Noname manuscript No.
(will be inserted by the editor)

On the Convergence of Asynchronous Parallel Iteration with Unbounded Delays

Zhimin Peng · Yangyang Xu · Ming Yan · Wotao Yin

the date of receipt and acceptance should be inserted later

Abstract Recent years have witnessed the surge of asynchronous parallel (async-parallel) iterative algorithms due to problems involving very large-scale data and a large number of decision variables. Because of asynchrony, the iterates are computed with outdated information, and the age of the outdated information, which we call *delay*, is the number of times it has been updated since its creation. Almost all recent works prove convergence under the assumption of a finite maximum delay and set their stepsize parameters accordingly. However, the maximum delay is practically unknown.

This paper presents convergence analysis of an async-parallel method from a probabilistic viewpoint, and it allows for large unbounded delays. An explicit formula of stepsize that guarantees convergence is given depending on delays' statistics. With $p+1$ identical processors, we empirically measured that delays closely follow the Poisson distribution with parameter p, matching our theoretical model, and thus the stepsize can be set accordingly. Simulations on both convex and nonconvex optimization problems demonstrate the validness of our analysis and also show that the existing maximum-delay induced stepsize is too conservative, often slowing down the convergence of the algorithm.

Keywords asynchronous unbounded delays, nonconvex, convex

Z. Peng and W. Yin
Department of Mathematics, University of California, Los Angeles, CA 90095
E-mail: zhiminp@gmail.com / wotaoyin@math.ucla.edu

Yangyang Xu
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180
E-mail: xuy21@rpi.edu

Ming Yan
Department of Computational Mathematics, Science and Engineering, Department of Mathematics, Michigan State University, East Lansing, MI 48824
E-mail: yanm@math.msu.edu

1 Introduction

In the "big data" era, the size of the dataset and the number of decision variables involved in many areas such as health care, the Internet, economics, and engineering are becoming tremendously large [34]. It motivates the development of new computational approaches by efficiently utilizing modern multi-core computers or computing clusters.

In this paper, we consider the block-structured optimization problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, F(\mathbf{x}) \equiv f(\mathbf{x}_1,\ldots,\mathbf{x}_m) + \sum_{i=1}^{m} r_i(\mathbf{x}_i), \tag{1}$$

where $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_m)$ is partitioned into m disjoint blocks, f has a Lipschitz continuous gradient (possibly nonconvex), and r_i's are (possibly nondifferentiable) proper closed convex functions. Note that r_i's can be extended-valued, and thus (1) can have block constraints $\mathbf{x}_i \in X_i$ by incorporating the indicator function of X_i in r_i for all i.

Many applications can be formulated in the form of (1), and they include classic machine learning problems: support vector machine (squared hinge loss and its dual formulation) [6], LASSO [30], and logistic regression (linear or multilinear) [37], and also subspace learning problems: sparse principal component analysis [38], nonnegative matrix or tensor factorization [5], just to name a few.

Toward solutions for these problems with extremely large-scale datasets and many variables, first-order methods and also stochastic methods become particularly popular because of their scalability to the problem size, such as FISTA [1], stochastic approximation [21], randomized coordinate descent [22], and their combinations [7,35]. Recently, lots of efforts have been made to the parallelization of these methods, and in particular, asynchronous parallel (async-parallel) methods attract more attention (e.g., [16,24]) over their synchronous counterparts partly due to the better speed-up performance.

This paper focuses on the async-parallel block coordinate update (async-BCU) method (see Algorithm 1) for solving (1). To the best of our knowledge, all works on async-BCU before 2013 consider a deterministic selection of blocks with an exception to [29], and thus they require strong conditions (like a contraction) for convergence. Recent works, e.g., [16,17,24,12], employ randomized block selection and significantly weaken the convergence requirement. However, all of them require bounded delays and/or are restricted to convex problems. The work [12] allows unbounded delays but requires convexity, and [8,3] do not assume convexity but require bounded delays. We consider unbounded delays and deal with nonconvex problems.

1.1 Algorithm

We describe the async-BCU method as follows. Assume there are $p+1$ processors, and the data and variable $\mathbf{x}$ are accessible to all processors. We let

all processors continuously and asynchronously update the variable $\mathbf{x}$ in parallel. At each time k, one processor reads the variable $\mathbf{x}$ as $\hat{\mathbf{x}}^k$ from the global memory, randomly picks a block $i_k \in \{1, 2, \cdots, m\}$, and renews $\mathbf{x}_{i_k}$ by a prox-linear update while keeping all the other blocks unchanged. The pseudocode is summarized in Algorithm 1, where the **prox** operator is defined in (3).

The algorithm first appeared in [16], where the age of $\hat{\mathbf{x}}^k$ relative to $\mathbf{x}^k$, which we call the *delay* of iteration k, was assumed to be bounded by a certain integer τ. For general convex problems, sublinear convergence was established, and for the strongly convex case, linear convergence was shown. However, its convergence for nonconvex problems and/or with unbounded delays was unknown. In addition, numerically, the stepsize is difficult to tune because it depends on τ, which is unknown before the algorithm completes.

Algorithm 1: Async-parallel block coordinate update

Input : Any point $\mathbf{x}^0 \in \mathbb{R}^n$ in the global memory, maximum number of iterations K, stepsize $\eta > 0$

while $k < K$, *each and all processors asynchronously* **do**

select i_k from $[m]$ uniformly at random;

$\hat{\mathbf{x}}^k \leftarrow$ read $\mathbf{x}$ from the global memory;

for all $i \in [m]$,

$$\mathbf{x}_i^{k+1} \leftarrow \begin{cases} \mathbf{prox}_{\eta r_i}\left(\mathbf{x}_i^k - \eta \nabla_i f(\hat{\mathbf{x}}^k)\right), & \text{if } i = i_k, \\ \mathbf{x}_i^k, & \text{otherwise;} \end{cases} \tag{2}$$

increase the global counter $k \leftarrow k + 1$;

end

1.2 Contributions

We summarize our contributions as follows.

- We analyze the convergence of Algorithm 1 and allow for large unbounded delays following a certain distribution. We require the delays to have certain bounded expected quantities (e.g., expected delay, variance of delay). Our results are more general than those requiring bounded delays such as [16, 17].
- Both nonconvex and convex problems are analyzed, and those problems include both smooth and nonsmooth functions. For nonconvex problems, we establish the global convergence in terms of first-order optimality conditions and show that any limit point of the iterates is a critical point almost surely. It appears to be the first result of an async-BCU method for general nonconvex problems and allowing unbounded delays. For weakly convex problems, we establish a sublinear convergence result, and for strongly convex problems, we show the linear convergence.

- We show that if all $p+1$ processors run at the same speed, the delay follows the Poisson distribution with parameter p. In this case, all the relevant expected quantities can be explicitly computed and are bounded. By setting appropriate stepsizes, we can reach a near-linear speedup if $p = o(\sqrt{m})$ for smooth cases and $p = o(\sqrt[4]{m})$ for nonsmooth cases.
- When the delay follows the Poisson distribution, we can explicitly set the stepsize based on the delay expectation (which equals p). We simulate the async-BCU method on one convex problem: LASSO, and one nonconvex problem: the nonnegative matrix factorization. The results demonstrate that async-BCU performs consistently better with a stepsize set based on the expected delay than on the maximum delay. The number of processors is known while the maximum delay is not. Hence, the setting based on expected delay is practically more useful.

Our algorithm updates one (block) coordinate of $\mathbf{x}$ in each step and is sharply different from stochastic gradient methods that sample one function in each step to update all coordinates of $\mathbf{x}$. While there are async-parallel algorithms in either classes and how to handle delays is important to both of their convergence, their basic lines of analysis are different with respect to how to absorb the delay-induced errors. The results of the two classes are in general not comparable. That said, for problems with certain proper structures, it is possible to apply both coordinate-wise update and stochastic sampling (e.g., [25,35, 20,8]), and our results apply to the coordinate part.

1.3 Notation and assumptions

Throughout the paper, bold lowercase letters $\mathbf{x}, \mathbf{y}, \ldots$, are used for vectors. We denote $\mathbf{x}_i$ as the i-th block of $\mathbf{x}$ and U_i as the i-th sampling matrix, i.e., $U_i\mathbf{x}$ is a vector with $\mathbf{x}_i$ as its i-th block and $\mathbf{0}$ for the remaining ones. $\mathbb{E}_{i_k}$ denotes the expectation with respect to i_k conditionally on all previous history, and $[m] = \{1, \ldots, m\}$.

We consider the Euclidean norm denoted by $\|\cdot\|$, but all our results can be directly extended to problems with general primal and dual norms in a Hilbert space.

The projection to a convex set X is defined as

$$\mathcal{P}_X(\mathbf{y}) = \arg\min_{\mathbf{x}\in X} \|\mathbf{x}-\mathbf{y}\|^2,$$

and the proximal mapping of a convex function h is defined as

$$\mathbf{prox}_h(\mathbf{y}) = \arg\min_{\mathbf{x}} h(\mathbf{x}) + \tfrac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2. \tag{3}$$

Definition 1 ***(Critical point)*** A point $\mathbf{x}^*$ is a critical point of (1) if $\mathbf{0} \in \nabla f(\mathbf{x}^*) + \partial R(\mathbf{x}^*)$, where $\partial R(\mathbf{x})$ denotes the subdifferential of R at $\mathbf{x}$ and

$$R(\mathbf{x}) = \sum_{i=1}^{m} r_i(\mathbf{x}_i). \tag{4}$$

Throughout our analysis, we make the following three assumptions to problem (1) and Algorithm 1. Other assumed conditions will be specified if needed.

Assumption 1 *The function F is lower bounded. The problem* (1) *has at least one solution, and the solution set is denoted as X^*.*

Assumption 2 *$\nabla f(\mathbf{x})$ is Lipschitz continuous with constant L_f, namely,*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_f \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \ \mathbf{y}. \tag{5}$$

In addition, for each $i \in [m]$, fixing all block coordinates but the i-th one, $\nabla f(\mathbf{x})$ and $\nabla_i f(\mathbf{x})$ are Lipschitz continuous about $\mathbf{x}_i$ with L_r and L_c, respectively, i.e., for any $\mathbf{x}$, $\mathbf{y}$, and i,

$$\begin{aligned} &\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x} + U_i \mathbf{y})\| \leq L_r \|\mathbf{y}_i\|, \\ &\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{x} + U_i \mathbf{y})\| \leq L_c \|\mathbf{y}_i\|. \end{aligned} \tag{6}$$

From (6), we have that for any $\mathbf{x}$, $\mathbf{y}$, and i,

$$f(\mathbf{x} + U_i \mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla_i f(\mathbf{x}), \mathbf{y}_i \rangle + \tfrac{L_c}{2} \|\mathbf{y}_i\|^2. \tag{7}$$

We denote $\kappa = \frac{L_r}{L_c}$ as the condition number.

Assumption 3 *For each $k \geq 1$, the reading $\hat{\mathbf{x}}^k$ is consistent and delayed by j_k, namely, $\hat{\mathbf{x}}^k = \mathbf{x}^{k - j_k}$. The delay j_k follows an identical distribution as a random variable $\mathbf{j}$*

$$\text{Prob}(\mathbf{j} = t) = q_t, \ t = 0, 1, 2 \ldots, \tag{8}$$

and is independent of i_k. We let

$$c_k := \textstyle\sum_{t=k}^{\infty} q_t, \qquad T := \mathbb{E}[\mathbf{j}], \qquad S := \mathbb{E}[\mathbf{j}^2].$$

Remark 1 Although the delay always satisfies $0 \leq j_k \leq k$, the assumption in (8) is without loss of generality if we make negative iterates and regard $\mathbf{x}^k = \mathbf{x}^0$, $\forall k < 0$. For simplicity, we make the identical distribution assumption, which is the same as that in *[29]*. Our results can still hold for non-identical distribution; see the analysis for the smooth nonconvex case in the arXiv version of the paper.

2 Related works

We briefly review block coordinate update (BCU) and async-parallel computing methods.

The BCU method is closely related to the Gauss-Seidel method for solving linear equations, which can date back to 1823. In the literature of optimization, BCU method first appeared in [13] as the block coordinate descent method, or more precisely, block minimization (BM), for quadratic programming. The convergence of BM was established early for both convex and nonconvex problems, for example [19,10,31]. However, in general, its convergence

rate result was only shown for strongly convex problems (e.g., [19]) until the recent work [14] that shows sublinear convergence for weakly convex cases. [33] proposed a new version of BCU methods, called coordinate gradient descent method, which mimics proximal gradient descent but only updates a block coordinate every time. The block coordinate gradient or block prox-linear update (BPU) becomes popular since [22] proposed to randomly select a block to update. The convergence rate of the randomized BPU is easier to show than the deterministic BPU. It was firstly established for convex smooth problems (both unconstrained and constrained) in [22] and then generalized to nonsmooth cases in [26,18]. Recently, [7,35] incorporated stochastic approximation into the BPU framework to deal with stochastic programming, and both established sublinear convergence for convex problems and also global convergence for nonconvex problems.

The async-parallel computing method (also called *chaotic relaxation*) first appeared in [28] to solve linear equations arising in electrical network problems. [4] first systematically analyzed (more general) asynchronous iterative methods for solving linear systems. Assuming bounded delays, it gave a necessary and sufficient condition for convergence. [2] proposed an asynchronous distributed iterative method for solving more general fixed-point problems and showed its convergence under a contraction assumption. [32] weakened the contraction assumption to pseudo-nonexpansiveness but made more other assumptions. [9] made a thorough review of asynchronous methods before 2000. It summarized convergence results under nested sets and synchronous convergence conditions, which are satisfied by P-contraction mappings and isotone mappings.

Since it was proposed in 1969, the async-parallel method has not attracted much attention until recent years when the size of data is increasing exponentially in many areas. Motivated by "big data" problems, [16,17] proposed the async-parallel stochastic coordinate descent method (i.e., Algorithm 1) for solving problems in the form of (1). Their analysis focuses on convex problems and assumes bounded delays. Specifically, they established sublinear convergence for weakly convex problems and linear convergence for strongly convex problems. In addition, near-linear speed up was achieved if $\tau = o(\sqrt{m})$ for unconstrained smooth convex problems and $\tau = o(\sqrt[4]{m})$ for constrained smooth or nonsmooth cases. For nonconvex problems, [8] introduced an async-parallel coordinate descent method, whose convergence was established under iterate boundedness assumptions and appropriate stepsizes.

