Journal of Machine Learning Research x (xxxx) xx-xx Submitted x/xx; Published xx/xx

On the Convergence of Asynchronous Parallel Iteration with Arbitrary Delays

Zhimin Peng ZHIMINP@GMAIL.COM
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA, 90095

Yangyang Xu YANGYANG.XU@UA.EDU
Department of Mathematics
University of Alabama
Tuscaloosa, AL, 35487

Ming Yan YANM@MATH.MSU.EDU
Department of Computational Mathematics, Science and Engineering
Department of Mathematics
Michigan State University
East Lansing, MI 48824

Wotao Yin WOTAOYIN@MATH.UCLA.EDU
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA, 90095

Editor: xxx

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 Because of asynchrony, the iterates are computed with outdated information, and the age of the outdated information, which we call *delay*, is the number of times it has been updated since its creation.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 arbitrarily large 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 parallel, arbitrary delay, Nonconvex, convex, convergence analysis

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 WhiteHouse (2014). 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

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

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 core machine learning problems: support vector machine (squared hinge loss and its dual formulation) Cortes and Vapnik (1995), LASSO Tibshirani (1996), and logistic regression (linear or multi-linear) Zhou et al. (2013), and also subspace learning problems: sparse principal component analysis Zou et al. (2006), nonnegative matrix or tensor factorization Cichocki et al. (2009), 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 Beck and Teboulle (2009), stochastic approximation Nemirovski et al. (2009), randomized coordinate descent Nesterov (2012), and their combinations Dang and Lan (2015); Xu and Yin (2015). Recently, lots of efforts have been made to the parallelization of these methods, and in particular, asynchronous parallel (async-parallel) methods attract more attention (for example, Liu et al. (2014); Peng et al. (2016)) 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 Strikwerda (2002), and thus they require strong conditions (like a contraction) for convergence. Recent works, e.g., Liu et al. (2014); Liu and Wright (2015); Peng et al. (2016); Hannah and Yin (2016), 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 Hannah and Yin (2016) allows unbounded delays but requires convexity, and Davis (2016); Davis et al. (2016); Cannelli et al. (2016) do not assume convexity but require bounded delays. We will allow arbitrarily large delays and also 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 $i_k \in [m] = \{1, 2, \cdots, m\}$, 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 Liu et al. (2014), 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 there, 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 arbitrarily large delays following a certain distribution. Our main assumption on the delay is the boundedness of certain expected quantities (e.g., expected delay, variance of delay), so our results are more general than those in Liu et al. (2014); Liu and Wright (2015), which assume only bounded delays.

- Both nonconvex and convex problems are analyzed for both smooth and nonsmooth cases. 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 have the same computing power, the delay follows the Poisson distribution with parameter p. Hence, all the expected quantities we use

can be explicitly computed and are bounded. In addition, by setting appropriate stepsizes, we can reach near-linear speedup if $p = o(\sqrt{m})$ for smooth cases and $p = o(\sqrt[4]{m})$ for nonsmooth cases.

- According to the Poisson distribution, we can explicitly set the stepsize based on the expected delay (that 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 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 each of these two 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., Xu and Yin (2015); Mokhtai et al. (2016); Davis (2016)), 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

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

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*

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

and $\partial R(\mathbf{x})$ denotes the subdifferential of R at $\mathbf{x}$.

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 constants L_r and L_c, respectively, namely, 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,

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

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, $\hat{\mathbf{x}}^k = \mathbf{x}^{k-j_k}$, and the delay follows an identical distribution*

$$\text{Prob}(j_k = t) = q_t,\ t = 0, 1, 2 \ldots, \forall k. \tag{8}$$

Remark 2 *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* Strikwerda (2002)*. Our results can still hold for non-identical distribution; see the analysis for the smooth nonconvex case in the supplementary materials.*

2. Related works

We briefly review block coordinate update (BCU) methods 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 Hildreth (1957) 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 Luo and Tseng (1992); Grippo and Sciandrone (2000); Tseng (2001). However, in general, its convergence rate result was only shown for strongly convex problems (e.g., Luo and Tseng (1992)) until the recent work Hong et al. (2016) that shows sublinear convergence for weakly convex cases. Tseng and Yun (2009) 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 Nesterov (2012) that 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 Nesterov (2012) and then generalized to nonsmooth cases in Richtárik and Takáč (2014); Lu and Xiao (2015). Recently, Dang and Lan (2015); Xu and Yin (2015) 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 Rosenfeld (1969) to solve linear equations arising in electrical network problems. Chazan and Miranker (1969) first systematically analyzed (more general) asynchronous iterative methods for solving linear systems. Assuming bounded delays, it gave a necessary and sufficient condition for convergence. Bertsekas (1983) proposed an asynchronous distributed iterative method for solving more general fixed-point problems and showed its convergence under a contraction assumption. Tseng et al. (1990) weakened the contraction assumption to pseudo-nonexpansiveness but made more other assumptions. Frommer and Szyld (2000) made a thorough review of asynchronous methods before 2000. It summarized convergence results under nested sets and synchronous convergence conditions, which are satisfied by max-norm 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, Liu et al. (2014); Liu and Wright (2015) 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 delays to be bounded by a finite integer τ. 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, Davis (2016); Davis et al. (2016) introduced an async-parallel coordinate descent method, whose convergence was established under iterate boundedness assumptions and with appropriate stepsize.

3. Convergence results for the smooth case

Throughout this section, we assume $r_i = 0$, $\forall i$, namely, we consider the smooth optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}). \tag{9}$$

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

We first establish a subsequence convergence result for the general (possibly nonconvex) case. The main result is summarized in the following theorem.

Theorem 3 (Convergence for the nonconvex smooth case) *Under Assumptions 1 through 3, let* $\{\mathbf{x}^k\}_{k\geq 1}$ *be generated from Algorithm 1. Assume*

$$T := \mathbb{E}[j_k] < \infty. \tag{10}$$

If the stepsize is taken as $0 < \eta < \frac{1/L_c}{1+2\kappa T/\sqrt{m}}$*, then*

$$\lim_{k\to\infty} \mathbb{E}\|\nabla f(\mathbf{x}^k)\| = 0, \tag{11}$$

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

Remark 4 *From the theorem, we see that if* $\mathbb{E}[j_k] = o(\sqrt{m})$*, then* η *only weakly depends on the delay.*

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.

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

$$M_\sigma := \mathbb{E}[\sigma^{j_k}] < \infty. \tag{12}$$

The condition in (12) is stronger than that in (10), 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 5 *Under Assumptions 2 through 4, we have that for any* $1 < \rho \leq \sigma$*, if the stepsize satisfies*

$$0 < \eta \leq \frac{(\rho-1)\sqrt{m}}{\rho L_r(1+M_\rho)}, \tag{13}$$

with M_ρ *defined in* (12)*, then for all* k*,*

$$\begin{aligned} \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 &\leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2 \\ \mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2 &\leq \rho\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2. \end{aligned} \tag{14}$$

The proof of Lemma 5 follows Liu et al. (2014). Using this lemma, we show that sufficient decrease of the objective can be made after each iteration.

Theorem 6 *Under Assumptions 1 through 4, let* $\{\mathbf{x}^k\}_{k\geq 1}$ *be the sequence generated from Algorithm 1. For a certain* $1 < \rho < \sigma$*, define*

$$N_\rho := \mathbb{E}[j_k\rho^{j_k}]. \tag{15}$$

Take the stepsize such that (13) *is satisfied and also*

$$0 < \eta < \frac{2/L_c}{M_\rho + \frac{\kappa(2N_\rho M_\rho + T)}{\sqrt{m}}}, \tag{16}$$

where T *and* M_ρ *are defined in* (10) *and* (12) *respectively. Let*

$$D = \frac{\eta}{2m}\left(2 - \frac{\eta L_r}{\sqrt{m}}(2N_\rho M_\rho + T) - \eta L_c M_\rho\right). \tag{17}$$

Then

$$\mathbb{E}f(\mathbf{x}^{k+1}) \leq \mathbb{E}f(\mathbf{x}^k) - D\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2. \tag{18}$$

Note that as t is sufficiently large, it holds that $t < t\rho^t < \sigma^t$ for any $1 < \rho < \sigma$. Hence, T and N_ρ are both finite if (12) holds.

Using (18) and assuming convexity of f, we establish convergence rate results of Algorithm 1 as follows.

Theorem 7 (Convergence rate for the convex smooth case) *Under the assumptions of Theorem 6, 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*

$$\mathbb{E}[f(\mathbf{x}^{k+1}) - f^*] \leq \frac{1}{[f(\mathbf{x}^0) - f^*]^{-1} + \frac{D(k+1)}{B^2}}, \tag{19}$$

   where f^* *denotes the minimum value of* (9)*.*

2. *If* f *is strongly convex with constant* μ*, then*

$$\mathbb{E}[f(\mathbf{x}^{k+1}) - f^*] \leq (1 - 2\mu D)\mathbb{E}[f(\mathbf{x}^k) - f^*], \tag{20}$$

where D *is given in* (17)*.*

Remark 8 *The sublinear convergence result in* (19) *for the weakly convex case assumes the boundedness of the iterates. This assumption can be reduced at the expense of possibly smaller stepsize; see Theorem 15.*

For linear convergence, the assumption on strongly convexity can be weakened to essential strong convexity*. The latter one is strictly weaker than the former one; see* Liu et al. (2014) *for more discussions.*

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 denote

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

a virtual full-update iterate, where R is defined in (4). 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 9 *Under Assumptions 1 through 3, let $\{\mathbf{x}^k\}_{k\geq 1}$ be the sequence generated from Algorithm 1. Assume*

$$S := \mathbb{E}[(j_k)^2] < \infty, \tag{21}$$

and the stepsize is taken as $0 < \eta < \frac{1/L_c}{1+\kappa^2 S/(2m)}$. Then

$$\sum_{k=0}^{\infty} \mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 < \infty. \tag{22}$$

Since $\big(\mathbb{E}[j_k]\big)^2 \leq \mathbb{E}[(j_k)^2]$, the condition in (21) implies that in (10). The result in (22) indicates that $\mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\| \to 0$ as $k \to \infty$. Together with (21), we are able to show $\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|$ also approaches to *zero*, as summarized in the following.

