153–180, 2018.
[45] P. Ochs, T. Brox, and T. Pock. ipiasco: inertial proximal algorithm for strongly convex optimization. *Journal of Mathematical Imaging and Vision*, 53(2):171–181, 2015.
[46] P. Ochs, Y. Chen, T. Brox, and T. Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.
[47] Z. Peng, Y. Xu, M. Yan, and W. Yin. On the convergence of asynchronous parallel iteration with unbounded delays. *Journal of the Operations Research Society of China*, 7(1):5–42, 2019.
[48] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.
[49] B. T. Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.
[50] B. Recht. Cs726-lyapunov analysis and the heavy ball method. 2010.
[51] 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*, pages 693–701, 2011.
[52] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
[53] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.
[54] J. V. Shi, Y. Xu, and R. G. Baraniuk. Sparse bilinear logistic regression. *arXiv preprint arXiv:1404.4104*, 2014.
[55] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
[56] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning*. Mit Press, 2012.
[57] S. Sra, A. W. Yu, M. Li, and A. Smola. Adadelay: Delay adaptive distributed stochastic optimization. In *Artificial Intelligence and Statistics*, pages 957–965, 2016.
[58] T. Sun, D. Li, Z. Quan, H. Jiang, S. Li, and Y. Dou. Heavy-ball algorithms always escape saddle points. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3520–3526. International Joint Conferences on Artificial Intelligence Organization, 7 2019.
[59] T. Sun, L. Qiao, and D. Li. Nonergodic complexity of proximal inertial gradient descents. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
[60] T. Sun, P. Yin, D. Li, C. Huang, L. Guan, and H. Jiang. Non-ergodic convergence analysis of heavy-ball algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5033–5040, 2019.
[61] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.
[62] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming (online first)*, 2021.
[63] P. Tseng. An incremental gradient (-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.
[64] H. Wang, X. Liao, T. Huang, and C. Li. Cooperative distributed optimization in multiagent networks with delays. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(2):363–369, 2014.
[65] Y. Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *arXiv preprint arXiv:2006.00425*, 2020.
[66] Y. Xu, Y. Xu, Y. Yan, C. Sutcher-Shepard, L. Grinberg, and J. Chen. Parallel and distributed asynchronous

adaptive stochastic gradient methods. *arXiv preprint arXiv:2002.09095*, 2020.

[67] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2955–2961. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[68] S. Zavriev and F. Kostyuk. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 4(4):336–341, 1993.