3 Convergence results for the smooth case

Throughout this section, let $r_i = 0$, $\forall i$, i.e., we consider the smooth optimization problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\mathbf{x}). \tag{9}$$

The general (possibly nonsmooth) case will be analyzed in the next section. The results for nonsmooth problems of course also hold for smooth ones. How-

ever, the smooth case requires weaker conditions for convergence than those required by the nonsmooth case, and their analysis techniques are different. Hence, we consider the two cases separately.

3.1 Convergence for the nonconvex case

In this subsection, we establish a subsequence convergence result for the general (possibly nonconvex) case. We begin with some technical lemmas. The first lemma deals with certain infinite sums that will appear later in our analysis.

Lemma 1 *For any k and $t \le k$, let*

$$\gamma_k = \tfrac{\eta^2 L_r}{2m\sqrt{m}} \textstyle\sum_{d=1}^{k-1}(c_{k-d} - c_k)c_d + \tfrac{\eta}{2m}c_k + \tfrac{\eta^2 L_c}{2m}c_k, \tag{10a}$$

$$\beta_k = \left(\tfrac{\eta}{m} - \tfrac{\eta^2 L_c}{2m}\right) q_0 - \tfrac{\eta}{2m}c_k \text{ for } k \ge 1, \quad (\text{and } \beta_0 = 0)\,, \tag{10b}$$

$$C_{t,k} = \left(\tfrac{\eta}{m} - \tfrac{\eta^2 L_c}{2m}\right) q_t - \tfrac{\eta^2 L_r}{2m\sqrt{m}} \left(tq_t + \textstyle\sum_{d=1}^{t}(c_d - c_k)q_{t-d}\right). \tag{10c}$$

Then

$$\textstyle\sum_{k=0}^{\infty} \gamma_k \le \tfrac{\eta^2 L_r}{2m\sqrt{m}} T^2 + \left(\tfrac{\eta}{2m} + \tfrac{\eta^2 L_c}{2m}\right)(1+T), \tag{11}$$

$$\beta_k + \textstyle\sum_{t=k+1}^{\infty} C_{t-k,t} \ge \tfrac{\eta}{2m} - \tfrac{\eta^2 L_c}{2m} - \tfrac{\eta^2 L_r T}{m\sqrt{m}}, \ \forall k. \tag{12}$$

Proof To bound $\sum_{k=0}^{\infty} \gamma_k$, we bound the first term $\sum_{d=1}^{k-1}(c_{k-d}-c_k)c_d$ in (10a). Specifically,

$$\sum_{k=0}^{\infty}\sum_{d=1}^{k-1}(c_{k-d} - c_k)c_d \le \sum_{k=0}^{\infty}\sum_{d=1}^{k-1} c_{k-d}c_d = \sum_{d=1}^{\infty}\sum_{k=d+1}^{\infty} c_{k-d}c_d = T^2,$$

where the last equality holds since $T := \mathbb{E}[\mathbf{j}] = \sum_{t=1}^{\infty} tq_t = \sum_{t=1}^{\infty}\sum_{d=1}^{t} q_t = \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} q_t = \sum_{d=1}^{\infty} c_d$. We obtain (11) by combining these two equations.

To prove (12), we will use

$$\sum_{t=1}^{\infty}\sum_{d=1}^{t}(c_d - c_{k+t})q_{t-d} \le \sum_{t=1}^{\infty}\sum_{d=1}^{t} c_d q_{t-d} = \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} c_d q_{t-d} = \sum_{d=1}^{\infty} c_d = T. \tag{13}$$

The above inequality yields

$$\begin{aligned}
&\beta_k + \textstyle\sum_{t=k+1}^{\infty} C_{t-k,t} = \beta_k + \sum_{t=1}^{\infty} C_{t,t+k} \\
=&\left(\tfrac{\eta}{m} - \tfrac{\eta^2 L_c}{2m}\right) q_0 - \tfrac{\eta}{2m}c_k \\
&+ \textstyle\sum_{t=1}^{\infty}\left(\left(\tfrac{\eta}{m} - \tfrac{\eta^2 L_c}{2m}\right) q_t - \tfrac{\eta^2 L_r}{2m\sqrt{m}} tq_t - \tfrac{\eta^2 L_r}{2m\sqrt{m}}\sum_{d=1}^{t}(c_d - c_{k+t})q_{t-d}\right) \\
\overset{(13)}{\ge}& \tfrac{\eta}{m} - \tfrac{\eta^2 L_c}{2m} - \tfrac{\eta}{2m}c_k - \tfrac{\eta^2 L_r T}{m\sqrt{m}} \ge \tfrac{\eta}{2m} - \tfrac{\eta^2 L_c}{2m} - \tfrac{\eta^2 L_r T}{m\sqrt{m}},
\end{aligned}$$

where the last inequality follows from $c_k \le 1$. □

The second lemma below bounds the cross term that appears in our analysis.

Lemma 2 (Cross term bound) *For any $k > 1$ and $t \leq k$, it holds that*

$$\begin{aligned} &\textstyle\sum_{t=1}^{k-1} q_t \mathbb{E}\Big[-\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t}), \nabla f(\mathbf{x}^{k-t})\rangle\Big] \\ \leq &\tfrac{\eta L_r}{2\sqrt{m}} \textstyle\sum_{t=1}^{k-1} \left(t q_t + \sum_{d=1}^{t} (c_d - c_k) q_{t-d} \right) \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 \\ &+ \tfrac{\eta L_r}{2\sqrt{m}} \textstyle\sum_{d=1}^{k-1} (c_{k-d} - c_k) c_d \|\nabla f(\mathbf{x}^0)\|^2. \end{aligned} \tag{14}$$

Proof Define $\Delta^d := \nabla f(\mathbf{x}^d) - \nabla f(\mathbf{x}^{d+1})$. Applying the Cauchy-Schwarz inequality with $\nabla f(\mathbf{x}^{k-t}) - \nabla f(\mathbf{x}^k) = \sum_{d=k-t}^{k-1} \Delta^d$ yields

$$-\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t}), \nabla f(\mathbf{x}^{k-t})\rangle \leq \textstyle\sum_{d=k-t}^{k-1} \|\Delta^d\| \cdot \|\nabla f(\mathbf{x}^{k-t})\|.$$

Since $\|\Delta^d\| \leq L_r \|\mathbf{x}^{d+1} - \mathbf{x}^d\| = \eta L_r \|\nabla_{i_d} f(\hat{\mathbf{x}}^d)\|$, by applying Young's inequality, we get

$$\begin{aligned} &-\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t}), \nabla f(\mathbf{x}^{k-t})\rangle \\ \leq &\tfrac{\eta L_r}{2} \textstyle\sum_{d=k-t}^{k-1} \Big(\sqrt{m} \|\nabla_{i_d} f(\hat{\mathbf{x}}^d)\|^2 + \tfrac{1}{\sqrt{m}} \|\nabla f(\mathbf{x}^{k-t})\|^2 \Big). \end{aligned} \tag{15}$$

By taking expectation, we have

$$\begin{aligned} \mathbb{E}_{i_d, j_d} \|\nabla_{i_d} f(\hat{\mathbf{x}}^d)\|^2 &= \tfrac{1}{m} \mathbb{E}_{j_d} \|\nabla f(\mathbf{x}^{d-j_d})\|^2 \\ &= \tfrac{1}{m} \Big(\textstyle\sum_{r=0}^{d-1} q_r \|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d \|\nabla f(\mathbf{x}^0)\|^2 \Big). \end{aligned}$$

Now taking expectation on both sides of (15) and using the above equation, we get

$$\begin{aligned} &\mathbb{E}[-\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t}), \nabla f(\mathbf{x}^{k-t})\rangle] \\ \leq &\tfrac{\eta L_r}{2\sqrt{m}} \textstyle\sum_{d=k-t}^{k-1} \Big(\sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d \|\nabla f(\mathbf{x}^0)\|^2 \Big) \\ &+ \tfrac{\eta L_r}{2\sqrt{m}} \textstyle\sum_{d=k-t}^{k-1} t \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2. \end{aligned} \tag{16}$$

Finally, (14) follows from

$$\begin{aligned} \textstyle\sum_{t=1}^{k-1} q_t \sum_{d=k-t}^{k-1} c_d \|\nabla f(\mathbf{x}^0)\|^2 &\overset{(83)}{=} \textstyle\sum_{d=1}^{k-1} \Big(\sum_{t=k-d}^{k-1} q_t \Big) c_d \|\nabla f(\mathbf{x}^0)\|^2 \\ &= \textstyle\sum_{d=1}^{k-1} (c_{k-d} - c_k) c_d \|\nabla f(\mathbf{x}^0)\|^2, \end{aligned}$$

and

$$\begin{aligned}
&\sum_{t=1}^{k-1} q_t \sum_{d=k-t}^{k-1} \sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 \\
=&\sum_{d=1}^{k-1} (c_{k-d} - c_k) \sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 \\
[\text{let } r \leftarrow d-r] \ =&\sum_{d=1}^{k-1} (c_{k-d} - c_k) \sum_{r=1}^{d} q_{d-r} \mathbb{E}\|\nabla f(\mathbf{x}^{r})\|^2 \\
\overset{(84)}{=}&\sum_{r=1}^{k-1} \left(\sum_{d=r}^{k-1} (c_{k-d} - c_k) q_{d-r} \right) \mathbb{E}\|\nabla f(\mathbf{x}^{r})\|^2 \\
[\text{let } t \leftarrow k-r,\ d \leftarrow k-d] \ =&\sum_{t=1}^{k-1} \left(\sum_{d=1}^{t} (c_d - c_k) q_{t-d} \right) \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2. \quad (17)
\end{aligned}$$

□

Using the above lemma, we show a result of running one iteration of the algorithm.

Theorem 1 (Fundamental bound) *Set γ_k, β_k and $C_{t,k}$ as in* (10)*. For any $k > 1$, we have*

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{k+1}) \leq& \mathbb{E}f(\mathbf{x}^k) + \gamma_k \|\nabla f(\mathbf{x}^0)\|^2 - \beta_k \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
&- \textstyle\sum_{t=1}^{k-1} C_{t,k} \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2. \quad (18)
\end{aligned}$$

Proof Since $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta U_{i_k} \nabla f(\mathbf{x}^{k-j_k})$, we have from (7) that

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \eta \langle \nabla f(\mathbf{x}^k), U_{i_k} \nabla f(\mathbf{x}^{k-j_k}) \rangle + \tfrac{L_c}{2} \|\eta U_{i_k} \nabla f(\mathbf{x}^{k-j_k})\|^2.$$

Taking conditional expectation on (i_k, j_k) gives

$$\begin{aligned}
&\mathbb{E}_{i_k,j_k} f(\mathbf{x}^{k+1}) \\
\leq& f(\mathbf{x}^k) - \tfrac{\eta}{m} \mathbb{E}_{j_k} \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-j_k}) \rangle + \tfrac{\eta^2 L_c}{2m} \mathbb{E}_{j_k} \|\nabla f(\mathbf{x}^{k-j_k})\|^2 \\
=& f(\mathbf{x}^k) - \tfrac{\eta}{m} \textstyle\sum_{t=0}^{k-1} q_t \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t}) \rangle - \tfrac{\eta}{m} c_k \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0) \rangle \\
&+ \tfrac{\eta^2 L_c}{2m} \textstyle\sum_{t=0}^{k-1} q_t \|\nabla f(\mathbf{x}^{k-t})\|^2 + \tfrac{\eta^2 L_c}{2m} c_k \|\nabla f(\mathbf{x}^0)\|^2. \quad (19)
\end{aligned}$$

For the first cross term in (19), we write each summand as

$$\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t}) \rangle = \langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t}), \nabla f(\mathbf{x}^{k-t}) \rangle + \|\nabla f(\mathbf{x}^{k-t})\|^2, \quad (20)$$

and we use Young's inequality to bound the second cross term by

$$-\tfrac{\eta}{m} c_k \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0) \rangle \leq \tfrac{\eta c_k}{2m} \left[\|\nabla f(\mathbf{x}^k)\|^2 + \|\nabla f(\mathbf{x}^0)\|^2 \right]. \quad (21)$$

Now taking expectation over both sides of (19), plugging in (20) and (21), and using Lemma 2, we have the desired result. □

We are now ready to show the main result in the following theorem.

Theorem 2 (Convergence for the nonconvex smooth case) *Under Assumptions 1 through 3, let* $\{\mathbf{x}^k\}_{k\geq 1}$ *be generated from Algorithm 1. Assume* $T<\infty$. *Take the stepsize as* $0<\eta<\frac{1/L_c}{1+2\kappa T/\sqrt{m}}$. *If* $q_0>0$ *or* $\nabla f(\mathbf{x})$ *is bounded for all* $\mathbf{x}$, *then*

$$\lim_{k\to\infty}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|=0, \tag{22}$$

and any limit point of $\{\mathbf{x}^k\}_{k\geq 1}$ *is almost surely a critical point of* (9).

Remark 2 If $T=\mathbb{E}[\mathbf{j}]=o(\sqrt{m})$, then η only weakly depends on the delay. The conditions $q_0>0$ or $\nabla f(\mathbf{x})$ being bounded can be dropped if $S=\mathbb{E}[\mathbf{j}^2]$ is bounded; see Theorem 5.

Proof Summing up (18) from $k=0$ through K and using (85), we have

$$\begin{aligned}\mathbb{E}f(\mathbf{x}^{K+1})\leq& f(\mathbf{x}^0)+\textstyle\sum_{k=0}^{K}\gamma_k\|\nabla f(\mathbf{x}^0)\|^2\\ &-\beta_K\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2-\textstyle\sum_{k=1}^{K-1}\left(\beta_k+\sum_{t=k+1}^{K}C_{t-k,t}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2.\end{aligned} \tag{23}$$

Note that $\beta_K\to(\frac{\eta}{m}-\frac{\eta^2L_c}{2m})q_0$ as $K\to\infty$. If $q_0>0$ or $\|\nabla f(\mathbf{x})\|$ is bounded, by letting $K\to\infty$ in (23) and using the lower boundedness of f, we have from Lemma 1 that

$$\textstyle\sum_{k=1}^{\infty}\left(\frac{\eta}{2m}-\frac{\eta^2L_c}{2m}-\frac{\eta^2L_rT}{m\sqrt{m}}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2<\infty.$$

Since $\eta<\frac{1/L_c}{1+2\kappa T/\sqrt{m}}$, we have (22) from the above inequality.