Lemma 10 *Under the assumptions of Lemma 9, we have*

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

Using Lemmas 9 and 10, we are ready to establish the almost sure global convergence of Algorithm 1.

Theorem 11 *Under the assumptions of Lemma 9, any limit point $\mathbf{x}^*$ of $\{\mathbf{x}^k\}_{k\geq 1}$ is a critical point of (1) almost surely.*

Remark 12 *From the theorem, we see that if $S = \mathbb{E}[(j_k)^2] = o(m)$, then the stepsize required for convergence only weakly depends on the delay.*

Remark 13 (Comparison of stepsize) *The works Davis (2016); Davis et al. (2016) also consider asynchronous coordinate descent for nonconvex problems. To have convergence to a critical point, 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 11 is larger if $\kappa^2 S \leq 2m$, where S is defined in (21) and smaller than τ^2, and that can lead to faster convergence.*

4.2 Convergence rate for the convex case

In this subsection, we establish convergence rates of Algorithm 1 for nonsmooth convex cases. Similar to (14), we first show that choosing an appropriate stepsize, the iterate difference does not change too fast.

Lemma 14 *Assume Assumptions 2 through 4. Then for any $1 < \rho < \sigma$, it holds that*

$$\begin{aligned} \gamma_{\rho,1} &:= \sum_{t=1}^{\infty} q_t \tfrac{\rho^{t/2}-1}{\rho^{1/2}-1} < \infty, \\ \gamma_{\rho,2} &:= \Big(\sum_{t=1}^{\infty} q_t t \tfrac{\rho^{t}-1}{1-\rho^{-1}}\Big)^{1/2} < \infty. \end{aligned} \tag{23}$$

In addition, if the stepsize is taken such that

$$0 < \eta \leq \frac{(1-\rho^{-1})\sqrt{m}-4}{2L_r(1+\gamma_{\rho,1}+\gamma_{\rho,2})}, \tag{24}$$

then for all $k \geq 1$,

$$\mathbb{E}\|\mathbf{x}^{k-1}-\bar{\mathbf{x}}^k\|^2 \leq \rho\mathbb{E}\|\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\|^2. \tag{25}$$

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

Theorem 15 (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* (24) *and also*

$$\eta \leq \frac{1}{L_c + \frac{2L_f\gamma_{\rho,2}^2}{m} + \frac{2L_r\gamma_{\rho,2}}{\sqrt{m}}}, \tag{26}$$

where $\gamma_{\rho,1}$ *and* $\gamma_{\rho,2}$ *are defined in* (23)*. We have*

1. *If the function* F *is convex, then*

$$\mathbb{E}[F(\mathbf{x}^k)-F^*] \leq \frac{m\Phi(\mathbf{x}^0)}{2\eta(m+k)}, \tag{27}$$

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*

$$\Phi(\mathbf{x}^k) \leq \left(1-\frac{\eta\mu}{m(1+\eta\mu)}\right)^k \Phi(\mathbf{x}^0). \tag{28}$$

Remark 16 *Similar to* (20)*, for the linear convergence result* (28)*, the strong convexity assumption can be weakened to* optimal strong convexity*. The latter one is strictly weaker than the former one; see* Liu and Wright (2015) *for more discussions.*

Remark 17 (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 Liu and Wright (2015).*

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 5.1 *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,*

$$\text{Prob}(j_k = t) = \frac{p^t e^{-p}}{t!},\ t = 0, 1, \ldots, \tag{29}$$

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 18 *Suppose there are $p+1$ processors and (29) holds. Then for any $\rho > 1$, we have that for all k,*

$$T = \mathbb{E}[j_k] = p, \tag{30a}$$
$$S = \mathbb{E}[j_k^2] = p(p+1), \tag{30b}$$
$$M_\rho = \mathbb{E}[\rho^{j_k}] = e^{p(\rho-1)}, \tag{30c}$$
$$N_\rho = \mathbb{E}[j_k \rho^{j_k}] = \rho p e^{p(\rho-1)}, \tag{30d}$$
$$\gamma_{\rho,1} = \frac{e^{p(\sqrt{\rho}-1)} - 1}{\sqrt{\rho} - 1}, \tag{30e}$$
$$\gamma_{\rho,2} = \sqrt{\frac{\rho p e^{p(\rho-1)} - p}{1 - \rho^{-1}}}, \tag{30f}$$

where $\gamma_{\rho,1}$ and $\gamma_{\rho,2}$ are defined in (23).

The proof of this proposition is not difficult by using the definition of Poisson distribution. From the values of quantities in (30) and the theorems we established in the previous sections, we have 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 3 and 11), 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 7 by setting $\eta \lessapprox \frac{2}{eL_c}$. Then $D \approx \frac{\eta}{m}$ in (17), 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 15 by setting $\eta \lessapprox \frac{1}{L_c}$ and thus achieve near-linear speedup for solving 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). 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 Liu et al. (2014); Liu and Wright (2015) 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 11, the following two stepsizes were used:[1]

$$\text{This paper}: \ \eta = \frac{1/L_c}{1+\kappa^2 p^2/(2m)}, \tag{31a}$$

$$\text{Max delay}: \ \eta = \frac{1/L_c}{1+\kappa^2 \tau^2/(2m)}, \tag{31b}$$

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

6.2 LASSO

In this subsection, we measure the performance of Algorithm 1 on solving the LASSO problem Tibshirani (1996)

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{32}$$

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.

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

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.

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. Since the async-BCU method has exactly the same per-epoch computational complexity as that of the serial method, the results demonstrate near-linear speedup of async-BCU when m 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 (31a). Some divergence behaviors are observed when using stepsizes larger than that in (31a).

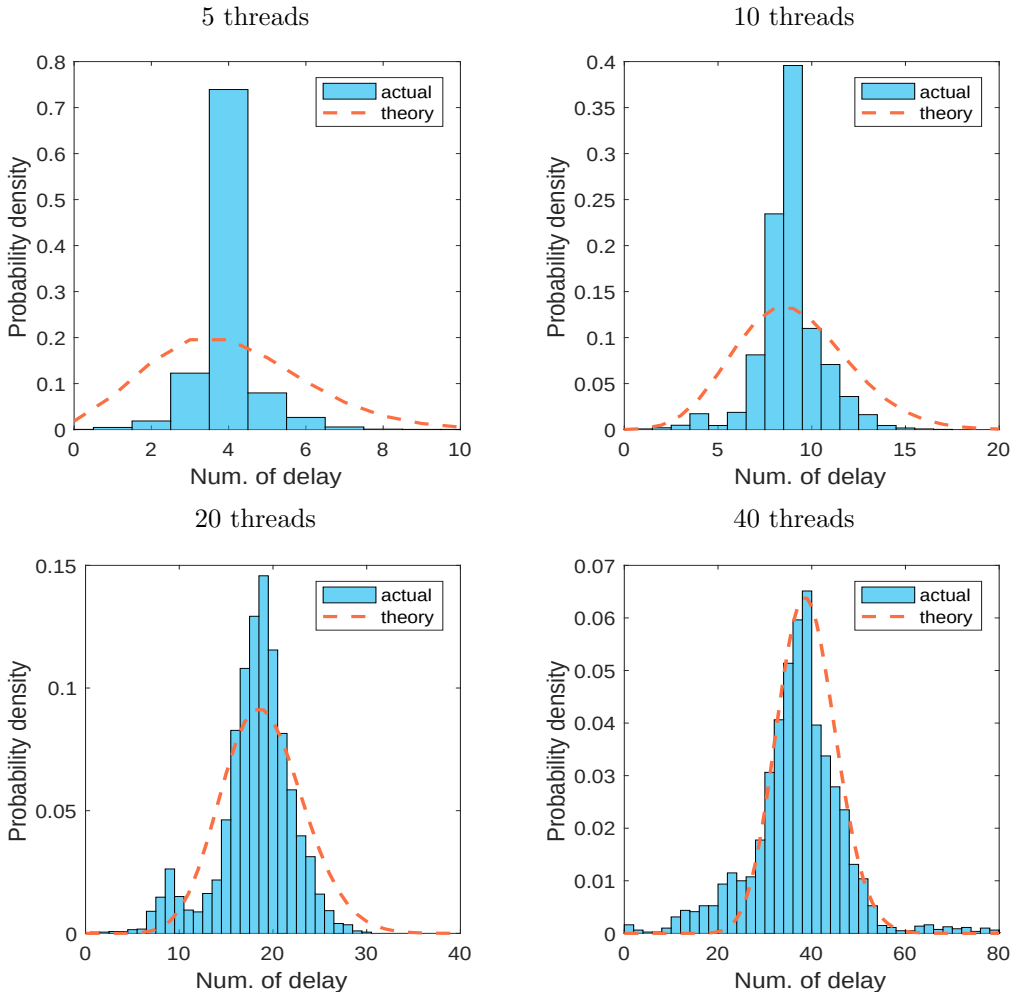

Figure 1: Delay distribution behaviors of Algorithm 1 for solving LASSO (32). The tested problem has $20,000$ coordinates, and it was running with 5, 10, 20, and 40 threads.

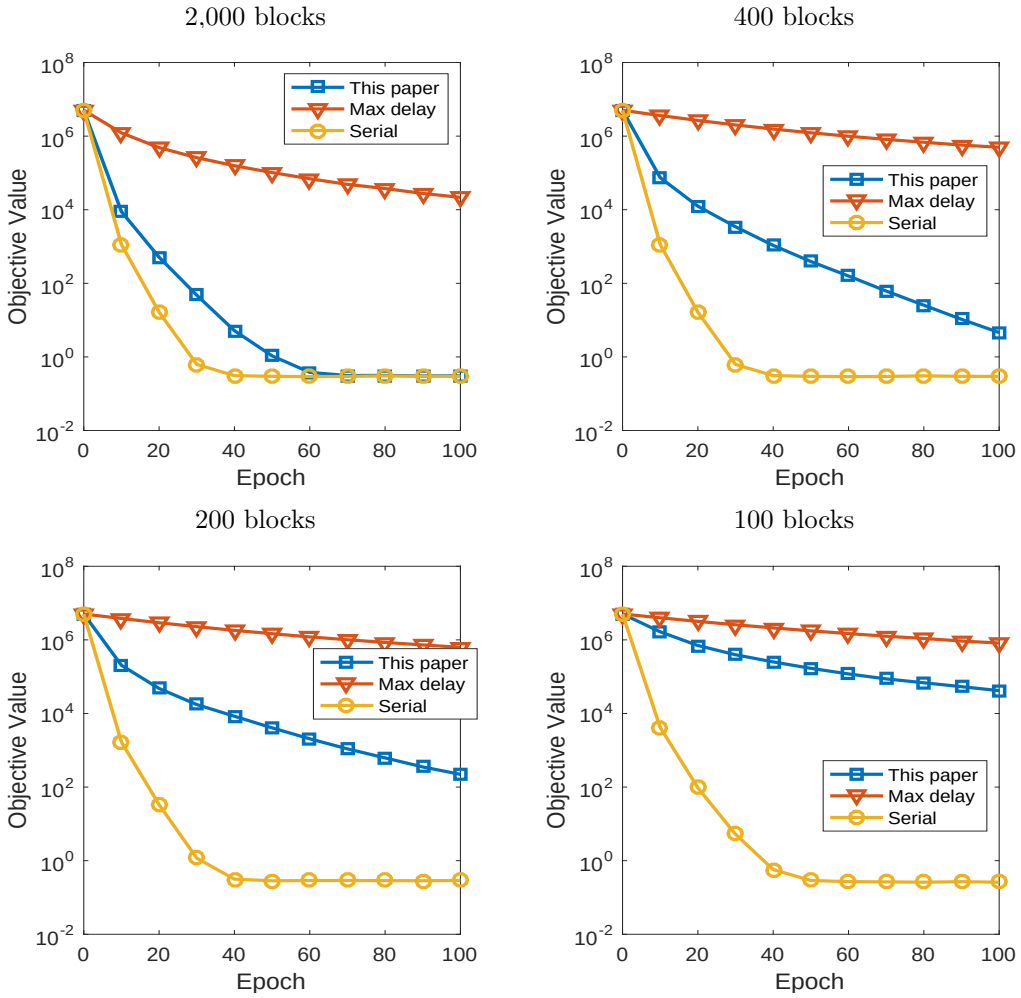

Figure 2: Convergence behaviors of Algorithm 1 for solving the LASSO problem (32) with the stepsize given in (31), and also the serial randomized coordinate descent method. The tested problem has $10,000$ samples and $20,000$ coordinates that are evenly partitioned into m blocks. It was simulated as running with 40 threads.

6.3 Nonnegative matrix factorization (NMF)

This section presents the numerical results of applying Algorithm 1 for solving the NMF problem Paatero and Tapper (1994)

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{Y}} \ \tfrac{1}{2}\|\mathbf{X}\mathbf{Y}^\top - \mathbf{Z}\|_F^2, \\ &\text{s.t. } \mathbf{X} \in \mathbb{R}_+^{M\times m}, \mathbf{Y} \in \mathbb{R}_+^{N\times m}, \end{aligned} \tag{33}$$

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 $m = 100$.

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 Xu and Yin (2014). 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 (31a), 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 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 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.

Acknowledgments

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

References

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Dimitri P Bertsekas. Distributed asynchronous computation of fixed points. *Mathematical Programming*, 27(1):107–120, 1983.

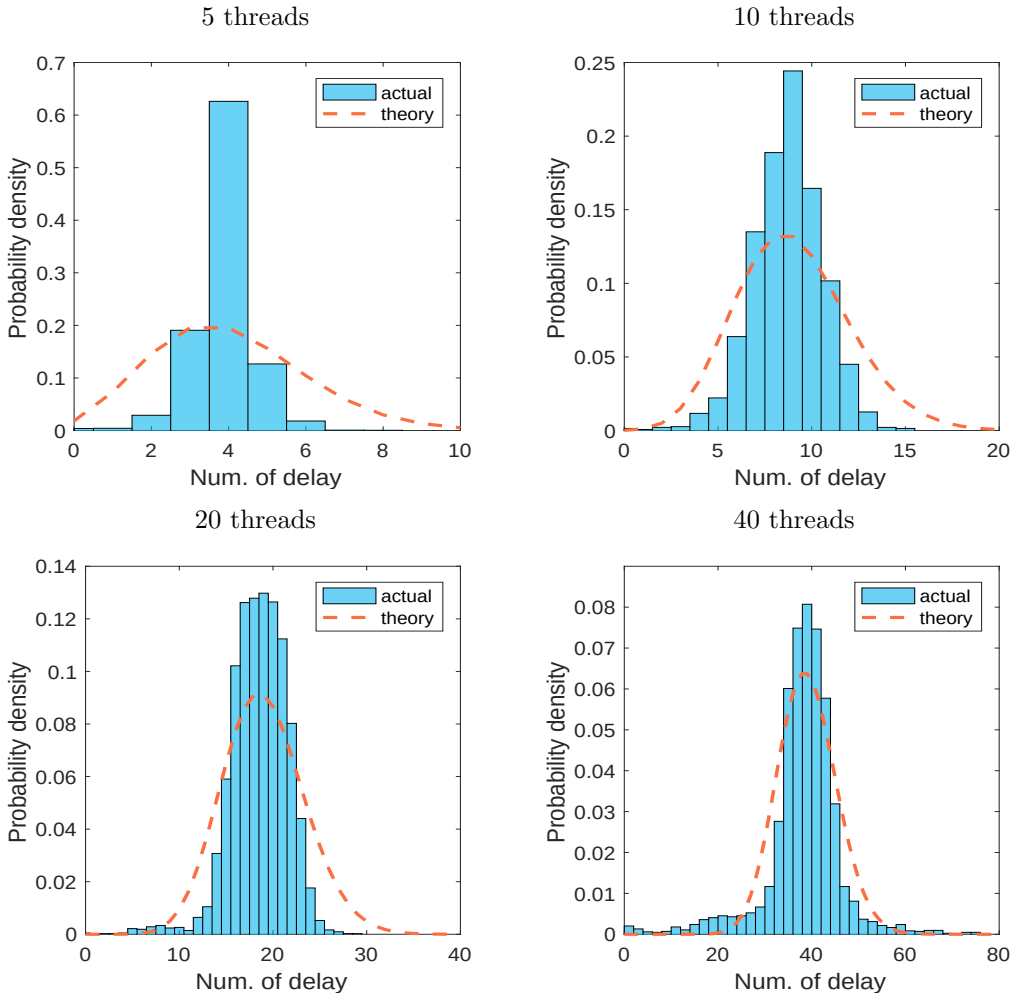

Figure 3: Delay distribution behaviors of Algorithm 1 for solving NMF (33). It was running with 5, 10, 20, and 40 threads.

Loris Cannelli, Francisco Facchinei, Vyacheslav Kungurtsev, and Gesualdo Scutari. Asynchronous parallel algorithms for nonconvex big-data optimization: Model and convergence. *arXiv preprint arXiv:1607.04818*, 2016.

Daniel Chazan and Willard Miranker. Chaotic relaxation. *Linear Algebra and its Applications*, 2 (2):199–222, 1969.

Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

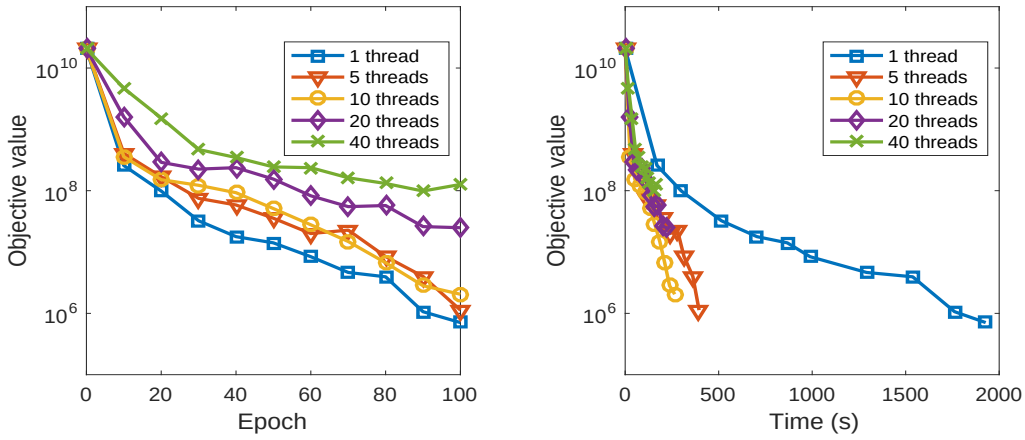

Figure 4: Convergence behaviors of Algorithm 1 for solving the NMF problem (33) with the stepsize set based on the expected delay. The size of the tested problem is $M = N = 10,000$ and $m = 100$, i.e., 200 block coordinates, and the algorithm was tested with 1, 5, 10, 20, and 40 threads.