From the Markov inequality, it follows that $\|\nabla f(\mathbf{x}^k)\|$ converges to *zero* with probability one. Let $\bar{\mathbf{x}}$ be a limit point of $\{\mathbf{x}^k\}_{k\geq 1}$, i.e., there is a subsequence $\{\mathbf{x}^k\}_{k\in\mathcal{K}}$ convergent to $\bar{\mathbf{x}}$. Hence, $\|\nabla f(\mathbf{x}^k)\|\to 0$ almost surely as $\mathcal{K}\ni k\to\infty$. By [11, Theorem 3.4, p.212], there is a subsubsequence $\{\mathbf{x}^k\}_{k\in\mathcal{K}'}$ such that $\|\nabla f(\mathbf{x}^k)\|\to 0$ almost surely as $\mathcal{K}'\ni k\to\infty$. This completes the proof. □

3.2 Convergence rate for the convex case

In this subsection, we assume the convexity of f and establish convergence rate results of Algorithm 1 for solving (9). Besides Assumptions 1 through 3, we make an additional assumption to the delay as follows. It means the delay follows a sub-exponential distribution.

Assumption 4 *There is a constant* $\sigma>1$ *such that*

$$M_\sigma:=\mathbb{E}[\sigma^{\mathbf{j}}]<\infty. \tag{24}$$

The condition in (24) is stronger than $T<\infty$, and both of them hold if the delay j_k is uniformly bounded by some number τ or follows the Poisson distribution; see the discussions in Section 5. Using this additional assumption and choosing an appropriate stepsize, we are able to control the gradient of f such that it changes not too fast.

Lemma 3 *Under Assumptions 2 through 4, for any* $1 < \rho \leq \sigma$*, if the stepsize satisfies*

$$0 < \eta \leq \tfrac{(\rho-1)\sqrt{m}}{\rho L_r(1+M_\rho)}, \tag{25}$$

with M_ρ *defined in* (24)*, then for all* k*, it holds that*

$$\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2 \quad \textit{and} \quad \mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2 \leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2. \tag{26}$$

The proof of Lemma 3 follows an argument similar to [16]. Since it is rather long, it is included in the appendix. Similar to Lemma 2, we can show the following result.

Lemma 4 *For any* k*, it holds that*

$$\begin{aligned}
&\textstyle\sum_{t=0}^{k-1} q_t \mathbb{E}[-\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t}) - \nabla f(\mathbf{x}^k)\rangle] \\
&- c_k \mathbb{E}\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0) - \nabla f(\mathbf{x}^k)\rangle \\
\leq &\tfrac{\eta L_r}{2\sqrt{m}} \sum_{d=1}^{k} c_{k-d}c_d \|\nabla f(\mathbf{x}^0)\|^2 + \tfrac{\eta L_r}{2\sqrt{m}} \sum_{t=1}^{k-1}\sum_{d=1}^{t} c_d q_{t-d} \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 \\
&+ \tfrac{\eta L_r}{2\sqrt{m}} \Big(\textstyle\sum_{t=0}^{k-1} t q_t + k c_k\Big) \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2.
\end{aligned} \tag{27}$$

Proof Following an argument similar to how (16) is obtained, we can show

$$\begin{aligned}
&\sum_{t=0}^{k-1} q_t \mathbb{E}[-\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t}) - \nabla f(\mathbf{x}^k)\rangle] \\
\leq &\tfrac{\eta L_r}{2\sqrt{m}} \sum_{t=0}^{k-1} q_t \Big(\sum_{d=k-t}^{k-1} \big(\sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d \|\nabla f(\mathbf{x}^0)\|^2\big) + t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \Big), \\
&- c_k \mathbb{E}\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0) - \nabla f(\mathbf{x}^k)\rangle \\
\leq &\tfrac{\eta L_r}{2\sqrt{m}} c_k \Big(\sum_{d=0}^{k-1} \Big(\sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d \|\nabla f(\mathbf{x}^0)\|^2 \Big) + k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \Big).
\end{aligned}$$

Using the above inequalities, we complete the proof by noting (17),

$$\begin{aligned}
\textstyle\sum_{t=0}^{k-1} q_t \sum_{d=k-t}^{k-1} c_d + c_k \sum_{d=0}^{k-1} c_d &= \textstyle\sum_{d=1}^{k-1} (c_{k-d} - c_k) c_d + c_k \sum_{d=0}^{k-1} c_d \\
&= \textstyle\sum_{d=1}^{k-1} c_{k-d} c_d + c_k = \sum_{d=1}^{k} c_{k-d} c_d,
\end{aligned} \tag{28}$$

and $c_k \sum_{d=0}^{k-1}\sum_{r=0}^{d-1} q_r \|\nabla f(\mathbf{x}^{d-r})\|^2 = \sum_{t=1}^{k-1}\sum_{d=1}^{t} c_k q_{t-d} \|\nabla f(\mathbf{x}^{k-t})\|^2$. □

Using the above two lemmas, we establish sufficient objective decrease.

Theorem 3 (Sufficient progress) *Under Assumptions 1 through 4, we let* $\{\mathbf{x}^k\}_{k\geq 1}$ *be the sequence generated from Algorithm 1. For a certain* $1 < \rho < \sigma$*, define*

$$N_\rho := \mathbb{E}[\mathbf{j}\rho^{\mathbf{j}}]. \tag{29}$$

Take the stepsize such that (25) *is satisfied and also*

$$0 < \eta < 2\Big(L_c\big(M_\rho + \tfrac{\kappa(2N_\rho M_\rho + T)}{\sqrt{m}}\big)\Big)^{-1}. \tag{30}$$

Let

$$D = \tfrac{\eta}{2m}\left(2 - \tfrac{\eta L_r}{\sqrt{m}}(2N_\rho M_\rho + T) - \eta L_c M_\rho\right). \tag{31}$$

Then,

$$\mathbb{E}f(\mathbf{x}^{k+1}) \le \mathbb{E}f(\mathbf{x}^k) - D\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2. \tag{32}$$

Proof First note that for any $\rho < \sigma$, $t\rho^t$ is dominated by σ^t as t is sufficiently large. Hence, $N_\rho < \infty$ from (24), and it is easy to see $T < \infty$. Also note that

$$\mathbb{E}[\mathbf{j}\rho^{\mathbf{j}}] = \sum_{t=1}^{\infty} tq_t\rho^t = \sum_{t=1}^{\infty}\sum_{d=1}^{t} q_t\rho^t = \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} q_t\rho^t \ge \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} q_t\rho^d = \sum_{d=1}^{\infty} c_d\rho^d. \tag{33}$$

We write the cross terms in (19) to

$$\langle\nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t})\rangle = \langle\nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t}) - \nabla f(\mathbf{x}^k)\rangle + \|\nabla f(\mathbf{x}^k)\|^2.$$

Taking expectation on both sides of (19) and using (27), we have

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{k+1}) \le& \mathbb{E}f(\mathbf{x}^k) + \tfrac{\eta^2 L_r}{2m\sqrt{m}}\sum_{d=1}^{k} c_{k-d}c_d\|\nabla f(\mathbf{x}^0)\|^2 \\
&+ \tfrac{\eta^2 L_r}{2m\sqrt{m}}\sum_{t=1}^{k-1}\sum_{d=1}^{t} c_d q_{t-d}\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 \\
&+ \tfrac{\eta^2 L_r}{2m\sqrt{m}}\left(\sum_{t=0}^{k-1} tq_t + kc_k\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \tfrac{\eta}{m}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
&+ \tfrac{\eta^2 L_c}{2m}\sum_{t=0}^{k-1} q_t\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 + \tfrac{\eta^2 L_c}{2m}c_k\|\nabla f(\mathbf{x}^0)\|^2.
\end{aligned} \tag{34}$$

The above inequality together with (26) implies

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{k+1}) \le& \mathbb{E}f(\mathbf{x}^k) + \tfrac{\eta^2 L_r}{2m\sqrt{m}}\textstyle\sum_{d=1}^{k} c_{k-d}c_d\rho^k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
&+ \tfrac{\eta^2 L_r}{2m\sqrt{m}}\textstyle\sum_{t=1}^{k-1}\sum_{d=1}^{t} c_d q_{t-d}\rho^t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
&+ \tfrac{\eta^2 L_r}{2m\sqrt{m}}\left(\textstyle\sum_{t=0}^{k-1} tq_t + kc_k\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \tfrac{\eta}{m}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
&+ \tfrac{\eta^2 L_c}{2m}\textstyle\sum_{t=0}^{k-1} q_t\rho^t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \tfrac{\eta^2 L_c}{2m}c_k\rho^k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2.
\end{aligned} \tag{35}$$

Note that $\sum_{t=1}^{k-1}\sum_{d=1}^{t} c_d q_{t-d}\rho^t \le \sum_{t=1}^{\infty}\sum_{d=1}^{t} c_d q_{t-d}\rho^t$, which by exchanging summations equals $\sum_{d=1}^{\infty} c_d\rho^d \sum_{t=d}^{\infty} q_{t-d}\rho^{t-d} \overset{(33)}{\le} N_\rho M_\rho$. Also note that $\sum_{d=1}^{k} c_{k-d}c_d\rho^k = \sum_{d=1}^{k} c_d\rho^d c_{k-d}\rho^{k-d} \le \sum_{d=1}^{k} c_d\rho^d\left(\sum_{r=0}^{\infty} q_r\rho^r\right) \le N_\rho M_\rho$. From these relations and (35), we obtain

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{k+1}) &\le \mathbb{E}f(\mathbf{x}^k) + \tfrac{\eta^2 L_r}{m\sqrt{m}}N_\rho M_\rho\|\nabla f(\mathbf{x}^k)\|^2 \\
&\quad + \tfrac{\eta^2 L_r}{2m\sqrt{m}}\left(\textstyle\sum_{t=0}^{k-1} tq_t + kc_k\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \tfrac{\eta}{m}\|\nabla f(\mathbf{x}^k)\|^2 \\
&\quad + \tfrac{\eta^2 L_c}{2m}\textstyle\sum_{t=0}^{k-1} q_t\rho^t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \tfrac{\eta^2 L_c}{2m}c_k\rho^k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
&\le \mathbb{E}f(\mathbf{x}^k) + \left(\tfrac{\eta^2 L_r}{2m\sqrt{m}}(2N_\rho M_\rho + T) + \tfrac{\eta^2 L_c}{2m}M_\rho - \tfrac{\eta}{m}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2,
\end{aligned}$$

which completes the proof. □

Using (32) and the convexity of f, we establish the following convergence rate.

Theorem 4 (Convergence rate for the convex smooth case) *Under the assumptions of Theorem 3, we have*

1. *If f is convex and $\|\mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k)\| \leq B$, $\forall k$ for a certain constant B, then*

$$\mathbb{E}[f(\mathbf{x}^{k+1}) - f^*] \leq \frac{1}{(f(\mathbf{x}^0)-f^*)^{-1}+(k+1)DB^{-2}}, \tag{36}$$

   where f^* denotes the minimum value of (9) *and D is given in* (31).
2. *If f is strongly convex with constant μ, then*

$$\mathbb{E}[f(\mathbf{x}^{k+1}) - f^*] \leq (1 - 2\mu D)\mathbb{E}[f(\mathbf{x}^k) - f^*], \tag{37}$$

   where D is given in (31).

Remark 3 For the sublinear rate in (36), we assume the boundedness of the iterates. This assumption can be relaxed if we use potentially smaller stepsize; see Theorem 6.

For the linear convergence, the assumption on strongly convexity can be weakened to *either essential or restrict strong convexity*, see [15] and [16].

Proof If $\|\mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k)\| \leq B$, then from $f(\mathbf{x}^k) - f(\mathcal{P}_{X^*}(\mathbf{x}^k)) \leq \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k)\rangle$, we have

$$|f(\mathbf{x}^k) - f^*| \leq \|\nabla f(\mathbf{x}^k)\| \cdot \|\mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k)\| \leq B\|\nabla f(\mathbf{x}^k)\|,$$

and thus

$$\|\nabla f(\mathbf{x}^k)\|^2 \geq \frac{1}{B^2}(f(\mathbf{x}^k) - f^*)^2. \tag{38}$$

Substituting (38) into (32) yields

$$\mathbb{E}f(\mathbf{x}^{k+1}) \leq \mathbb{E}f(\mathbf{x}^k) - \frac{D}{B^2}\mathbb{E}(f(\mathbf{x}^k) - f^*)^2.$$

Hence,

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}^{k+1}) - f^*] \leq \mathbb{E}[f(\mathbf{x}^k) - f^*] - \frac{D}{B^2}\mathbb{E}(f(\mathbf{x}^k) - f^*)^2 \\
\Rightarrow\ &\frac{1}{\mathbb{E}[f(\mathbf{x}^{k+1})-f^*]} \geq \frac{1}{\mathbb{E}[f(\mathbf{x}^k)-f^*]} + \frac{D}{B^2}\frac{\mathbb{E}[f(\mathbf{x}^k)-f^*]}{\mathbb{E}[f(\mathbf{x}^{k+1})-f^*]} \geq \frac{1}{\mathbb{E}[f(\mathbf{x}^k)-f^*]} + \frac{D}{B^2} \\
\Rightarrow\ &\frac{1}{\mathbb{E}[f(\mathbf{x}^{k+1})-f^*]} \geq \frac{1}{[f(\mathbf{x}^0)-f^*]} + \frac{D(k+1)}{B^2},
\end{aligned}$$

and thus (36) holds.

If f is strongly convex with constant μ, then

$$-\frac{1}{2\mu}\|\nabla f(\mathbf{x}^k)\|^2 \leq f^* - f(\mathbf{x}^k).$$

We immediately have (37) from (32) and the above inequality. This completes the proof. □

4 Convergence results for the nonsmooth case

In this section, we analyze the convergence of Algorithm 1 for possibly nonsmooth cases. Throughout this section, we let

$$\bar{\mathbf{x}}^{k+1} = \mathbf{prox}_{\eta R}\left(\mathbf{x}^k - \eta \nabla f(\mathbf{x}^{k-j_k})\right)$$

a virtual full-update iterate, where R is defined in (4), and denote

$$\mathbf{d}^k = \bar{\mathbf{x}}^{k+1} - \mathbf{x}^k.$$

Due to more generality, we will make stronger assumptions on the delay than those made in the previous section. But all these assumptions are satisfied if the delay is uniformly bounded or follows the Poisson distribution, as shown in Section 5.