Cong D. Dang and Guanghui Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015.

Damek Davis. The asynchronous PALM algorithm for nonsmooth nonconvex problems. *arXiv preprint arXiv:1604.00526*, 2016.

Damek Davis, Brent Edmunds, and Madeleine Udell. The sound of APALM clapping: Faster nonsmooth nonconvex optimization with stochastic asynchronous PALM. In *NIPS: Proceedings of Neural Information Processing Systems*, 2016.

Andreas Frommer and Daniel B Szyld. On asynchronous iterations. *Journal of Computational and Applied Mathematics*, 123(1):201–216, 2000.

L. Grippo and M. Sciandrone. On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Operations Research Letters*, 26(3):127–136, 2000.

Allan Gut. *Probability: A Graduate Course: A Graduate Course*. Springer Science & Business Media, 2006.

Robert Hannah and Wotao Yin. On unbounded delays in asynchronous parallel fixed-point algorithms. *arXiv preprint arXiv:1609.04746*, 2016.

C. Hildreth. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85, 1957.

Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn, and Zhi-Quan Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, in press, 2016.

Ji Liu and Stephen J Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

Ji Liu, Steve Wright, Christopher Re, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 469–477, 2014.

Zhaosong Lu and Lin Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015.

Zhi-Quan Luo and Paul Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

Aryan Mokhtai, Alec Koppel, and Alejandro Ribeiro. A class of parallel doubly stochastic algorithms for large-scale learning. *arXiv preprint arXiv:1606.04991*, 2016.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

Zhimin Peng, Yangyang Xu, Ming Yan, and Wotao Yin. Arock: An algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing*, 38(5):A2851–A2879, 2016.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

Jack L Rosenfeld. A case study in programming for parallel-processors. *Communications of the ACM*, 12(12):645–655, 1969.

John C. Strikwerda. A probabilistic analysis of asynchronous iteration. *Linear Algebra and its Applications*, 349(13):125 – 154, 2002.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

Paul Tseng, Dimitri P Bertsekas, and John N Tsitsiklis. Partially asynchronous, parallel algorithms for network flow and other problems. *SIAM Journal on Control and Optimization*, 28(3):678–710, 1990.

WhiteHouse. *Big Data: Seizing Opportunities Preserving Values*. 2014.

Yangyang Xu and Wotao Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *arXiv preprint arXiv:1410.1386*, 2014.

Yangyang Xu and Wotao Yin. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM Journal on Optimization*, 25(3):1686–1716, 2015.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

Throughout our analysis, we let $\mathbb{E}_{i_k,j_k}$ and $\mathbb{E}_{j_k}$ denote the conditional expectation about (i_k, j_k) and j_k, respectively, with respect to previous history $\{\mathbf{x}^0, \mathbf{x}^1, \cdots, \mathbf{x}^k\}$. Let $c_t := \sum_{d=t}^{\infty} q_d$, where $q_d = \text{Prob}(\mathbf{j} = d)$.

Appendix A. Technical details

The following result will be used several times, and it can be verified straightforward.

Lemma 19 *For any number sequence $\{a_{i,j}\}$, it holds that*

$$\sum_{t=1}^{k-1}\sum_{d=k-t}^{k-1} a_{d,t} = \sum_{d=1}^{k-1}\sum_{t=k-d}^{k-1} a_{d,t},\ \forall k \geq 0, \tag{34}$$

$$\sum_{t=1}^{k}\sum_{d=0}^{t-1} a_{d,t} = \sum_{d=0}^{k-1}\sum_{t=d+1}^{k} a_{d,t},\ \forall k \geq 0. \tag{35}$$

A.1 Proof of Theorem 3

Proof It is easy to see that

$$T := \mathbb{E}[\mathbf{j}] = \sum_{t=1}^{\infty} t q_t = \sum_{t=1}^{\infty}\sum_{d=1}^{t} q_t = \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} q_t = \sum_{d=1}^{\infty} c_d. \tag{36}$$

For any integers k, we have (by regarding $\mathbf{x}^k = \mathbf{x}^0$ if $k < 0$)

$$\begin{aligned}
\mathbb{E}_{i_k,j_k} f(\mathbf{x}^{k+1}) =& \mathbb{E}_{i_k,j_k} f(\mathbf{x}^k - \eta U_{i_k}\nabla f(\mathbf{x}^{k-j_k})) \\
\overset{(7)}{\leq}& f(\mathbf{x}^k) - \eta\mathbb{E}_{i_k,j_k}\langle \nabla f(\mathbf{x}^k), U_{i_k}\nabla f(\mathbf{x}^{k-j_k})\rangle + \frac{L_c}{2}\mathbb{E}_{i_k,j_k}\|\eta U_{i_k}\nabla f(\mathbf{x}^{k-j_k})\|^2 \\
=& f(\mathbf{x}^k) - \frac{\eta}{m}\mathbb{E}_{j_k}\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-j_k})\rangle + \frac{\eta^2 L_c}{2m}\mathbb{E}_{j_k}\|\nabla f(\mathbf{x}^{k-j_k})\|^2 \\
=& f(\mathbf{x}^k) - \frac{\eta}{m}\sum_{t=0}^{k-1} q_t\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t})\rangle - \frac{\eta}{m}c_k\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0)\rangle \\
&+ \frac{\eta^2 L_c}{2m}\sum_{t=0}^{k-1} q_t\|\nabla f(\mathbf{x}^{k-t})\|^2 + \frac{\eta^2 L_c}{2m}c_k\|\nabla f(\mathbf{x}^0)\|^2 \qquad (37) \\
=& f(\mathbf{x}^k) - \frac{\eta}{m}\sum_{t=0}^{k-1} q_t\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t}), \nabla f(\mathbf{x}^{k-t})\rangle - \frac{\eta}{m}c_k\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0)\rangle \\
&- \left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right)\sum_{t=0}^{k-1} q_t\|\nabla f(\mathbf{x}^{k-t})\|^2 + \frac{\eta^2 L_c}{2m}c_k\|\nabla f(\mathbf{x}^0)\|^2. \qquad (38)
\end{aligned}$$

When $k=0$ and $k=1$, we have

$$\mathbb{E}f(\mathbf{x}^1) \leq f(\mathbf{x}^0) - \frac{\eta}{m}\|\nabla f(\mathbf{x}^0)\|^2 + \frac{\eta^2 L_c}{2m}\|\nabla f(\mathbf{x}^0)\|^2, \tag{39}$$

$$\begin{aligned}\mathbb{E}f(\mathbf{x}^2) \leq& \mathbb{E}f(\mathbf{x}^1) - \frac{\eta}{m}c_1\mathbb{E}\langle\nabla f(\mathbf{x}^1), \nabla f(\mathbf{x}^0)\rangle - \left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right) q_0 \mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2 + \frac{\eta^2 L_c}{2m}c_1\|\nabla f(\mathbf{x}^0)\|^2 \\ \leq& \mathbb{E}f(\mathbf{x}^1) - \left[\left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right) q_0 - \frac{\eta}{2m}c_1\right]\mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2 + \left(\frac{\eta}{2m}c_1 + \frac{\eta^2 L_c}{2m}c_1\right)\|\nabla f(\mathbf{x}^0)\|^2.\end{aligned} \tag{40}$$

Next, we are going to find the inequalities for $k \geq 2$. Note that

$$\begin{aligned}&\mathbb{E}[-\langle\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t}), \nabla f(\mathbf{x}^{k-t})\rangle] \\ \leq& \mathbb{E}\|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t})\| \cdot \|\nabla f(\mathbf{x}^{k-t})\| \\ \leq& \sum_{d=k-t}^{k-1} \mathbb{E}\|\nabla f(\mathbf{x}^{d+1}) - \nabla f(\mathbf{x}^d)\| \cdot \|\nabla f(\mathbf{x}^{k-t})\| \\ \leq& L_r \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{x}^{d+1} - \mathbf{x}^d\| \cdot \|\nabla f(\mathbf{x}^{k-t})\| \\ =& \eta L_r \sum_{d=k-t}^{k-1} \mathbb{E}\|\nabla_{i_d} f(\mathbf{x}^{d-j_d})\| \cdot \|\nabla f(\mathbf{x}^{k-t})\| \\ \leq& \frac{\eta L_r}{2} \sum_{d=k-t}^{k-1} \left(\mathbb{E}\sqrt{m}\|\nabla_{i_d} f(\mathbf{x}^{d-j_d})\|^2 + \frac{1}{\sqrt{m}}\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2\right) \\ =& \frac{\eta L_r}{2\sqrt{m}} \sum_{d=k-t}^{k-1} \left(\mathbb{E}\|\nabla f(\mathbf{x}^{d-j_d})\|^2 + \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2\right) \\ =& \frac{\eta L_r}{2\sqrt{m}} \left(\sum_{d=k-t}^{k-1} \left(\sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d\|\nabla f(\mathbf{x}^0)\|^2\right) + t\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2\right),\end{aligned} \tag{41}$$

where in the third inequality, we have used the Lipschitz continuity of ∇f and $\mathbf{x}^{d+1}$ differs from $\mathbf{x}^d$ only at one block of coordinates. Substituting (41) into (38) and taking expectation gives

$$\begin{aligned}\mathbb{E}f(\mathbf{x}^{k+1}) \leq& \mathbb{E}f(\mathbf{x}^k) + \frac{\eta^2 L_r}{2m\sqrt{m}} \sum_{t=1}^{k-1} t q_t \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 \\ &+ \frac{\eta^2 L_r}{2m\sqrt{m}} \sum_{t=1}^{k-1} q_t \sum_{d=k-t}^{k-1} \left(\sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d\|\nabla f(\mathbf{x}^0)\|^2\right) - \frac{\eta}{m}c_k\mathbb{E}\langle\nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0)\rangle \\ &- \left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right) \sum_{t=0}^{k-1} q_t \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 + \frac{\eta^2 L_c}{2m}c_k\mathbb{E}\|\nabla f(\mathbf{x}^0)\|^2.\end{aligned} \tag{42}$$

Note that

$$\sum_{t=1}^{k-1} q_t \sum_{d=k-t}^{k-1} c_d \|\nabla f(\mathbf{x}^0)\|^2 \overset{(34)}{=} \sum_{d=1}^{k-1} \left(\sum_{t=k-d}^{k-1} q_t \right) c_d \|\nabla f(\mathbf{x}^0)\|^2 = \sum_{d=1}^{k-1} (c_{k-d} - c_k) c_d \|\nabla f(\mathbf{x}^0)\|^2, \tag{43}$$

and

$$\begin{aligned}
\sum_{t=1}^{k-1} q_t \sum_{d=k-t}^{k-1} \sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 &= \sum_{d=1}^{k-1} (c_{k-d} - c_k) \sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 \\
[\text{let } r \leftarrow d-r] \quad &= \sum_{d=1}^{k-1} (c_{k-d} - c_k) \sum_{r=1}^{d} q_{d-r} \mathbb{E}\|\nabla f(\mathbf{x}^{r})\|^2 \\
&\overset{(35)}{=} \sum_{r=1}^{k-1} \left(\sum_{d=r}^{k-1} (c_{k-d} - c_k) q_{d-r} \right) \mathbb{E}\|\nabla f(\mathbf{x}^{r})\|^2 \\
[\text{let } t \leftarrow k-r,\ d \leftarrow k-d] \quad &= \sum_{t=1}^{k-1} \left(\sum_{d=1}^{t} (c_d - c_k) q_{t-d} \right) \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2.
\end{aligned} \tag{44}$$

In addition,

$$-c_k \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0) \rangle \le \frac{c_k}{2} \|\nabla f(\mathbf{x}^k)\|^2 + \frac{c_k}{2} \|\nabla f(\mathbf{x}^0)\|^2. \tag{45}$$

Let

$$\gamma_k = \frac{\eta^2 L_r}{2m\sqrt{m}} \sum_{d=1}^{k-1} (c_{k-d} - c_k) c_d + \frac{\eta}{2m} c_k + \frac{\eta^2 L_c}{2m} c_k, \tag{46a}$$

$$\beta_k = \left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m} \right) q_0 - \frac{\eta}{2m} c_k \text{ for } k \ge 1, \quad (\text{and } \beta_0 = 0)\,, \tag{46b}$$

$$C_{t,k} = \left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m} \right) q_t - \frac{\eta^2 L_r}{2m\sqrt{m}} t q_t - \frac{\eta^2 L_r}{2m\sqrt{m}} \sum_{d=1}^{t} (c_d - c_k) q_{t-d}. \tag{46c}$$

Then substituting (43) through (45) into (42) and combining terms, we have

$$\mathbb{E} f(\mathbf{x}^{k+1}) \le \mathbb{E} f(\mathbf{x}^k) + \gamma_k \|\nabla f(\mathbf{x}^0)\|^2 - \beta_k \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \sum_{t=1}^{k-1} C_{t,k} \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2. \tag{47}$$

For any integer $K \geq 1$, it holds that

$$\begin{aligned}
&\sum_{k=0}^{K}\sum_{t=1}^{k-1} C_{t,k}\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 \\
[\text{let } t \leftarrow k-t] \quad &= \sum_{k=1}^{K}\sum_{t=1}^{k-1} C_{k-t,k}\mathbb{E}\|\nabla f(\mathbf{x}^{t})\|^2 \\
&\overset{(35)}{=} \sum_{t=1}^{K-1}\sum_{k=t+1}^{K} C_{k-t,k}\mathbb{E}\|\nabla f(\mathbf{x}^{t})\|^2 \\
[k \leftrightarrow t] \quad &= \sum_{k=1}^{K-1}\sum_{t=k+1}^{K} C_{t-k,t}\mathbb{E}\|\nabla f(\mathbf{x}^{k})\|^2.
\end{aligned}$$

Summing up (47) from $k=0$ through K and using the above equality, we have

$$\mathbb{E}f(\mathbf{x}^{K+1}) \leq f(\mathbf{x}^0) + \sum_{k=0}^{K}\gamma_k\|\nabla f(\mathbf{x}^0)\|^2 - \beta_K\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 - \sum_{k=1}^{K-1}\left(\beta_k + \sum_{t=k+1}^{K} C_{t-k,t}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2. \tag{48}$$

Note that

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

Hence,

$$\sum_{k=0}^{K}\gamma_k \leq \left(\frac{\eta^2 L_r}{2m\sqrt{m}}\right)T^2 + \left(\frac{\eta}{2m} + \frac{\eta^2 L_c}{2m}\right)(1+T).$$

In addition,

$$\sum_{t=1}^{\infty}\sum_{d=1}^{t}(c_d - c_{k+t})q_{t-d} \leq \sum_{t=1}^{\infty}\sum_{d=1}^{t} c_d q_{t-d} = \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} c_d q_{t-d} = \sum_{d=1}^{\infty} c_d \overset{(36)}{=} T, \tag{49}$$

and thus

$$\begin{aligned}
&\beta_k + \sum_{t=k+1}^{\infty} C_{t-k,t} = \beta_k + \sum_{t=1}^{\infty} C_{t,t+k} \\
=&(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m})q_0 - \frac{\eta}{2m}c_k + \sum_{t=1}^{\infty}\left(\left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right)q_t - \frac{\eta^2 L_r}{2m\sqrt{m}}tq_t - \frac{\eta^2 L_r}{2m\sqrt{m}}\sum_{d=1}^{t}(c_d - c_{k+t})q_{t-d}\right) \\
\overset{(49)}{\geq}& \left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right) - \frac{\eta}{2m}c_k - \frac{\eta^2 L_r T}{m\sqrt{m}} \geq \left(\frac{\eta}{2m} - \frac{\eta^2 L_c}{2m}\right) - \frac{\eta^2 L_r T}{m\sqrt{m}},
\end{aligned}$$

where the last inequality follows from $c_k \leq 1$. Letting $K \to \infty$ in (48) and using the lower boundedness of f, we have

$$\lim_{K\to\infty}\beta_K\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 + \sum_{k=1}^{\infty}\left(\frac{\eta}{2m} - \frac{\eta^2 L_c}{2m} - \frac{\eta^2 L_r T}{m\sqrt{m}}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 < \infty.$$

Since $\eta < \frac{1/L_c}{1+2\kappa T/\sqrt{m}}$, we have (11) 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 (Gut, 2006, 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. ■

A.2 Proof of Lemma 5

Proof Following the proof of Theorem 1 in Liu et al. (2014), we have

$$\begin{aligned}
&\mathbb{E}\big[\|\nabla f(\mathbf{x}^t)\|^2 - \|\nabla f(\mathbf{x}^{t+1})\|^2\big]\\
\leq& 2\mathbb{E}\big[\|\nabla f(\mathbf{x}^t)\| \cdot \|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^{t+1})\|\big] \quad (\text{from } \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2 \leq 2\|\mathbf{u}\| \cdot \|\mathbf{u}-\mathbf{v}\|)\\
\leq& 2L_r\mathbb{E}\big[\|\nabla f(\mathbf{x}^t)\| \cdot \|\mathbf{x}^t - \mathbf{x}^{t+1}\|\big]\\
=& 2\eta L_r\mathbb{E}\big[\|\nabla f(\mathbf{x}^t)\| \cdot \|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|\big]\\
\leq& \eta L_r\left(\frac{1}{\sqrt{m}}\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \sqrt{m}\mathbb{E}\|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|^2\right)\\
=& \frac{\eta L_r}{\sqrt{m}}\left(\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \mathbb{E}\|\nabla f(\mathbf{x}^{t-j_t})\|^2\right)\\
=& \frac{\eta L_r}{\sqrt{m}}\left(\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \sum_{r=0}^{t-1} q_r\mathbb{E}\|\nabla f(\mathbf{x}^{t-r})\|^2 + c_t\|\nabla f(\mathbf{x}^0)\|^2\right)
\end{aligned} \tag{50}$$

and

$$\begin{aligned}
&\mathbb{E}\big[\|\nabla f(\mathbf{x}^{t+1})\|^2 - \|\nabla f(\mathbf{x}^t)\|^2\big]\\
\leq& \mathbb{E}\big[\|\nabla f(\mathbf{x}^{t+1}) + \nabla f(\mathbf{x}^t)\| \cdot \|\nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t)\|\big]\\
\leq& L_r\mathbb{E}\left[\big(2\|\nabla f(\mathbf{x}^t)\| + \|\nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t)\|\big)\|\mathbf{x}^{t+1} - \mathbf{x}^t\|\right]\\
\leq& L_r\mathbb{E}\left[2\|\nabla f(\mathbf{x}^t)\| \cdot \|\mathbf{x}^{t+1} - \mathbf{x}^t\| + L_r\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2\right]\\
=& L_r\mathbb{E}\left[2\eta\|\nabla f(\mathbf{x}^t)\| \cdot \|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\| + \eta^2 L_r\|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|^2\right]\\
\leq& L_r\mathbb{E}\left[\frac{\eta}{\sqrt{m}}\|\nabla f(\mathbf{x}^t)\|^2 + \eta\sqrt{m}\|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|^2 + \eta^2 L_r\|U_{i_t}\nabla f(\mathbf{x}^{t-j_t})\|^2\right]\\
=& \frac{\eta L_r}{\sqrt{m}}\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \left(\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)\mathbb{E}\|\nabla f(\mathbf{x}^{t-j_t})\|^2\\
=& \frac{\eta L_r}{\sqrt{m}}\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 + \left(\frac{\eta L_r}{\sqrt{m}} + \frac{\eta^2 L_r^2}{m}\right)\left(\sum_{r=0}^{t-1} q_r\mathbb{E}\|\nabla f(\mathbf{x}^{t-r})\|^2 + c_t\|\nabla f(\mathbf{x}^0)\|^2\right).
\end{aligned} \tag{51}$$