4.1 Convergence for the nonconvex case

We first establish the almost sure global convergence for possibly nonconvex cases starting with the following square summable result.

Lemma 5 (Square summability) *Under Assumptions 1 through 3, we let $\{\mathbf{x}^k\}_{k\geq 1}$ be the sequence generated in Algorithm 1. Assume $S < \infty$, and the stepsize is taken as $0 < \eta < \frac{1/L_c}{1+\kappa^2 S/(2m)}$. Then*

$$\sum_{k=0}^{\infty} \mathbb{E}\|\mathbf{d}^k\|^2 < \infty. \tag{39}$$

Proof By the definition of $\bar{\mathbf{x}}^{k+1}$, we have $-\nabla f(\mathbf{x}^{k-j_k}) - \frac{1}{\eta}\mathbf{d}^k \in \partial R(\bar{\mathbf{x}}^{k+1})$, which together with the convexity of R implies that, for any $\mathbf{x}$,

$$R(\bar{\mathbf{x}}^{k+1}) - R(\mathbf{x}) \leq -\langle \nabla f(\mathbf{x}^{k-j_k}) + \tfrac{1}{\eta}\mathbf{d}^k, \bar{\mathbf{x}}^{k+1} - \mathbf{x}\rangle. \tag{40}$$

By $\mathbf{x}^{k+1} = \mathbf{x}^k + U_{i_k}\mathbf{d}^k$ and (7), we get $F(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla_{i_k} f(\mathbf{x}^k), \mathbf{d}^k_{i_k}\rangle + \frac{L_c}{2}\|\mathbf{d}^k_{i_k}\|^2 + R(\mathbf{x}^{k+1})$. To this inequality, take conditional expectation on i_k:

$$\mathbb{E}_{i_k} F(\mathbf{x}^{k+1}) \leq F(\mathbf{x}^k) + \tfrac{1}{m}\left(\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k\rangle + \tfrac{L_c}{2}\|\mathbf{d}^k\|^2 + R(\bar{\mathbf{x}}^{k+1}) - R(\mathbf{x}^k)\right).$$

To bound the right-hand side, we split the cross term as

$$\langle \nabla f(\mathbf{x}^k), \mathbf{d}^k\rangle = \langle \nabla f(\mathbf{x}^{k-j_k}), \mathbf{d}^k\rangle + \langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-j_k}), \mathbf{d}^k\rangle$$

and apply (40) with $\mathbf{x} = \mathbf{x}^k$, arriving at

$$\mathbb{E}_{i_k} F(\mathbf{x}^{k+1}) \leq F(\mathbf{x}^k) + \tfrac{1}{m}(\tfrac{L_c}{2} - \tfrac{1}{\eta})\|\mathbf{d}^k\|^2 + \tfrac{1}{m}\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-j_k}), \mathbf{d}^k\rangle. \tag{41}$$

Following a similar argument in the proof of Lemma 2 and Young's inequality, we get

$$\begin{aligned}\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-j_k}), \mathbf{d}^k\rangle \leq & L_r \sum_{d=k-j_k}^{k-1} \|\mathbf{x}^{d+1} - \mathbf{x}^d\| \cdot \|\mathbf{d}^k\| \\ \leq & \tfrac{L_r}{2\kappa}\|\mathbf{d}^k\|^2 + \tfrac{\kappa L_r}{2}\Big(j_k \sum_{d=k-j_k}^{k-1} \|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2\Big). \quad (42)\end{aligned}$$

Note that

$$\begin{aligned}&\mathbb{E}\Big[j_k \sum_{d=k-j_k}^{k-1} \|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2\Big] \\ =& \sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2 + \sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2 \\ =& \tfrac{1}{m}\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{d}^d\|^2 + \tfrac{1}{m}\sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{d}^d\|^2. \quad (43)\end{aligned}$$

Hence, taking expectation yields

$$\begin{aligned}&\mathbb{E}\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-j_k}), \mathbf{d}^k\rangle \\ \leq & \tfrac{L_r}{2}\Big[\tfrac{1}{\kappa}\mathbb{E}\|\mathbf{d}^k\|^2 + \tfrac{\kappa}{m}\Big(\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{d}^d\|^2 + \sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{d}^d\|^2\Big)\Big]. \quad (44)\end{aligned}$$

Taking expectation on both sides of (41) and substituting (44) yield

$$\begin{aligned}&\mathbb{E}[F(\mathbf{x}^{k+1}) - F(\mathbf{x}^k)] + \tfrac{1}{m}\Big(\tfrac{1}{\eta} - L_c\Big)\mathbb{E}\|\mathbf{d}^k\|^2 \\ \leq & \tfrac{\kappa L_r}{2m^2}\Big(\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{d}^d\|^2 + \sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{d}^d\|^2\Big). \quad (45)\end{aligned}$$

From Lemma 12, we have that for any $K \geq 0$,

$$\begin{aligned}\sum_{k=0}^{K}\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{d}^d\|^2 &\overset{(83)}{=} \sum_{k=0}^{K}\sum_{d=1}^{k-1}\Big(\sum_{t=k-d}^{k-1} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2 \\ &\overset{(84)}{=} \sum_{d=1}^{K-1}\sum_{k=d+1}^{K}\Big(\sum_{t=k-d}^{k-1} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2 \\ [k \leftrightarrow d] \quad &= \sum_{k=1}^{K-1}\Big(\sum_{d=k+1}^{K}\sum_{t=d-k}^{d-1} q_t t\Big)\mathbb{E}\|\mathbf{d}^k\|^2, \quad (46)\end{aligned}$$

and

$$\begin{aligned}\sum_{k=0}^{K}\sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{d}^d\|^2 &= \sum_{k=1}^{K}\sum_{d=0}^{k-1}\Big(\sum_{t=k}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^k\|^2 \\ &\overset{(84)}{=} \sum_{d=0}^{K-1}\sum_{k=d+1}^{K}\Big(\sum_{t=k}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2 \\ [k \leftrightarrow d] \quad &= \sum_{k=0}^{K-1}\Big(\sum_{d=k+1}^{K}\sum_{t=d}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^k\|^2. \quad (47)\end{aligned}$$

Summing up (45) from $k = 0$ through K and substituting (46) and (47), we have

$$\begin{aligned}&\mathbb{E}[F(\mathbf{x}^{K+1}) - F(\mathbf{x}^0)] + \tfrac{1}{m}(\tfrac{1}{\eta} - L_c)\sum_{k=0}^{K}\mathbb{E}\|\mathbf{d}^k\|^2\\ \le&\tfrac{\kappa L_r}{2m^2}\sum_{k=0}^{K-1}\Big(\sum_{d=k+1}^{K}\sum_{t=d-k}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^k\|^2. \end{aligned}\tag{48}$$

Note that

$$\sum_{d=k+1}^{K}\sum_{t=d-k}^{\infty} q_t t = \sum_{d=1}^{K-k}\sum_{t=d}^{\infty} q_t t \le \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} q_t t = \sum_{t=1}^{\infty} t^2 q_t = S.$$

Since F is lower bounded, we have (39) from (48) by letting $K \to \infty$. □

Since $(\mathbb{E}[\mathbf{j}])^2 \le \mathbb{E}[\mathbf{j}^2]$, the condition $S < \infty$ implies $T < \infty$. Equation (39) indicates that $\mathbb{E}\|\mathbf{d}^k\| \to 0$ as $k \to \infty$. Together with $S < \infty$, we are able to show $\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|$ also approaches *zero*, as summarized in the following.

Lemma 6 *Under the assumptions of Lemma 5, we have*

$$\lim_{k\to\infty}\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\| = 0.$$

Proof Pick any $\epsilon > 0$. From (39), there must exist an integer $J > 0$ such that

$$\sum_{d=J}^{\infty}\mathbb{E}\|\mathbf{d}^d\|^2 \le m\epsilon\Big(3\sum_{t=1}^{\infty} q_t t\Big)^{-1}. \tag{49}$$

For the above J, there must exist an integer $K > J$ such that, for any $k \ge K$,

$$\sum_{t=k-J}^{\infty} q_t t \le m\epsilon\Big(3\sum_{d=0}^{\infty}\mathbb{E}\|\mathbf{d}^d\|^2\Big)^{-1}. \tag{50}$$

From Young's inequality, it follows that $\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2 \le j_k\sum_{d=k-j_k}^{k-1}\|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2$. Hence, for any $k \ge K$, using (43) and (83), we have

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2 \le&\tfrac{1}{m}\Big[\sum_{d=1}^{k-1}\Big(\sum_{t=k-d}^{k-1} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2 + \sum_{d=0}^{k-1}\Big(\sum_{t=k}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2\Big]\\ =&\tfrac{1}{m}\sum_{d=1}^{J}\Big(\sum_{t=k-d}^{k-1} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2\\ &+\tfrac{1}{m}\Big[\sum_{d=J+1}^{k-1}\Big(\sum_{t=k-d}^{k-1} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2 + \sum_{d=0}^{k-1}\Big(\sum_{t=k}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2\Big]\\ \le&\tfrac{1}{m}\sum_{d=1}^{J}\Big(\sum_{t=k-J}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2\\ &+\tfrac{1}{m}\Big[\sum_{d=J+1}^{k-1}\Big(\sum_{t=1}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2 + \sum_{d=0}^{k-1}\Big(\sum_{t=k-J}^{\infty} q_t t\Big)\mathbb{E}\|\mathbf{d}^d\|^2\Big],\end{aligned}$$

which implies $\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2 \le \epsilon$ under (49) and (50). We have $\lim_{k\to\infty}\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2 = 0$ as ϵ is arbitrary. Now note $\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\| \le \sqrt{\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2}$ to complete the proof. □

Using Lemmas 5 and 6, we establish the almost sure global convergence of Algorithm 1.

Theorem 5 *Under the assumptions of Lemma 5, any limit point* $\mathbf{x}^*$ *of* $\{\mathbf{x}^k\}$ *is a critical point of* (1) *almost surely.*

Before proving this theorem, we make two remarks as follows.

Remark 4 From the theorem, we see that if $S = \mathbb{E}[\mathbf{j}^2] = o(m)$, then the stepsize required for convergence only weakly depends on the delay.

Remark 5 (Comparison of stepsize) The works [8] consider asynchronous coordinate descent for nonconvex problems. To have convergence to critical points, they assume delays bounded by a number τ. Also, they require the boundedness of iterates and the stepsize less than $\frac{1/L_c}{1+2\kappa\tau/\sqrt{m}}$. Note that our stepsize in Theorem 5 is larger if $\kappa^2 S \leq 16m$, where $S = \mathbb{E}[\mathbf{j}^2] < \tau^2$, and that can lead to faster convergence.

Proof Let $\{\mathbf{x}^k\}_{k\in\mathcal{K}}$ be a subsequence that converges to $\mathbf{x}^*$. Since $\mathbb{E}\|\mathbf{d}^k\| \to 0$ as $\mathcal{K} \ni k \to \infty$, from the Markov inequality, $\|\mathbf{d}^k\|$ converges to *zero* in probability as $\mathcal{K} \ni k \to \infty$. By [11, Theorem 3.4, pp.212], there is a subsubsequence $\{\mathbf{x}^k\}_{k\in\mathcal{K}'}$ such that $\|\mathbf{d}^k\|$ almost surely converges to *zero* as $\mathcal{K}' \ni k \to \infty$. Hence, $\bar{\mathbf{x}}^{k+1}$ almost surely converges to $\mathbf{x}^*$ as $\mathcal{K}' \ni k \to \infty$.

Since $-\nabla f(\mathbf{x}^{k-j_k}) - \frac{1}{\eta}\mathbf{d}^k \in \partial R(\bar{\mathbf{x}}^{k+1})$, we have

$$\mathrm{dist}(\mathbf{0}, \partial F(\bar{\mathbf{x}}^{k+1})) \leq \|\nabla f(\bar{\mathbf{x}}^{k+1}) - \nabla f(\mathbf{x}^{k-j_k}) - \tfrac{1}{\eta}\mathbf{d}^k\|.$$

Using triangle inequality and the Lipschitz continuity of ∇f, and taking expectation give

$$\mathbb{E}\mathrm{dist}(\mathbf{0}, \partial F(\bar{\mathbf{x}}^{k+1})) \leq L_f\mathbb{E}\|\mathbf{d}^k\| + L_f\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\| + \tfrac{1}{\eta}\mathbb{E}\|\mathbf{d}^k\|.$$

From Lemmas 5 and 6, it follows that the right-hand side approaches to *zero* as $k \to \infty$. Hence, $\mathbb{E}\mathrm{dist}(\mathbf{0}, \partial F(\bar{\mathbf{x}}^{k+1})) \to 0$ as $k \to \infty$. If necessary, passing to another subsequence, we use Markov inequality and [11, Theorem 3.4, pp.212] again to have $\mathrm{dist}(\mathbf{0}, \partial F(\bar{\mathbf{x}}^{k+1}))$ almost surely converges to *zero* as $\mathcal{K}' \ni k \to \infty$. Now use the outer semicontinuity [27] of $\mathrm{dist}(\mathbf{0}, \partial F(\mathbf{x}))$ to obtain the desired result. □

4.2 Convergence rate for the convex case

In this subsection, we establish convergence rates of Algorithm 1 for nonsmooth convex cases. Similar to (26), we first show that choosing an appropriate stepsize, the iterate difference does not change too fast.

Lemma 7 (Fundamental bounds) *Assume Assumptions 2 through 4. Then for any $1 < \rho < \sigma$, it holds that*

$$\gamma_{\rho,1} := \sum_{t=1}^{\infty} q_t \tfrac{\rho^{t/2}-1}{\rho^{1/2}-1} < \infty \quad \text{and} \quad \gamma_{\rho,2} := \Big(\sum_{t=1}^{\infty} q_t t \tfrac{\rho^{t}-1}{1-\rho^{-1}}\Big)^{1/2} < \infty. \tag{51}$$

In addition, if the stepsize is taken such that

$$0 < \eta \leq \tfrac{(1-\rho^{-1})\sqrt{m}-4}{2L_r(1+\gamma_{\rho,1}+\gamma_{\rho,2})}, \tag{52}$$

then, for all $k \geq 1$,

$$\mathbb{E}\|\mathbf{d}^{k-1}\|^2 \leq \rho \mathbb{E}\|\mathbf{d}^{k}\|^2. \tag{53}$$

Proof It is easy to show (51) by noting that $t\rho^t$ is dominated by σ^t as t is sufficiently large. Next we show (53) by induction.