We first show the first inequality in (14). Note that (13) gives us

$$\frac{1}{1-(1+M_\rho)\frac{\eta L_r}{\sqrt{m}}} \leq \rho. \tag{52}$$

When $t=0$, we have from (50) that

$$\|\nabla f(\mathbf{x}^0)\|^2 - \mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2 \le \frac{2\eta L_r}{\sqrt{m}}\|\nabla f(\mathbf{x}^0)\|^2 \le (1+M_\rho)\frac{\eta L_r}{\sqrt{m}}\|\nabla f(\mathbf{x}^0)\|^2.$$

Hence, $\|\nabla f(\mathbf{x}^0)\|^2 \le \rho\mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2$ from (52). Now assume $\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 \le \rho\mathbb{E}\|\nabla f(\mathbf{x}^{t+1})\|^2$ for all $t \le k-1$. For $t=k$, it holds from (50) and the induction assumption that

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

Hence, we have $\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \le \rho\mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2$ from (52). Therefore, we finish the induction step, and thus the first inequality of (14) holds.

Next we show the second inequality of (14). It is easy to verify that (13) implies $\eta \le \frac{\rho-1}{\frac{L_r}{\sqrt{m}}\left(1+M_\rho+\frac{(\rho-1)M_\rho}{\rho(1+M_\rho)}\right)}$ and thus

$$\begin{aligned}1+\frac{\eta L_r}{\sqrt{m}}+\left(\frac{\eta L_r}{\sqrt{m}}+\frac{\eta^2 L_r^2}{m}\right)M_\rho \overset{(13)}{\le}& 1+\frac{\eta L_r}{\sqrt{m}}(1+M_\rho)+M_\rho\frac{\eta L_r^2}{m}\frac{(\rho-1)\sqrt{m}}{\rho L_r(1+M_\rho)}\\ =&1+\frac{\eta L_r}{\sqrt{m}}\left(1+M_\rho+\frac{(\rho-1)M_\rho}{\rho(1+M_\rho)}\right)\le\rho.\end{aligned}\tag{53}$$

When $t=0$, we have from (51) that

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

Hence, $\mathbb{E}\|\nabla f(\mathbf{x}^1)\|^2 \le \rho\|\nabla f(\mathbf{x}^0)\|^2$ holds from (53). Assume $\mathbb{E}\|\nabla f(\mathbf{x}^{t+1})\|^2 \le \rho\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2$ for all $t \le k-1$. It follows from (51) and the induction assumption that

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

Hence, from (53), $\mathbb{E}\|\nabla f(\mathbf{x}^{k+1})\|^2 \le \rho\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2$ holds, and we complete the proof. ■

A.3 Proof of Theorem 6

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 (12), and it is easy to see $T < \infty$. Also note that

$$\mathbb{E}[\mathbf{j}\rho^{\mathbf{j}}] = \sum_{t=1}^{\infty} t q_t \rho^t = \sum_{t=1}^{\infty}\sum_{d=1}^{t} q_t\rho^t = \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} q_t\rho^t \geq \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} q_t\rho^d = \sum_{d=1}^{\infty} c_d\rho^d. \tag{54}$$

Taking expectation on both sides of (37) gives

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{k+1}) \leq& \mathbb{E}f(\mathbf{x}^k) - \frac{\eta}{m}\sum_{t=0}^{k-1} q_t \mathbb{E}\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t})\rangle - \frac{\eta}{m} c_k \mathbb{E}\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0)\rangle \\
&+ \frac{\eta^2 L_c}{2m}\sum_{t=0}^{k-1} q_t \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 + \frac{\eta^2 L_c}{2m} c_k \|\nabla f(\mathbf{x}^0)\|^2 \\
=& \mathbb{E}f(\mathbf{x}^k) - \frac{\eta}{m}\sum_{t=0}^{k-1} q_t \mathbb{E}\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t}) - \nabla f(\mathbf{x}^k)\rangle - \frac{\eta}{m} c_k \mathbb{E}\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0) - \nabla f(\mathbf{x}^k)\rangle \\
&- \frac{\eta}{m}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta^2 L_c}{2m}\sum_{t=0}^{k-1} q_t \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 + \frac{\eta^2 L_c}{2m} c_k \|\nabla f(\mathbf{x}^0)\|^2.
\end{aligned} \tag{55}$$

Through the same arguments of showing (41), we have for any $t \leq k$,

$$\begin{aligned}
&\mathbb{E}[-\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k-t}) - \nabla f(\mathbf{x}^k)\rangle] \\
\leq& \frac{\eta L_r}{2\sqrt{m}}\left(\sum_{d=k-t}^{k-1}\left(\sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d\|\nabla f(\mathbf{x}^0)\|^2\right) + t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2\right).
\end{aligned} \tag{56}$$

Substituting (56) into (55) gives

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}^{k+1}) \leq& \mathbb{E}f(\mathbf{x}^k) + \frac{\eta^2 L_r}{2m\sqrt{m}}\sum_{t=0}^{k-1} q_t\left(\sum_{d=k-t}^{k-1}\left(\sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d\|\nabla f(\mathbf{x}^0)\|^2\right) + t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2\right) \\
&+ \frac{\eta^2 L_r}{2m\sqrt{m}} c_k\left(\sum_{d=0}^{k-1}\left(\sum_{r=0}^{d-1} q_r \mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2 + c_d\|\nabla f(\mathbf{x}^0)\|^2\right) + k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2\right) \\
&- \frac{\eta}{m}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta^2 L_c}{2m}\sum_{t=0}^{k-1} q_t \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 + \frac{\eta^2 L_c}{2m} c_k \|\nabla f(\mathbf{x}^0)\|^2.
\end{aligned} \tag{57}$$

From (43), it follows that

$$\sum_{t=0}^{k-1} q_t \sum_{d=k-t}^{k-1} c_d + c_k \sum_{d=0}^{k-1} c_d = \sum_{d=1}^{k-1}(c_{k-d} - c_k)c_d + c_k\sum_{d=0}^{k-1} c_d = \sum_{d=1}^{k-1} c_{k-d}c_d + c_k. \tag{58}$$

Following the arguments for showing (44), we have

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

Substituting the above equalities and also (44) into (57) gives

$$\begin{aligned}\mathbb{E}f(\mathbf{x}^{k+1}) \leq & \mathbb{E}f(\mathbf{x}^k) + \frac{\eta^2 L_r}{2m\sqrt{m}}\left(\sum_{d=1}^{k-1} c_{k-d}c_d + c_k\right)\|\nabla f(\mathbf{x}^0)\|^2 + \frac{\eta^2 L_r}{2m\sqrt{m}}\sum_{t=1}^{k-1}\left(\sum_{d=1}^{t} c_d q_{t-d}\right)\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 \\ & + \frac{\eta^2 L_r}{2m\sqrt{m}}\left(\sum_{t=0}^{k-1} t q_t + k c_k\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\ & - \frac{\eta}{m}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta^2 L_c}{2m}\sum_{t=0}^{k-1} q_t \mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 + \frac{\eta^2 L_c}{2m} c_k \|\nabla f(\mathbf{x}^0)\|^2 \\ \overset{(14)}{\leq} & \mathbb{E}f(\mathbf{x}^k) + \frac{\eta^2 L_r}{2m\sqrt{m}}\left(\sum_{d=1}^{k-1} c_{k-d}c_d + c_k\right)\rho^k\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta^2 L_r}{2m\sqrt{m}}\sum_{t=1}^{k-1}\left(\sum_{d=1}^{t} c_d q_{t-d}\right)\rho^t\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\ & + \frac{\eta^2 L_r}{2m\sqrt{m}}\left(\sum_{t=0}^{k-1} t q_t + k c_k\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\ & - \frac{\eta}{m}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta^2 L_c}{2m}\sum_{t=0}^{k-1} q_t \rho^t \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta^2 L_c}{2m} c_k \rho^k \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2. \quad (59)\end{aligned}$$

Note that

$$\sum_{t=1}^{k-1}\left(\sum_{d=1}^{t} c_d q_{t-d}\right)\rho^t \leq \sum_{t=1}^{\infty}\left(\sum_{d=1}^{t} c_d q_{t-d}\right)\rho^t = \sum_{d=1}^{\infty} c_d \rho^d \sum_{t=d}^{\infty} q_{t-d}\rho^{t-d} \overset{(54)}{\leq} N_\rho M_\rho,$$

and

$$\left(\sum_{d=1}^{k-1} c_{k-d}c_d + c_k\right)\rho^k = \sum_{d=1}^{k} c_d c_{k-d}\rho^k = \sum_{d=1}^{k} c_d\rho^d c_{k-d}\rho^{k-d} \leq \sum_{d=1}^{k} c_d\rho^d\left(\sum_{r=0}^{\infty} q_r \rho^r\right) \leq N_\rho M_\rho.$$

Hence, (59) together with the above two inequalities implies

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

which completes the proof. ■

A.4 Proof of Theorem 7

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

$$\|\nabla f(\mathbf{x}^k)\|^2 \geq \frac{(f(\mathbf{x}^k) - f^*)^2}{B^2}. \tag{60}$$

Substituting (60) into (18) yields

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

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 (19) holds.