Using the inequality $\|\mathbf{u}\|^2 - \|\mathbf{v}\|^2 \leq 2\|\mathbf{u}\|\cdot\|\mathbf{v}-\mathbf{u}\|$, we have

$$\|\mathbf{d}^{k-1}\|^2 - \|\mathbf{d}^{k}\|^2 \leq 2\|\mathbf{d}^{k-1}\|\cdot\|\mathbf{d}^{k}-\mathbf{d}^{k-1}\|, \ \forall k. \tag{54}$$

In addition, for all k,

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^{k-1}-\mathbf{x}^{k}\|\|\mathbf{d}^{k-1}\| &\leq \tfrac{1}{2}\mathbb{E}\left[\sqrt{m}\|\mathbf{x}^{k-1}-\mathbf{x}^{k}\|^2 + \tfrac{1}{\sqrt{m}}\|\mathbf{d}^{k-1}\|^2\right] \\ &= \tfrac{1}{\sqrt{m}}\mathbb{E}\|\mathbf{d}^{k-1}\|^2. \end{aligned} \tag{55}$$

Furthermore, from $\mathbf{d}^k - \mathbf{d}^{k-1} = \mathbf{x}^k - \mathbf{prox}_{\eta R}(\mathbf{x}^k - \eta\nabla f(\mathbf{x}^{k-j_k})) - \mathbf{x}^{k-1} + \mathbf{prox}_{\eta R}(\mathbf{x}^{k-1} - \eta\nabla f(\mathbf{x}^{k-1-j_{k-1}}))$, the nonexpansiveness of $\mathbf{prox}_{\eta R}$, and the triangle inequality, we have

$$\begin{aligned}&\|\mathbf{d}^k - \mathbf{d}^{k-1}\| \\ \leq&\|\mathbf{x}^k - \mathbf{x}^{k-1}\| + \|\mathbf{x}^k - \eta\nabla f(\mathbf{x}^{k-j_k}) - \mathbf{x}^{k-1} + \eta\nabla f(\mathbf{x}^{k-1-j_{k-1}})\| \\ \leq&2\|\mathbf{x}^k - \mathbf{x}^{k-1}\| + \eta\|\nabla f(\mathbf{x}^{k-j_k}) - \nabla f(\mathbf{x}^{k-1-j_{k-1}})\| \qquad (56)\\ \leq&2\|\mathbf{x}^k - \mathbf{x}^{k-1}\| + \eta\|\nabla f(\mathbf{x}^{k-j_k}) - \nabla f(\mathbf{x}^{k})\| \\ &+ \eta\|\nabla f(\mathbf{x}^{k}) - \nabla f(\mathbf{x}^{k-1-j_{k-1}})\|. \qquad (57)\end{aligned}$$

When $k = 1$, we have $j_0 = 0$ and $j_1 \in \{0, 1\}$ because $j_k \leq k, \forall k$. Hence, from (56),

$$\|\mathbf{d}^1 - \mathbf{d}^0\| \leq 2\|\mathbf{x}^1 - \mathbf{x}^0\| + \eta\|\nabla f(\mathbf{x}^1) - \nabla f(\mathbf{x}^0)\| \leq (2 + \eta L_r)\|\mathbf{x}^1 - \mathbf{x}^0\|,$$

which together with (54) and (55) implies

$$\mathbb{E}[\|\mathbf{d}^0\|^2 - \|\mathbf{d}^1\|^2] \leq (4 + 2\eta L_r)\mathbb{E}[\|\mathbf{d}^0\|\cdot\|\mathbf{x}^0 - \mathbf{x}^1\|] \leq \tfrac{4+2\eta L_r}{\sqrt{m}}\mathbb{E}\|\mathbf{d}^0\|^2.$$

Hence,

$$\mathbb{E}\|\mathbf{d}^0\|^2 \leq (1 - \tfrac{4+2\eta L_r}{\sqrt{m}})^{-1}\mathbb{E}\|\mathbf{d}^1\|^2 \overset{(52)}{\leq} \rho\mathbb{E}\|\mathbf{d}^1\|^2.$$

Assume (53) holds for all $k \leq K-1$. We show it holds for $k = K$. First, for any $d \leq K-1$,

$$\begin{aligned}\mathbb{E}\|\mathbf{d}^{K-1}\|\cdot\|\mathbf{x}^d - \mathbf{x}^{d+1}\| &\leq \tfrac{1}{2}\mathbb{E}\Big[\tfrac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\|\mathbf{d}^{K-1}\|^2 + \tfrac{\sqrt{m}}{\rho^{\frac{K-1-d}{2}}}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|^2\Big] \\ &= \tfrac{1}{2}\mathbb{E}\Big[\tfrac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\|\mathbf{d}^{K-1}\|^2 + \tfrac{1}{\sqrt{m}\rho^{\frac{K-1-d}{2}}}\|\mathbf{d}^{d}\|^2\Big] \\ &\leq \tfrac{1}{2}\mathbb{E}\Big[\tfrac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\|\mathbf{d}^{K-1}\|^2 + \tfrac{\rho^{K-1-d}}{\sqrt{m}\rho^{\frac{K-1-d}{2}}}\|\mathbf{d}^{K-1}\|^2\Big] \\ &= \tfrac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\mathbb{E}\|\mathbf{d}^{K-1}\|^2. \end{aligned} \tag{58}$$

Secondly, we have

$$\begin{aligned}&\mathbb{E}[\|\mathbf{d}^{K-1}\|^2 - \|\mathbf{d}^{K}\|^2] \overset{(54)}{\leq} 2\mathbb{E}\|\mathbf{d}^{K-1}\|\|\mathbf{d}^K - \mathbf{d}^{K-1}\| \\ &\overset{(57)}{\leq} 4\mathbb{E}\|\mathbf{d}^{K-1}\|\|\mathbf{x}^K - \mathbf{x}^{K-1}\| + 2\eta\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^K) - \nabla f(\mathbf{x}^{K-1})\| \\ &\quad + 2\eta\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^{K-j_K}) - \nabla f(\mathbf{x}^K)\| \\ &\quad + 2\eta\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^{K-1}) - \nabla f(\mathbf{x}^{K-1-j_{K-1}})\| \\ &\overset{(55)}{\leq} \tfrac{4+2\eta L_r}{\sqrt{m}}\mathbb{E}\|\mathbf{d}^{K-1}\|^2 + 2\eta\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^{K-j_K}) - \nabla f(\mathbf{x}^K)\| \\ &\quad + 2\eta\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^{K-1}) - \nabla f(\mathbf{x}^{K-1-j_{K-1}})\|. \end{aligned} \tag{59}$$

Note that

$$\begin{aligned}\mathbb{E}_{j_K}\|\nabla f(\mathbf{x}^{K-j_K}) - \nabla f(\mathbf{x}^K)\| &= \sum_{t=1}^{K-1} q_t\|\nabla f(\mathbf{x}^{K-t}) - \nabla f(\mathbf{x}^K)\| \\ &\quad + c_K\|\nabla f(\mathbf{x}^0) - \nabla f(\mathbf{x}^K)\|. \end{aligned}$$

By the triangle inequality and the Lipschitz of ∇f, it follows that, for any $1 \leq t \leq K$,

$$\begin{aligned}\|\nabla f(\mathbf{x}^{K-t}) - \nabla f(\mathbf{x}^K)\| &\leq \textstyle\sum_{d=K-t}^{K-1}\|\nabla f(\mathbf{x}^d) - \nabla f(\mathbf{x}^{d+1})\| \\ &\leq L_r\textstyle\sum_{d=K-t}^{K-1}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|. \end{aligned} \tag{60}$$

Since $\|\mathbf{d}^{K-1}\|$ is independent of j_K, we have from the above two equations that

$$\begin{aligned}&\mathbb{E}\|\mathbf{d}^{K-1}\|\cdot\|\nabla f(\mathbf{x}^{K-j_K}) - \nabla f(\mathbf{x}^K)\| \\ &\leq L_r\sum_{t=1}^{K-1} q_t\mathbb{E}\|\mathbf{d}^{K-1}\|\sum_{d=K-t}^{K-1}\|\mathbf{x}^d - \mathbf{x}^{d+1}\| + L_r c_K\mathbb{E}\|\mathbf{d}^{K-1}\|\sum_{d=0}^{K-1}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|. \end{aligned}$$

Using (58), the definition of $\gamma_{\rho,1}$ in (51) and $\sum_{d=K-t}^{K-1}\rho^{\frac{K-1-d}{2}} = \frac{\rho^{t/2}-1}{\rho^{1/2}-1}$, $\forall 1 \leq t \leq K$, we have

$$\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^{K-j_K}) - \nabla f(\mathbf{x}^K)\| \leq \tfrac{L_r}{\sqrt{m}}\gamma_{\rho,1}\mathbb{E}\|\mathbf{d}^{K-1}\|^2. \tag{61}$$

Also, using Young's inequality and (60) with K replaced by $K-1$ and $t = j_{K-1}$, we have, for any $\beta > 0$,

$$\begin{aligned}&\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^{K-1}) - \nabla f(\mathbf{x}^{K-1-j_{K-1}})\| \\ &\leq \tfrac{L_r}{2\beta}\mathbb{E}\|\mathbf{d}^{K-1}\|^2 + \tfrac{L_r\beta}{2}\mathbb{E}\Big[\sum_{d=K-1-j_{K-1}}^{K-2}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|\Big]^2. \end{aligned} \tag{62}$$

Note that

$$\begin{aligned}&\mathbb{E}\Big[\sum_{d=K-1-j_{K-1}}^{K-2}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|\Big]^2 \\ &= \sum_{t=1}^{K-2} q_t \mathbb{E}\Big[\sum_{d=K-1-t}^{K-2}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|\Big]^2 + c_{K-1}\mathbb{E}\Big[\sum_{d=0}^{K-2}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|\Big]^2 \\ &\leq \sum_{t=1}^{K-2} q_t t \sum_{d=K-1-t}^{K-2}\mathbb{E}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|^2 + c_{K-1}(K-1)\sum_{d=0}^{K-2}\mathbb{E}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|^2.\end{aligned}$$

Substituting this inequality into (62), noting $\mathbb{E}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|^2 = \frac{1}{m}\mathbb{E}\|\mathbf{d}^d\|^2$, and applying (53) for all $k \leq K-1$, we have

$$\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^{K-1}) - \nabla f(\mathbf{x}^{K-1-j_{K-1}})\| \leq C\mathbb{E}\|\mathbf{d}^{K-1}\|^2$$

where $C = \frac{L_r}{2\beta} + \frac{L_r\beta}{2m}\sum_{t=1}^{K-2} q_t t \sum_{d=K-1-t}^{K-2}\rho^{K-1-d} + \frac{L_r\beta}{2m}c_{K-1}(K-1)\sum_{d=0}^{K-2}\rho^{K-1-d}$. Now let $\beta = \sqrt{m}(\sum_{t=1}^{K-2} q_t t\frac{\rho^t-1}{1-\rho^{-1}} + c_{K-1}(K-1)\frac{\rho^{K-1}-1}{1-\rho^{-1}})^{-1/2}$ and recall the definition of $\gamma_{\rho,2}$ in (51). From the above inequality, we have

$$\mathbb{E}\|\mathbf{d}^{K-1}\|\|\nabla f(\mathbf{x}^{K-1}) - \nabla f(\mathbf{x}^{K-1-j_{K-1}})\| \leq \tfrac{L_r\gamma_{\rho,2}}{\sqrt{m}}\mathbb{E}\|\mathbf{d}^{K-1}\|^2. \tag{63}$$

Substituting (61) and (63) into (59) gives

$$\mathbb{E}[\|\mathbf{d}^{K-1}\|^2 - \|\mathbf{d}^K\|^2] \leq \tfrac{4+2\eta L_r(1+\gamma_{\rho,1}+\gamma_{\rho,2})}{\sqrt{m}}\mathbb{E}\|\mathbf{d}^{K-1}\|^2,$$

and thus

$$\mathbb{E}\|\mathbf{d}^{K-1}\|^2 \leq \Big(1 - \tfrac{4+2\eta L_r(1+\gamma_{\rho,1}+\gamma_{\rho,2})}{\sqrt{m}}\Big)^{-1}\mathbb{E}\|\mathbf{d}^K\|^2 \overset{(52)}{\leq} \rho\mathbb{E}\|\mathbf{d}^K\|^2.$$

Therefore, by induction, it follows that (53) holds for all k, and we complete the proof. □

By this lemma, we are able to establish the convergence rate result of Algorithm 1 for solving (1) if the problem is convex.

Theorem 6 (Convergence rate for the nonsmooth convex case) *Under Assumptions 1 through 4, let $\{\mathbf{x}^k\}_{k\geq 1}$ be the sequence generated from Algorithm 1 with stepsize satisfying* (52) *and also*

$$\eta \leq \Big(L_c + \tfrac{2L_f\gamma_{\rho,2}^2}{m} + \tfrac{2L_r\gamma_{\rho,2}}{\sqrt{m}}\Big)^{-1}, \tag{64}$$

where $\gamma_{\rho,1}$ and $\gamma_{\rho,2}$ are defined in (51)*. We have*

1. If the function F is convex, then

$$\mathbb{E}[F(\mathbf{x}^k) - F^*] \leq \frac{m\Phi(\mathbf{x}^0)}{2\eta(m+k)}, \tag{65}$$

where

$$\Phi(\mathbf{x}^k) = \mathbb{E}\left\|\mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k)\right\|^2 + 2\eta\mathbb{E}[F(\mathbf{x}^k) - F^*].$$

2. If F is strongly convex with constant μ, then

$$\Phi(\mathbf{x}^k) \leq \left(1 - \frac{\eta\mu}{m(1+\eta\mu)}\right)^k \Phi(\mathbf{x}^0). \tag{66}$$

Before proving this theorem, we make two remarks and present a few lemmas below.

Remark 6 Similar to (37), for the linear convergence result (66), the strong convexity assumption can be weakened to *optimal strong convexity*. The latter one is strictly weaker than the former one; see *[17]* for more discussions.

Remark 7 (Comparison of stepsize) For the special case that the delay is bounded by $\tau = o(\sqrt[4]{m})$, choosing $\rho = O(1 + \frac{1}{\tau})$, we have both $\gamma_{\rho,1}$ and $\gamma_{\rho,2}$ are $O(\tau)$. Thus we can take stepsize almost $\frac{1}{L_c}$, which is larger than the stepsize $\frac{1}{2L_c}$ given in [17].

Lemma 8 *Let $\gamma_{\rho,2}$ be defined in* (51)*. We have*

$$\mathbb{E}\langle\nabla_{i_k} f(\mathbf{x}^{k-j_k}) - \nabla_{i_k} f(\mathbf{x}^k), \mathbf{x}_{i_k}^k - \mathbf{x}_{i_k}^{k+1}\rangle \leq \frac{L_r\gamma_{\rho,2}}{m\sqrt{m}}\mathbb{E}\|\mathbf{d}^k\|^2. \tag{67}$$

Proof It is proved via the Cauchy-Schwarz inequality, the bound (63), and $\mathbb{E}\langle\nabla_{i_k} f(\mathbf{x}^{k-j_k}) - \nabla_{i_k} f(\mathbf{x}^k), \mathbf{x}_{i_k}^k - \mathbf{x}_{i_k}^{k+1}\rangle = \frac{1}{m}\mathbb{E}\langle\nabla f(\mathbf{x}^{k-j_k}) - \nabla f(\mathbf{x}^k), \mathbf{d}^k\rangle$. □

Lemma 9 *It holds that*

$$\begin{aligned}&\mathbb{E}\left[f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) + r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}) - r_{i_k}(\mathbf{x}_{i_k}^{k+1})\right]\\ =&\mathbb{E}[F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})] + \frac{1}{m}\mathbb{E}[R(\mathcal{P}_{X^*}(\mathbf{x}^k)) - R(\mathbf{x}^k)].\end{aligned} \tag{68}$$

Proof Equation (68) is a direct consequence of $r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}) - r_{i_k}(\mathbf{x}_{i_k}^{k+1}) = r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}) - r_{i_k}(\mathbf{x}_{i_k}^k) + R(\mathbf{x}^k) - R(\mathbf{x}^{k+1})$. □

Lemma 10 *Let $\gamma_{\rho,2}$ be defined in* (51)*. It holds that*

$$\begin{aligned}\mathbb{E}\langle\nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}_{i_k}^k\rangle \leq&\frac{1}{m}\mathbb{E}[f(\mathcal{P}_{X^*}(\mathbf{x}^k)) - f(\mathbf{x}^k)]\\ &+ \frac{L_f\gamma_{\rho,2}^2}{m^2}\mathbb{E}\|\mathbf{d}^k\|^2.\end{aligned} \tag{69}$$

Proof Since i_k is uniformly distributed and independent of j_k, we have

$$\mathbb{E}_{i_k}\langle\nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}_{i_k}^k\rangle = \frac{1}{m}\langle\nabla f(\mathbf{x}^{k-j_k}), \mathcal{P}_{X^*}(\mathbf{x}^k) - \mathbf{x}^k\rangle. \tag{70}$$

We split the term and apply the convexity of f and Lipschitz continuity of ∇f to get

$$\begin{aligned}
&\langle \nabla f(\mathbf{x}^{k-j_k}), \mathcal{P}_{X^*}(\mathbf{x}^k) - \mathbf{x}^k \rangle \\
=&\langle \nabla f(\mathbf{x}^{k-j_k}), \mathcal{P}_{X^*}(\mathbf{x}^k) - \mathbf{x}^{k-j_k} \rangle \\
&+ \langle \nabla f(\mathbf{x}^k) + \nabla f(\mathbf{x}^{k-j_k}) - \nabla f(\mathbf{x}^k), \mathbf{x}^{k-j_k} - \mathbf{x}^k \rangle \\
\leq&[f(\mathcal{P}_{X^*}(\mathbf{x}^k)) - f(\mathbf{x}^{k-j_k}) + f(\mathbf{x}^{k-j_k}) - f(\mathbf{x}^k)] \\
&+ \langle \nabla f(\mathbf{x}^{k-j_k}) - \nabla f(\mathbf{x}^k), \mathbf{x}^{k-j_k} - \mathbf{x}^k \rangle \\
\leq&[f(\mathcal{P}_{X^*}(\mathbf{x}^k)) - f(\mathbf{x}^k)] + L_f \|\mathbf{x}^{k-j_k} - \mathbf{x}^k\|^2. \qquad (71)
\end{aligned}$$

Substituting (71) into (70) and taking expectation yield

$$\begin{aligned}
&\mathbb{E}\langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}^k_{i_k} \rangle \\
\leq&\tfrac{1}{m}\mathbb{E}[f(\mathcal{P}_{X^*}(\mathbf{x}^k)) - f(\mathbf{x}^k)] + \tfrac{L_f}{m}\mathbb{E}\|\mathbf{x}^{k-j_k} - \mathbf{x}^k\|^2.
\end{aligned}$$

Noting $\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2 \leq j_k \sum_{d=k-j_k}^{k-1} \|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2$, applying (43) and (53) and using the definition of $\gamma_{\rho,2}$, we complete the proof of (69). □

Lemma 11 *Under the assumptions of Theorem 6, we have* $\mathbb{E}[F(\mathbf{x}^{k+1})] \leq \mathbb{E}[F(\mathbf{x}^k)]$, $\forall k$.

Proof Taking expectation on both side of (41) and using (67) yield

$$\mathbb{E}[F(\mathbf{x}^{k+1})] \leq \mathbb{E}[F(\mathbf{x}^k)] + \tfrac{1}{m}\left(\tfrac{L_c}{2} - \tfrac{1}{\eta} + \tfrac{L_r \gamma_{\rho,2}}{\sqrt{m}}\right)\mathbb{E}\|\mathbf{d}^k\|^2,$$

which implies $\mathbb{E}[F(\mathbf{x}^{k+1})] \leq \mathbb{E}[F(\mathbf{x}^k)]$ from the condition on η in (64). □

Now we are ready to prove Theorem 6.

Proof (of Theorem 6) From the update of $\mathbf{x}^{k+1}$, we have

$$\mathbf{0} \in \nabla_{i_k} f(\mathbf{x}^{k-j_k}) + \tfrac{1}{\eta}(\mathbf{x}^{k+1}_{i_k} - \mathbf{x}^k_{i_k}) + \partial r_{i_k}(\mathbf{x}^{k+1}_{i_k}),$$

and thus for any $\mathbf{x}_{i_k}$, it holds from the convexity of r_{i_k} that

$$r_{i_k}(\mathbf{x}_{i_k}) \geq r_{i_k}(\mathbf{x}^{k+1}_{i_k}) - \langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}) + \tfrac{1}{\eta}(\mathbf{x}^{k+1}_{i_k} - \mathbf{x}^k_{i_k}), \mathbf{x}_{i_k} - \mathbf{x}^{k+1}_{i_k} \rangle. \qquad (72)$$

Since $\mathbf{x}^{k+1} = \mathbf{x}^k + U_{i_k}(\mathbf{x}^{k+1} - \mathbf{x}^k)$, we have

$$\begin{aligned}
\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^k)\|^2 =&\|\mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k)\|^2 - \|\mathbf{x}^{k+1}_{i_k} - \mathbf{x}^k_{i_k}\|^2 \\
&+ 2\langle \mathbf{x}^{k+1}_{i_k} - (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}, \mathbf{x}^{k+1}_{i_k} - \mathbf{x}^k_{i_k} \rangle. \qquad (73)
\end{aligned}$$

From the definition of $\mathcal{P}_{X^*}$, it follows that $\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^{k+1})\|^2 \leq \|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^k)\|^2$. Then using (72) and (73), we have

$$\begin{aligned}
\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^{k+1})\|^2 \leq&\|\mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k)\|^2 - \|\mathbf{x}^{k+1}_{i_k} - \mathbf{x}^k_{i_k}\|^2 \\
&+ 2\eta\left(r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}) - r_{i_k}(\mathbf{x}^{k+1}_{i_k})\right) \\
&+ \langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}^{k+1}_{i_k} \rangle. \qquad (74)
\end{aligned}$$

We split the cross term to have

$$\begin{aligned}\langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}_{i_k}^{k+1}\rangle &= \langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}_{i_k}^{k}\rangle \\ &+\langle \nabla_{i_k} f(\mathbf{x}^{k}), \mathbf{x}_{i_k}^{k} - \mathbf{x}_{i_k}^{k+1}\rangle + \langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}) - \nabla_{i_k} f(\mathbf{x}^{k}), \mathbf{x}_{i_k}^{k} - \mathbf{x}_{i_k}^{k+1}\rangle.\end{aligned}$$

From (7), it follows that

$$\langle \nabla_{i_k} f(\mathbf{x}^k), \mathbf{x}_{i_k}^k - \mathbf{x}_{i_k}^{k+1}\rangle \le f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) + \tfrac{L_c}{2}\|\mathbf{x}_{i_k}^k - \mathbf{x}_{i_k}^{k+1}\|^2.$$

Plugging the above two equations into (74) gives

$$\begin{aligned}&\left\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^{k+1})\right\|^2 \\ \le& \left\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 - (1-\eta L_c)\|\mathbf{x}_{i_k}^{k+1} - \mathbf{x}_{i_k}^{k}\|^2 \\ &+ 2\eta\langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}_{i_k}^{k}\rangle \\ &+ 2\eta\langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}) - \nabla_{i_k} f(\mathbf{x}^{k}), \mathbf{x}_{i_k}^{k} - \mathbf{x}_{i_k}^{k+1}\rangle \\ &+ 2\eta\left[f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) + r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}) - r_{i_k}(\mathbf{x}_{i_k}^{k+1})\right].\end{aligned} \tag{75}$$

Substituting (67) through (69) into (75) and rearranging terms yield

$$\begin{aligned}&\mathbb{E}\left\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^{k+1})\right\|^2 \\ \le& \mathbb{E}\left\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 - \tfrac{1}{m}\left[1 - \eta L_c - \tfrac{2\eta L_f \gamma_{\rho,2}^2}{m} - \tfrac{2\eta L_r \gamma_{\rho,2}}{\sqrt{m}}\right]\mathbb{E}\|\mathbf{d}^k\|^2 \\ &+ \tfrac{2\eta}{m}\mathbb{E}[F^* - F(\mathbf{x}^k)] + 2\eta\mathbb{E}[F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})]\end{aligned}$$

The above inequality together with (64) implies

$$\begin{aligned}&\mathbb{E}\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^{k+1})\|^2 \\ \le& \mathbb{E}\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\|^2 + \tfrac{2\eta}{m}\mathbb{E}[F^* - F(\mathbf{x}^k)] + 2\eta\mathbb{E}[F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})]\end{aligned}$$

and thus, with the monotonicity of $\mathbb{E}[F(\mathbf{x}^k)]$ in Lemma 11,

$$\begin{aligned}&\mathbb{E}\left\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^{k+1})\right\|^2 + 2\eta\mathbb{E}[F(\mathbf{x}^{k+1}) - F^*] \\ \le& \mathbb{E}\left\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 + 2\eta\mathbb{E}[F(\mathbf{x}^k) - F^*] - \tfrac{2\eta}{m}\mathbb{E}[F(\mathbf{x}^k) - F^*] \qquad (76)\\ \le& \left\|\mathbf{x}^{0} - \mathcal{P}_{X^*}(\mathbf{x}^{0})\right\|^2 + 2\eta\mathbb{E}[F(\mathbf{x}^0) - F^*] - \tfrac{2\eta}{m}\textstyle\sum_{t=0}^{k}\mathbb{E}[F(\mathbf{x}^t) - F^*] \\ \le& \left\|\mathbf{x}^{0} - \mathcal{P}_{X^*}(\mathbf{x}^{0})\right\|^2 + 2\eta\mathbb{E}[F(\mathbf{x}^0) - F^*] - \tfrac{2\eta}{m}(k+1)\mathbb{E}[F(\mathbf{x}^{k+1}) - F^*]. \qquad (77)\end{aligned}$$

Hence, (65) follows.

When F is strongly convex with constant μ, we have

$$F(\mathbf{x}^k) - F^* \ge \tfrac{\mu}{2}\|\mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k)\|^2,$$

and thus from (76), it follows that

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}^{k+1}-\mathcal{P}_{X^*}(\mathbf{x}^{k+1})\right\|^2+2\eta\mathbb{E}[F(\mathbf{x}^{k+1})-F^*]\\
\leq&\mathbb{E}\left\|\mathbf{x}^{k}-\mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2+\left(2\eta-\tfrac{2\eta^2\mu}{m(1+\eta\mu)}\right)\mathbb{E}[F(\mathbf{x}^{k})-F^*]\\
&-\left(\tfrac{2\eta}{m}-\tfrac{2\eta^2\mu}{m(1+\eta\mu)}\right)\tfrac{\mu}{2}\mathbb{E}\|\mathbf{x}^{k}-\mathcal{P}_{X^*}(\mathbf{x}^{k})\|^2\\
=&\left(1-\tfrac{\eta\mu}{m(1+\eta\mu)}\right)\left(\mathbb{E}\left\|\mathbf{x}^{k}-\mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2+2\eta\mathbb{E}[F(\mathbf{x}^{k})-F^*]\right).
\end{aligned}$$

Therefore, (66) follows, and we complete the proof. □

5 Poisson distribution

We can treat the asynchronous reading and writing as a queueing system. Assume the $p+1$ processors have the same computing power (i.e., the same speed of reading and writing). At any time k, suppose the update to $\mathbf{x}_{i_k}$ is performed by the p_k-th processor, which can be treated as the server with speed (or service rate) *one* of reading and writing. All the other p processors can be treated as customers, each with speed (or arrival rate) *one*, where any update to $\mathbf{x}$ from the p processors can be regarded as one customer's arrival. Under this setting, from the p_k-th processor starts reading $\mathbf{x}$ until it finishes updating $\mathbf{x}_{i_k}$, there would be p customers in the queue in average, namely, the delay j_k follows the Poisson distribution with parameter p. Summarizing the above discussion, we have the following result.

Claim Suppose Algorithm 1 runs on a system with $p+1$ processors, which have the same speed of reading and writing during the iterations. Then the delay j_k follows the Poisson distribution with parameter p, i.e., for all k,

$$\text{Prob}(j_k=t)=\frac{p^t e^{-p}}{t!},\ t=0,1,\ldots, \tag{78}$$

which implies no delay if $p=0$.

In general, if the processors have different computing power, j_k would follow Poisson distribution with a parameter being the speed ratio of the other p processors to the p_k-th one. However, in a multi-core workstation with shared memory, the processors are usually of the same style and can have the same computing ability. In the following, we assume the distribution in (8) to be Poisson distribution with parameter p and discuss the convergence results we obtained in the previous sections. First we give the values of the expected quantities we used before.