If f is strongly convex with constant μ, then

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

and we immediately have (20) from (18) and the above inequality. This completes the proof. ■

A.5 Proof of Lemma 9

Proof First note that from the definition of $\bar{\mathbf{x}}^{k+1}$, we have

$$-\nabla f(\mathbf{x}^{k-j_k}) - \frac{1}{\eta}(\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k) \in \partial R(\bar{\mathbf{x}}^{k+1}), \tag{61}$$

which together with the convexity of R implies that for any $\mathbf{x}$,

$$R(\bar{\mathbf{x}}^{k+1}) - R(\mathbf{x}) \leq -\langle \nabla f(\mathbf{x}^{k-j_k}) + \frac{1}{\eta}(\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k), \bar{\mathbf{x}}^{k+1} - \mathbf{x}\rangle. \tag{62}$$

We have

$$\begin{aligned}
\mathbb{E}_{i_k} F(\mathbf{x}^{k+1}) =& \mathbb{E}_{i_k}\left[f(\mathbf{x}^k + U_{i_k}(\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k)) + R(\mathbf{x}^{k+1})\right] \\
\leq& \mathbb{E}_{i_k}\left[f(\mathbf{x}^k) + \langle \nabla_{i_k} f(\mathbf{x}^k), \bar{\mathbf{x}}_{i_k}^{k+1} - \mathbf{x}_{i_k}^k\rangle + \frac{L_c}{2}\|\bar{\mathbf{x}}_{i_k}^{k+1} - \mathbf{x}_{i_k}^k\|^2 + R(\mathbf{x}^{k+1})\right] \\
=& f(\mathbf{x}^k) + \frac{m-1}{m} R(\mathbf{x}^k) + \frac{1}{m}\left(\langle \nabla f(\mathbf{x}^k), \bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\rangle + \frac{L_c}{2}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + R(\bar{\mathbf{x}}^{k+1})\right) \\
=& F(\mathbf{x}^k) + \frac{1}{m}\left(\langle \nabla f(\mathbf{x}^k), \bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\rangle + \frac{L_c}{2}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + R(\bar{\mathbf{x}}^{k+1}) - R(\mathbf{x}^k)\right) \\
=& F(\mathbf{x}^k) + \frac{1}{m}\left(\langle \nabla f(\mathbf{x}^{k-j_k}), \bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\rangle + \frac{L_c}{2}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + R(\bar{\mathbf{x}}^{k+1}) - R(\mathbf{x}^k)\right) \\
&+ \frac{1}{m}\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-j_k}), \bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\rangle \\
\overset{(62)}{\leq}& F(\mathbf{x}^k) + \frac{1}{m}\left(\frac{L_c}{2} - \frac{1}{\eta}\right)\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 \\
&+ \frac{1}{m}\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-j_k}), \bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\rangle. \qquad (63)
\end{aligned}$$

In addition,

$$\begin{aligned}
&\mathbb{E}\langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-j_k}), \bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\rangle \\
\leq& \mathbb{E}\|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-j_k})\| \cdot \|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\| \\
\leq& \mathbb{E}\sum_{d=k-j_k}^{k-1} \|\nabla f(\mathbf{x}^{d+1}) - \nabla f(\mathbf{x}^d)\| \cdot \|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\| \\
\leq& L_r \mathbb{E}\sum_{d=k-j_k}^{k-1} \|\mathbf{x}^{d+1} - \mathbf{x}^d\| \cdot \|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\| \\
\leq& \frac{L_r}{2}\left(\frac{1}{\kappa}\mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + \kappa\mathbb{E}\left[j_k \sum_{d=k-j_k}^{k-1} \|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2\right]\right) \\
=& \frac{L_r}{2}\left(\frac{1}{\kappa}\mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + \kappa\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2 + \kappa\sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{x}^{d+1} - \mathbf{x}^d\|^2\right) \\
=& \frac{L_r}{2}\left(\frac{1}{\kappa}\mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + \frac{\kappa}{m}\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\bar{\mathbf{x}}^{d+1} - \mathbf{x}^d\|^2 + \frac{\kappa}{m}\sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\bar{\mathbf{x}}^{d+1} - \mathbf{x}^d\|^2\right) \\
\leq& \frac{L_r}{2}\left[\frac{1}{\kappa}\mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|^2 + \frac{\kappa}{m}\left(\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\bar{\mathbf{x}}^{d+1} - \mathbf{x}^d\|^2 + \sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\bar{\mathbf{x}}^{d+1} - \mathbf{x}^d\|^2\right)\right]. \qquad (64)
\end{aligned}$$

Taking expectation on both sides of (63) and substituting (64) yield

$$\begin{aligned}&\mathbb{E}[F(\mathbf{x}^{k+1})-F(\mathbf{x}^k)]+\frac{1}{m}\left(\frac{1}{\eta}-L_c\right)\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2\\ \le&\frac{\kappa L_r}{2m^2}\left(\sum_{t=1}^{k-1}q_t t\sum_{d=k-t}^{k-1}\mathbb{E}\|\bar{\mathbf{x}}^{d+1}-\mathbf{x}^d\|^2+\sum_{t=k}^{\infty}q_t t\sum_{d=0}^{k-1}\mathbb{E}\|\bar{\mathbf{x}}^{d+1}-\mathbf{x}^d\|^2\right).\end{aligned}\tag{65}$$

From Lemma 19, we have that for any $K\ge 0$,

$$\begin{aligned}\sum_{k=0}^{K}\sum_{t=1}^{k-1}q_t t\sum_{d=k-t}^{k-1}\mathbb{E}\|\bar{\mathbf{x}}^{d+1}-\mathbf{x}^d\|^2&\overset{(34)}{=}\sum_{k=0}^{K}\sum_{d=1}^{k-1}\left(\sum_{t=k-d}^{k-1}q_t t\right)\mathbb{E}\|\bar{\mathbf{x}}^{d+1}-\mathbf{x}^d\|^2\\ &\overset{(35)}{=}\sum_{d=1}^{K-1}\sum_{k=d+1}^{K}\left(\sum_{t=k-d}^{k-1}q_t t\right)\mathbb{E}\|\bar{\mathbf{x}}^{d+1}-\mathbf{x}^d\|^2\\ [k\leftrightarrow d]\quad&=\sum_{k=1}^{K-1}\left(\sum_{d=k+1}^{K}\sum_{t=d-k}^{d-1}q_t t\right)\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2,\end{aligned}\tag{66}$$

and

$$\begin{aligned}\sum_{k=0}^{K}\sum_{t=k}^{\infty}q_t t\sum_{d=0}^{k-1}\mathbb{E}\|\bar{\mathbf{x}}^{d+1}-\mathbf{x}^d\|^2&=\sum_{k=1}^{K}\sum_{d=0}^{k-1}\left(\sum_{t=k}^{\infty}q_t t\right)\mathbb{E}\|\bar{\mathbf{x}}^{d+1}-\mathbf{x}^d\|^2\\ &\overset{(35)}{=}\sum_{d=0}^{K-1}\sum_{k=d+1}^{K}\left(\sum_{t=k}^{\infty}q_t t\right)\mathbb{E}\|\bar{\mathbf{x}}^{d+1}-\mathbf{x}^d\|^2\\ [k\leftrightarrow d]\quad&=\sum_{k=0}^{K-1}\left(\sum_{d=k+1}^{K}\sum_{t=d}^{\infty}q_t t\right)\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2.\end{aligned}\tag{67}$$

Summing up (65) from $k=0$ through K and substituting (66) and (67), we have

$$\mathbb{E}[F(\mathbf{x}^{K+1})-F(\mathbf{x}^0)]+\frac{1}{m}\left(\frac{1}{\eta}-L_c\right)\sum_{k=0}^{K}\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2\le\frac{\kappa L_r}{2m^2}\sum_{k=0}^{K-1}\left(\sum_{d=k+1}^{K}\sum_{t=d-k}^{\infty}q_t t\right)\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2.\tag{68}$$

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 (22) from (68) by letting $K\to\infty$. ■

A.6 Proof of Lemma 10

Proof Let $\epsilon > 0$ be any positive number. From (22), there must exist an integer $J > 0$ such that

$$\sum_{d=J}^{\infty} \mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 \leq \frac{m\epsilon}{3\sum_{t=1}^{\infty} q_t t}. \tag{69}$$

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

$$\sum_{t=k-J}^{\infty} q_t t \leq \frac{m\epsilon}{3\sum_{d=0}^{\infty} \mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2}. \tag{70}$$