Proposition 1 *Suppose there are $p+1$ processors and* (78) *holds. Then for any $\rho > 1$, we have that for all k,*

$$\begin{aligned} &T = \mathbb{E}[\mathbf{j}] = p, && S = \mathbb{E}[\mathbf{j}^2] = p(p+1), \\ &M_\rho = \mathbb{E}[\rho^{\mathbf{j}}] = e^{p(\rho-1)}, && N_\rho = \mathbb{E}[\mathbf{j}\rho^{\mathbf{j}}] = \rho p e^{p(\rho-1)}, \\ &\gamma_{\rho,1} = \tfrac{e^{p(\sqrt{\rho}-1)}-1}{\sqrt{\rho}-1}, && \gamma_{\rho,2} = \left(\tfrac{\rho p e^{p(\rho-1)}-p}{1-\rho^{-1}}\right)^{-1}. \end{aligned} \tag{79}$$

where $\gamma_{\rho,1}$ and $\gamma_{\rho,2}$ are defined in (51).

The proof of this proposition is standard. From the quantities in (79) and the theorems we established in the previous sections, we make the following observations:

1. If $p = o(\sqrt{m})$, we can guarantee the convergence of Algorithm 1 for both smooth and nonsmooth problems by setting $\eta \lessapprox \frac{1}{L_c}$ (see Theorems 2 and 5), where $\lessapprox$ means "less than but close to";
2. If $2e^2(p+1)+p = o(\sqrt{m})$, then choosing $\rho = 1+\frac{1}{p}$, we have the convergence rate of Algorithm 1 obtained in Theorem 4 by setting $\eta \lessapprox \frac{2}{eL_c}$. Then $D \approx \frac{\eta}{m}$ in (31), and thus near-linear speedup is achieved for solving convex smooth problems;
3. If $p = o(\sqrt[4]{m})$, we can guarantee the convergence rate of Algorithm 1 in Theorem 6 by setting $\eta \lessapprox \frac{1}{L_c}$ and thus a near-linear speedup for convex nonsmooth problems.

6 Numerical experiments

In this section, we evaluate the numerical performance of Algorithm 1 on solving two problems: the LASSO problem and the nonnegative matrix factorization (NMF). The tests were carried out on a machine with 64GB of memory and two Intel Xeon E5-2690 v2 processors (20 cores, 40 threads). All of the experiments were coded in C++ and its threading library was used for parallelization. We use the Eigen library for numerical linear algebra operations. To measure the delay, we use an atomic variable to track the number of iterations as defined in the paper. The atomic variable will be incremented by one for each update. For each thread, the delay is calculated based on the difference of the iteration counters before and after the update. For LASSO, two different settings were used. The first one sets the stepsize by the expected delay according to the analysis of this paper, and the other one used the maximum delay from [16,17] and is dubbed as AsySCD. We compared the async-BCU to the serial BCU, which can be regarded as a special case of Algorithm 1 with the delay $j_k \equiv 0$, $\forall k$. For NMF, we set the stepszie by the expected delay and test its convergence behavior with different numbers of threads.

6.1 Parameter settings

According to Theorem 5, the following two stepsizes were used:[1]

$$\text{This paper}: \ \eta = \frac{1/L_c}{1+\kappa^2 p^2/(2m)}, \tag{80a}$$

$$\text{Max delay}: \ \eta = \frac{1/L_c}{1+\kappa^2 \tau^2/(2m)}, \tag{80b}$$

where τ equals the maximum number of the generated sequence of delays.

6.2 LASSO

We measure the performance of Algorithm 1 on the LASSO problem [30]

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \ \tfrac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_1, \tag{81}$$

where $\mathbf{A}\in\mathbb{R}^{N\times n}$, $\mathbf{b}\in\mathbb{R}^N$, and λ is a parameter balancing the fitting term and the regularization term. We randomly generated $\mathbf{A}$ and $\mathbf{b}$ following the standard normal distribution. The size was fixed to $n = 2N$ and $N = 10,000$, and $\lambda = \frac{1}{N}$ was used. The Lipschitz constant $L_c = \max\{\|(\mathbf{A}_i^T\mathbf{A}_i)\|^2, \ \forall i\}$, where $\mathbf{A}_i$ represents the ith column block of $\mathbf{A}$.

Figure 1 shows the delay distribution of Algorithm 1 with different numbers of threads. The blue bars are the normalized histogram so that the bar heights add to 1. Orange curve is the probability density function of Poisson distribution. By using 5 and 10 threads, we observe that the number of delays is concentrated on 4, and 9 respectively. When the number of threads is relatively large, the actual delay distribution closely matches with the theoretical distribution as we discussed in Section 5. For 20 threads, an interesting observation is that, the actual probability density is higher than the theoretical probability density when the number of delays is around 9. We think this is due to the architecture of the testing environment, i.e., the average delay within a CPU is smaller than the average delay across two different CPUs. We observe a similar behavior when 40 threads are used.

Figure 2 plots the convergence behavior of Algorithm 1 running on 40 threads with different block sizes. We partition $\mathbf{x}$ into m equal-sized blocks with block sizes varying among $\{10, \ 50, \ 100, \ 500\}$. The results of the serial randomized coordinate descent method is also plotted for comparison. Here, one epoch is equivalent to updating all coordinates once. Comparing to the serial method, we observe that the delay does affect the convergence speed, and the affect becomes weaker as m increases. Hence, Algorithm 1 can have nearly linear speed-up when the number of blocks is large. In addition, we note that the stepsize setting of AsySCD is too conservative, and Algorithm 1 with stepsize set by the expected delay converges significantly faster. However, we observed that, in general, we could not take larger stepsize than that in (80a).

[1] For the NMF problem, L_c cannot be determined in the beginning, so instead of using a uniform L_c, we used the gradient Lipschitz constant adaptive to the iterate.

Some divergence behaviors are observed when using stepsizes larger than that in (80a).

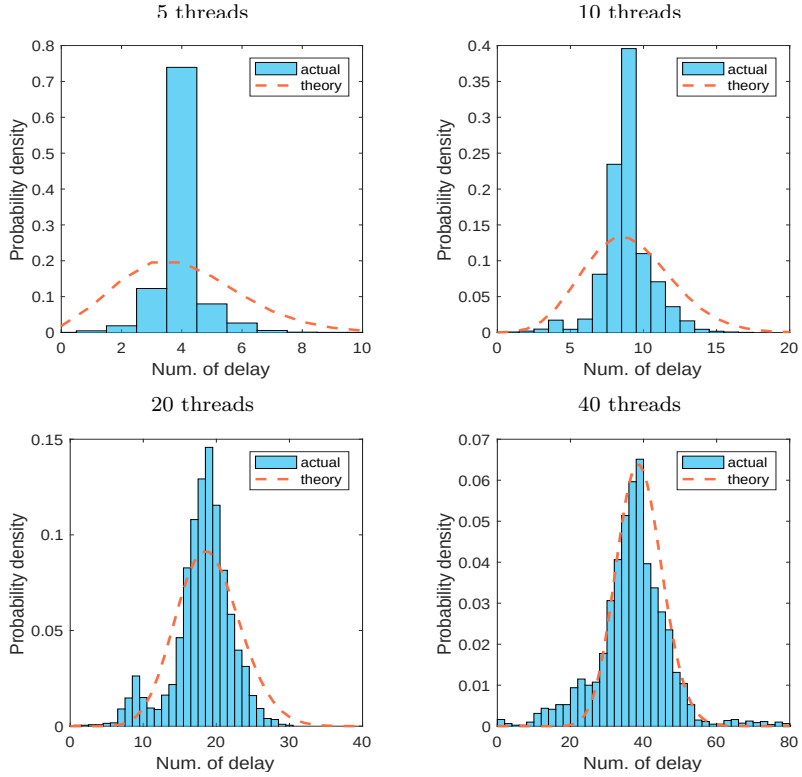

Fig. 1 Delay distribution behaviors of Algorithm 1 for solving LASSO (81). The tested problem has 20,000 coordinates, and it was running with 5, 10, 20, and 40 threads.

6.3 Nonnegative matrix factorization (NMF)

This section presents the numerical results of applying Algorithm 1 for solving the NMF problem [23]

$$\begin{aligned} &\underset{\mathbf{X},\mathbf{Y}}{\text{minimize}} \ \ \tfrac{1}{2}\|\mathbf{X}\mathbf{Y}^\top - \mathbf{Z}\|_F^2, \\ &\text{s.t. } \mathbf{X} \in \mathbb{R}_+^{M\times m}, \mathbf{Y} \in \mathbb{R}_+^{N\times m}, \end{aligned} \tag{82}$$

where $\mathbf{Z} \in \mathbb{R}_+^{M\times N}$ is a given nonnegative matrix. We generated $\mathbf{Z} = \mathbf{Z}_L\mathbf{Z}_R^\top$ with the elements of $\mathbf{Z}_L$ and $\mathbf{Z}_R$ first drawn from the standard normal distribution and then projected into the nonnegative orthant. The size was fixed to $M = N = 10,000$ and $m = 100$.

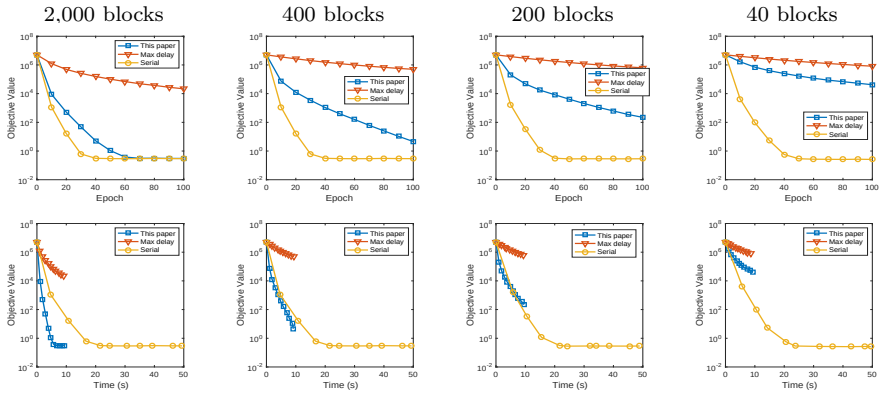

Fig. 2 Convergence behaviors of Algorithm 1 for solving the LASSO problem (81) with the stepsize given in (80), and also the serial randomized coordinate descent method. The tested problem has $10,000$ samples and $20,000$ coordinates that are evenly partitioned into m blocks. It was simulated as running with 40 threads. We run 100 epochs for each experiments.

We treated one column of $\mathbf{X}$ or $\mathbf{Y}$ as one block coordinate, and during the iterations, every column of $\mathbf{X}$ was kept with unit norm. Therefore, the partial gradient Lipschitz constant equals *one* if one column of $\mathbf{Y}$ is selected to update and $\|\mathbf{y}_{i_k}^k\|_2^2$ if the i_k-th column of $\mathbf{X}$ is selected. Since $\|\mathbf{y}_{i_k}^k\|_2^2$ could approach to *zero*, we set the Lipschitz constant to $\max(0.001, \|\mathbf{y}_{i_k}^k\|_2^2)$. This modification can guarantee the whole sequence convergence of the coordinate descent method [36]. Due to nonconvexity, global optimality cannot be guaranteed. Thus, we set the starting point close to $\mathbf{Z}_L$ and $\mathbf{Z}_R$. Specifically, we let $\mathbf{X}^0 = \mathbf{Z}_L + 0.5\mathbf{\Xi}_L$ and $\mathbf{Y}^0 = \mathbf{Z}_R + 0.5\mathbf{\Xi}_R$ with the elements of $\mathbf{\Xi}_L$ and $\mathbf{\Xi}_R$ following the standard normal distribution. All methods used the same starting point.

Figure 3 shows the delay distribution behavior of Algorithm 1 for solving NMF. The observation is similar to Figure 1. Figure 4 plots the convergence results of Algorithm 1 running with 1, 5, 10, 20 and 40 threads. From the results, we observe that Algorithm 1 scales up to 10 threads for the tested problem. Degenerated convergence is observed with 20 and 40 threads. This is mostly due to the following three reasons: (1) since the number of blocks is relatively small ($m = 200$), as shown in (80a), using more threads leads to smaller stepsize, hence, slower convergence; (2) the gradient used for the current update is more staled when a relative large number of threads are used, which also leads to slow convergence; (3) high cache miss rates and false sharing also downgrade the speedup performance.

7 Conclusions

We have analyzed the convergence of the async-BCU method for solving both convex and nonconvex problems in a probabilistic way. We showed that the

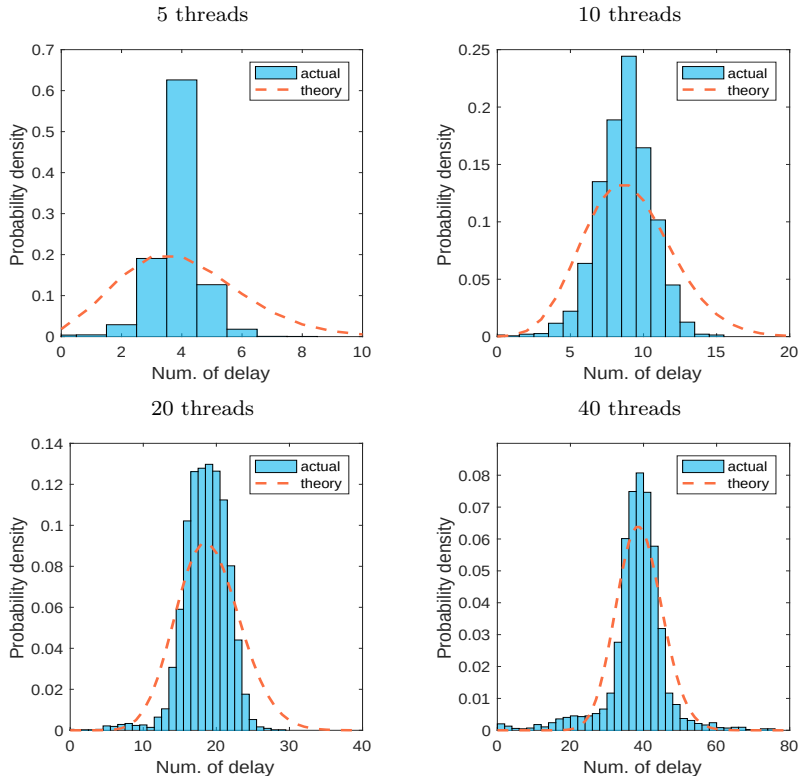

Fig. 3 Delay distribution behaviors of Algorithm 1 for solving NMF (82). It was running with 5, 10, 20, and 40 threads.

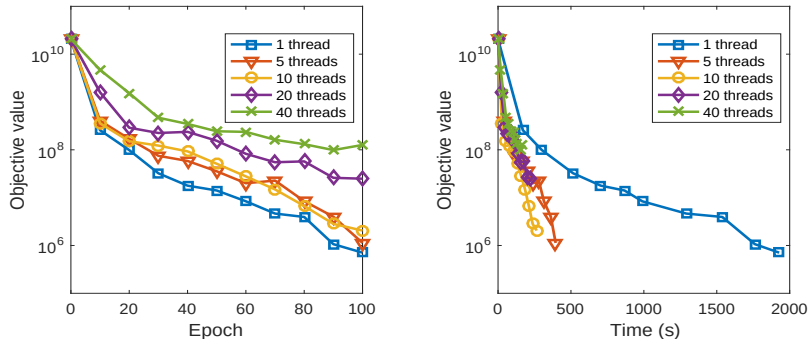

Fig. 4 Convergence behaviors of Algorithm 1 for solving the NMF problem (82) with the stepsize set based on the expected delay. The size of the tested problem is $M = N = 10,000$ and $m = 100$, i.e., 200 block coordinates, and the algorithm was tested with 1, 5, 10, 20, and 40 threads.

algorithm is guaranteed to converge for smooth problems if the expected delay

is finite and for nonsmooth problems if the variance of the delay is also finite. In addition, we established sublinear convergence of the method for weakly convex problems and linear convergence for strongly convex ones. The stepsize we obtained depends on certain expected quantities. Assuming the given $p+1$ processors perform identically, we showed that the delay follows a Poisson distribution with parameter p and thus fully determined the stepsize. We have simulated the performance of the algorithm with our determined stepsize on solving LASSO and the nonnegative matrix factorization, and the numerical results validated our analysis.

Acknowledgements

We would like to acknowledge support for this project from the National Science Foundation (NSF EAGER ECCS-1462397, DMS-1621798, and DMS-1719549).

References

1. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009)
2. Bertsekas, D.P.: Distributed asynchronous computation of fixed points. Mathematical Programming **27**(1), 107–120 (1983)
3. Cannelli, L., Facchinei, F., Kungurtsev, V., Scutari, G.: Asynchronous parallel algorithms for nonconvex big-data optimization: Model and convergence. arXiv preprint arXiv:1607.04818 (2016)
4. Chazan, D., Miranker, W.: Chaotic relaxation. Linear Algebra and its Applications **2**(2), 199–222 (1969)
5. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.i.: Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons (2009)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
7. Dang, C.D., Lan, G.: Stochastic block mirror descent methods for nonsmooth and stochastic optimization. SIAM Journal on Optimization **25**(2), 856–881 (2015)
8. Davis, D.: The asynchronous PALM algorithm for nonsmooth nonconvex problems. arXiv preprint arXiv:1604.00526 (2016)
9. Frommer, A., Szyld, D.B.: On asynchronous iterations. Journal of Computational and Applied Mathematics **123**(1), 201–216 (2000)
10. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. Operations Research Letters **26**(3), 127–136 (2000)
11. Gut, A.: Probability: A Graduate Course: A Graduate Course. Springer Science & Business Media (2006)
12. Hannah, R., Yin, W.: On unbounded delays in asynchronous parallel fixed-point algorithms. arXiv preprint arXiv:1609.04746 (2016)
13. Hildreth, C.: A quadratic programming procedure. Naval Research Logistics Quarterly **4**(1), 79–85 (1957)
14. Hong, M., Wang, X., Razaviyayn, M., Luo, Z.Q.: Iteration complexity analysis of block coordinate descent methods. Mathematical Programming **163**(1-2), 85–114 (2017)
15. Lai, M.J., Yin, W.: Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. SIAM Journal on Imaging Sciences **6**(2), 1059–1091 (2013)

16. Liu, J., Wright, S., Re, C., Bittorf, V., Sridhar, S.: An asynchronous parallel stochastic coordinate descent algorithm. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 469–477 (2014)
17. Liu, J., Wright, S.J.: Asynchronous stochastic coordinate descent: Parallelism and convergence properties. SIAM Journal on Optimization **25**(1), 351–376 (2015)
18. Lu, Z., Xiao, L.: On the complexity analysis of randomized block-coordinate descent methods. Mathematical Programming **152**(1-2), 615–642 (2015)
19. Luo, Z.Q., Tseng, P.: On the convergence of the coordinate descent method for convex differentiable minimization. Journal of Optimization Theory and Applications **72**(1), 7–35 (1992)
20. Mokhtai, A., Koppel, A., Ribeiro, A.: A class of parallel doubly stochastic algorithms for large-scale learning. arXiv preprint arXiv:1606.04991 (2016)
21. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization **19**(4), 1574–1609 (2009)
22. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization **22**(2), 341–362 (2012)
23. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994)
24. Peng, Z., Xu, Y., Yan, M., Yin, W.: Arock: An algorithmic framework for asynchronous parallel coordinate updates. SIAM Journal on Scientific Computing **38**(5), A2851–A2879 (2016)
25. Recht, B., Re, C., Wright, S., Niu, F.: Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In: Advances in Neural Information Processing Systems, pp. 693–701 (2011)
26. Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Mathematical Programming **144**(1-2), 1–38 (2014)
27. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer Science & Business Media (2009)
28. Rosenfeld, J.L.: A case study in programming for parallel-processors. Communications of the ACM **12**(12), 645–655 (1969)
29. Strikwerda, J.C.: A probabilistic analysis of asynchronous iteration. Linear Algebra and its Applications **349**(13), 125 – 154 (2002)
30. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
31. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications **109**(3), 475–494 (2001)
32. Tseng, P., Bertsekas, D.P., Tsitsiklis, J.N.: Partially asynchronous, parallel algorithms for network flow and other problems. SIAM Journal on Control and Optimization **28**(3), 678–710 (1990)
33. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming **117**(1-2), 387–423 (2009)
34. WhiteHouse: Big Data: Seizing Opportunities Preserving Values (2014)
35. Xu, Y., Yin, W.: Block stochastic gradient iteration for convex and nonconvex optimization. SIAM Journal on Optimization **25**(3), 1686–1716 (2015)
36. Xu, Y., Yin, W.: A globally convergent algorithm for nonconvex optimization based on block coordinate update. Journal of Scientific Computing **72**(2), 700–734 (2017)
37. Zhou, H., Li, L., Zhu, H.: Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association **108**(502), 540–552 (2013)
38. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics **15**(2), 265–286 (2006)

A Proofs of lemmas

The following lemma is used in other proofs several times, and it is easy to verify.

Lemma 12 *For any scalar sequences* $\{a_{i,j}\}$ *and* $\{b_i\}$*, it holds that*

$$\sum_{t=1}^{k-1}\sum_{d=k-t}^{k-1} a_{d,t} = \sum_{d=1}^{k-1}\sum_{t=k-d}^{k-1} a_{d,t},\ \forall k \geq 0, \tag{83}$$

$$\sum_{t=1}^{k}\sum_{d=0}^{t-1} a_{d,t} = \sum_{d=0}^{k-1}\sum_{t=d+1}^{k} a_{d,t},\ \forall k \geq 0. \tag{84}$$

$$\sum_{t=1}^{k}\sum_{d=1}^{t-1} a_{d,t}b_{t-d} = \sum_{t=1}^{k-1}\Big(\sum_{d=t+1}^{k} a_{d-t,d}\Big)b_t,\ \forall k \geq 0. \tag{85}$$

A.1 Proof of Lemma 3

Proof Following the proof of Theorem 1 in [16], we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla f(\mathbf{x}^t)\|^2 - \|\nabla f(\mathbf{x}^{t+1})\|^2] \\
\leq& 2\mathbb{E}[\|\nabla f(\mathbf{x}^t)\| \cdot \|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^{t+1})\|] \quad (\text{from } \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2 \leq 2\|\mathbf{u}\| \cdot \|\mathbf{u}-\mathbf{v}\|) \\
\leq& 2L_r\mathbb{E}[\|\nabla f(\mathbf{x}^t)\| \cdot \|\mathbf{x}^t - \mathbf{x}^{t+1}\|] = 2\eta L_r\mathbb{E}[\|\nabla f(\mathbf{x}^t)\| \cdot \|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|] \\
\leq& \eta L_r\left(\frac{1}{\sqrt{m}}\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \sqrt{m}\mathbb{E}\|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|^2\right) \\
=& \frac{\eta L_r}{\sqrt{m}}\left(\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \mathbb{E}\|\nabla f(\mathbf{x}^{t-j_t})\|^2\right) \\
=& \frac{\eta L_r}{\sqrt{m}}\Big(\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \sum_{r=0}^{t-1} q_r\mathbb{E}\|\nabla f(\mathbf{x}^{t-r})\|^2 + c_t\|\nabla f(\mathbf{x}^0)\|^2\Big)
\end{aligned} \tag{86}$$

and

$$\begin{aligned}
&\mathbb{E}[\|\nabla f(\mathbf{x}^{t+1})\|^2 - \|\nabla f(\mathbf{x}^t)\|^2] \\
\leq& \mathbb{E}[\|\nabla f(\mathbf{x}^{t+1}) + \nabla f(\mathbf{x}^t)\| \cdot \|\nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t)\|] \\
\leq& L_r\mathbb{E}\left[(2\|\nabla f(\mathbf{x}^t)\| + \|\nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t)\|)\|\mathbf{x}^{t+1} - \mathbf{x}^t\|\right] \\
\leq& L_r\mathbb{E}\left[2\|\nabla f(\mathbf{x}^t)\| \cdot \|\mathbf{x}^{t+1} - \mathbf{x}^t\| + L_r\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2\right] \\
=& L_r\mathbb{E}\left[2\eta\|\nabla f(\mathbf{x}^t)\| \cdot \|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\| + \eta^2 L_r\|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|^2\right] \\
\leq& L_r\mathbb{E}\left[\frac{\eta}{\sqrt{m}}\|\nabla f(\mathbf{x}^t)\|^2 + \eta\sqrt{m}\|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|^2 + \eta^2 L_r\|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|^2\right] \\
=& \frac{\eta L_r}{\sqrt{m}}\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \left(\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)\mathbb{E}\|\nabla f(\mathbf{x}^{t-j_t})\|^2 \\
=& \frac{\eta L_r}{\sqrt{m}}\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \left(\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)\left(\sum_{r=0}^{t-1} q_r\mathbb{E}\|\nabla f(\mathbf{x}^{t-r})\|^2 + c_t\|\nabla f(\mathbf{x}^0)\|^2\right).
\end{aligned} \tag{87}$$

We first show the first inequality in (26). Note that (25) gives us

$$\frac{1}{1-(1+M_\rho)\frac{\eta L_r}{\sqrt{m}}} \leq \rho. \tag{88}$$

When $t = 0$, we have from (86) that $\|\nabla f(\mathbf{x}^0)\|^2 - \mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2 \leq \frac{2\eta L_r}{\sqrt{m}}\|\nabla f(\mathbf{x}^0)\|^2 \leq (1 + M_\rho)\frac{\eta L_r}{\sqrt{m}}\|\nabla f(\mathbf{x}^0)\|^2$. Hence, $\|\nabla f(\mathbf{x}^0)\|^2 \leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2$ from (88). Now we assume that

$\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 \leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^{t+1})\|^2$ for all $t \leq k-1$. For $t=k$, it holds from (86) and the induction assumption that

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2 \\
\leq &\frac{\eta L_r}{\sqrt{m}}\left(\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \sum_{t=0}^{k-1} q_t\rho^t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + c_k\rho^k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2\right) \\
= &\frac{\eta L_r}{\sqrt{m}}\left(1 + \sum_{t=0}^{k-1} q_t\rho^t + c_k\rho^k\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \leq \frac{\eta L_r}{\sqrt{m}}(1+M_\rho)\cdot\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2.
\end{aligned}$$

Hence, we have $\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2$ from (88). Therefore, we finish the induction step, and thus the first inequality of (26) holds.

Next we show the second inequality of (26). Since (25) implies $\eta \leq \frac{\rho-1}{\frac{L_r}{\sqrt{m}}\left(1+M_\rho+\frac{(\rho-1)M_\rho}{\rho(1+M_\rho)}\right)}$,

$$\begin{aligned}
1 + \frac{\eta L_r}{\sqrt{m}} + \left(\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right) M_\rho &\overset{(25)}{\leq} 1 + \frac{\eta L_r}{\sqrt{m}}(1+M_\rho) + M_\rho\frac{\eta L_r^2}{m}\frac{(\rho-1)\sqrt{m}}{\rho L_r(1+M_\rho)} \\
&= 1 + \frac{\eta L_r}{\sqrt{m}}\left(1 + M_\rho + \frac{(\rho-1)M_\rho}{\rho(1+M_\rho)}\right) \leq \rho. \qquad (89)
\end{aligned}$$

When $t=0$, we have from (87) that

$$\begin{aligned}
\mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2 - \|\nabla f(\mathbf{x}^0)\|^2 &\leq \left(\frac{2\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)\|\nabla f(\mathbf{x}^0)\|^2 \\
&\leq \left((1+M_\rho)\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)\|\nabla f(\mathbf{x}^0)\|^2.
\end{aligned}$$

Hence, $\mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2 \leq \rho\|\nabla f(\mathbf{x}^0)\|^2$ holds from (89). Assume $\mathbb{E}\|\nabla f(\mathbf{x}^{t+1})\|^2 \leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2$ for all $t \leq k-1$. It follows from (87) and the induction assumption that

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2 - \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
\leq &\frac{\eta L_r}{\sqrt{m}}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \left(\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)\left(\sum_{t=0}^{k-1} q_t\rho^t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + c_k\rho^k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2\right) \\
= &\left(\frac{\eta L_r}{\sqrt{m}} + \left(\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)\left(\sum_{t=0}^{k-1} q_t\rho^t + c_k\rho^k\right)\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
\leq &\left(\frac{\eta L_r}{\sqrt{m}} + \left(\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)M_\rho\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2.
\end{aligned}$$

Hence, from (89), $\mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2 \leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2$ holds, and we complete the proof. □