Hence, for any $k \geq K$,

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2 \\
\leq&\mathbb{E}\left(j_k \sum_{d=k-j_k}^{k-1} \|\mathbf{x}^d - \mathbf{x}^{d+1}\|^2\right) \\
=&\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|^2 + \sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|^2 \\
=&\frac{1}{m}\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 + \frac{1}{m}\sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 \\
=&\frac{1}{m}\left(\sum_{t=1}^{k-1} q_t t \sum_{d=k-t}^{k-1} \mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 + \sum_{t=k}^{\infty} q_t t \sum_{d=0}^{k-1} \mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2\right) \qquad (71) \\
\overset{(34)}{=}&\frac{1}{m}\left[\sum_{d=1}^{k-1}\left(\sum_{t=k-d}^{k-1} q_t t\right)\mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 + \sum_{d=0}^{k-1}\left(\sum_{t=k}^{\infty} q_t t\right)\mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2\right] \\
=&\frac{1}{m}\left[\sum_{d=1}^{J}\left(\sum_{t=k-d}^{k-1} q_t t\right)\mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 + \sum_{d=J+1}^{k-1}\left(\sum_{t=k-d}^{k-1} q_t t\right)\mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 + \sum_{d=0}^{k-1}\left(\sum_{t=k}^{\infty} q_t t\right)\mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2\right] \\
\leq&\frac{1}{m}\left[\sum_{d=1}^{J}\left(\sum_{t=k-J}^{\infty} q_t t\right)\mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 + \sum_{d=J+1}^{k-1}\left(\sum_{t=1}^{\infty} q_t t\right)\mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2 + \sum_{d=0}^{k-1}\left(\sum_{t=k-J}^{\infty} q_t t\right)\mathbb{E}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2\right] \\
\leq&\epsilon, \qquad (72)
\end{aligned}$$

where the last inequality is from (69) and (70). Since ϵ is arbitrary, we have $\lim_{k\to\infty} \mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2 = 0$. Now note $\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\| \leq \sqrt{\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\|^2}$ to complete the proof. ■

A.7 Proof of Theorem 11

Proof Let $\{\mathbf{x}^k\}_{k\in\mathcal{K}}$ be a subsequence that converges to $\mathbf{x}^*$. Since $\mathbb{E}\|\mathbf{x}^k - \bar{\mathbf{x}}^{k+1}\| \to 0$ as $\mathcal{K} \ni k \to \infty$, from Markov inequality, $\|\mathbf{x}^k - \bar{\mathbf{x}}^{k+1}\|$ converges to *zero* in probabilty as $\mathcal{K} \ni k \to \infty$. By (Gut, 2006, Theorem 3.4, pp.212), there is a subsubsequence $\{\mathbf{x}^k\}_{k\in\mathcal{K}'}$

such that $\|\mathbf{x}^k - \bar{\mathbf{x}}^{k+1}\|$ 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 $\mathcal{K}' \ni k \to \infty$.

Note that

$$\begin{aligned}
&\lim_{k\to\infty} \mathbb{E}\mathrm{dist}\big(\mathbf{0}, \partial F(\bar{\mathbf{x}}^{k+1})\big)\\
\overset{(61)}{\leq} &\lim_{k\to\infty} \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{k+1}) - \nabla f(\mathbf{x}^{k-j_k}) - \frac{1}{\eta}(\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k)\|\\
\leq &\lim_{k\to\infty} \left(L_f \mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^{k-j_k}\| + \frac{1}{\eta}\mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|\right)\\
\leq &\lim_{k\to\infty} \left(L_f \mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\| + L_f\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-j_k}\| + \frac{1}{\eta}\mathbb{E}\|\bar{\mathbf{x}}^{k+1} - \mathbf{x}^k\|\right)\\
= &0
\end{aligned}$$

If necessary, passing to another subsequence, we use Markov inequality and (Gut, 2006, Theorem 3.4, pp.212) again to have $\mathrm{dist}\big(\mathbf{0}, \partial F(\bar{\mathbf{x}}^{k+1})\big)$ almost surely converges to *zero* as $\mathcal{K}' \ni k \to \infty$. Now use the outer semicontinuity Rockafellar and Wets (2009) of $\mathrm{dist}\big(\mathbf{0}, \partial F(\mathbf{x})\big)$ to obtain the desired result. ■

A.8 Proof of Lemma 14

Proof It is easy to show (23) by noting that $t\rho^t$ is dominated by σ^t as t is sufficiently large. Next we show (25) by induction.

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

$$\|\mathbf{x}^{k-1} - \bar{\mathbf{x}}^k\|^2 - \|\mathbf{x}^k - \bar{\mathbf{x}}^{k+1}\|^2 \leq 2\|\mathbf{x}^{k-1} - \bar{\mathbf{x}}^k\| \cdot \|\mathbf{x}^k - \bar{\mathbf{x}}^{k+1} - \mathbf{x}^{k-1} + \bar{\mathbf{x}}^k\|, \forall k. \tag{73}$$

In addition, for all k,

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{k-1} - \mathbf{x}^k\| \cdot \|\mathbf{x}^{k-1} - \bar{\mathbf{x}}^k\|\\
\leq &\frac{1}{2}\mathbb{E}\left(\sqrt{m}\|\mathbf{x}^{k-1} - \mathbf{x}^k\|^2 + \frac{1}{\sqrt{m}}\|\mathbf{x}^{k-1} - \bar{\mathbf{x}}^k\|^2\right)\\
= &\frac{1}{\sqrt{m}}\mathbb{E}\|\mathbf{x}^{k-1} - \bar{\mathbf{x}}^k\|^2.
\end{aligned} \tag{74}$$

Furthermore,

$$\begin{aligned}
&\|\mathbf{x}^k - \bar{\mathbf{x}}^{k+1} - \mathbf{x}^{k-1} + \bar{\mathbf{x}}^k\|\\
= &\left\|\mathbf{x}^k - \mathbf{prox}_{\eta R}\big(\mathbf{x}^k - \eta\nabla f(\mathbf{x}^{k-j_k})\big) - \mathbf{x}^{k-1} + \mathbf{prox}_{\eta R}\big(\mathbf{x}^{k-1} - \eta\nabla f(\mathbf{x}^{k-1-j_{k-1}})\big)\right\|\\
\leq &\|\mathbf{x}^k - \mathbf{x}^{k-1}\| + \|\mathbf{x}^k - \eta\nabla f(\mathbf{x}^{k-j_k}) - \mathbf{x}^{k-1} + \eta\nabla f(\mathbf{x}^{k-1-j_{k-1}})\|\\
\leq &2\|\mathbf{x}^k - \mathbf{x}^{k-1}\| + \eta\|\nabla f(\mathbf{x}^{k-j_k}) - \nabla f(\mathbf{x}^{k-1-j_{k-1}})\| &(75)\\
\leq &2\|\mathbf{x}^k - \mathbf{x}^{k-1}\| + \eta\big(\|\nabla f(\mathbf{x}^{k-j_k}) - \nabla f(\mathbf{x}^k)\| + \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1-j_{k-1}})\|\big). &(76)
\end{aligned}$$

When $k = 1$, we have $j_0 = 0$ and $j_1 \in \{0, 1\}$ because $j_k \leq k$, $\forall k$. Hence, from (75), it holds that

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

which together with (73) and (74) implies

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

Hence,

$$\mathbb{E}\|\mathbf{x}^0 - \bar{\mathbf{x}}^1\|^2 \leq \left(1 - \frac{4 + 2\eta L_r}{\sqrt{m}}\right)^{-1} \mathbb{E}\|\mathbf{x}^1 - \bar{\mathbf{x}}^2\|^2 \overset{(24)}{\leq} \rho\mathbb{E}\|\mathbf{x}^1 - \bar{\mathbf{x}}^2\|^2.$$

Assume (25) holds for all $k \leq K - 1$, and we show it also holds for $k = K$. First, we have for any $d \leq K - 1$,

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\| \cdot \|\mathbf{x}^d - \mathbf{x}^{d+1}\| \\
\leq &\frac{1}{2}\mathbb{E}\left[\frac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\|^2 + \frac{\sqrt{m}}{\rho^{\frac{K-1-d}{2}}}\|\mathbf{x}^d - \mathbf{x}^{d+1}\|^2\right] \\
= &\frac{1}{2}\mathbb{E}\left[\frac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\|^2 + \frac{1}{\sqrt{m}\rho^{\frac{K-1-d}{2}}}\|\mathbf{x}^d - \bar{\mathbf{x}}^{d+1}\|^2\right] \\
\leq &\frac{1}{2}\mathbb{E}\left[\frac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\|^2 + \frac{\rho^{K-1-d}}{\sqrt{m}\rho^{\frac{K-1-d}{2}}}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\|^2\right] \\
= &\frac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\mathbb{E}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\|^2.
\end{aligned} \tag{77}$$

Secondly, we have

$$\begin{aligned}
&\mathbb{E}\big[\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\|^2 - \|\mathbf{x}^K - \bar{\mathbf{x}}^{K+1}\|^2\big] \\
\overset{(73)}{\leq} &2\mathbb{E}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\| \cdot \|\mathbf{x}^K - \bar{\mathbf{x}}^{K+1} - \mathbf{x}^{K-1} + \bar{\mathbf{x}}^K\| \\
\overset{(76)}{\leq} &4\mathbb{E}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\| \cdot \|\mathbf{x}^K - \mathbf{x}^{K-1}\| \\
&+ 2\eta\mathbb{E}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\|\big(\|\nabla f(\mathbf{x}^{K-j_K}) - \nabla f(\mathbf{x}^K)\| + \|\nabla f(\mathbf{x}^K) - \nabla f(\mathbf{x}^{K-1})\| \\
&+ \|\nabla f(\mathbf{x}^{K-1}) - \nabla f(\mathbf{x}^{K-1-j_{K-1}})\|\big) \\
\overset{(74)}{\leq} &\frac{4 + 2\eta L_r}{\sqrt{m}}\mathbb{E}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\|^2 + 2\eta\mathbb{E}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\| \cdot \|\nabla f(\mathbf{x}^{K-j_K}) - \nabla f(\mathbf{x}^K)\| \\
&+ 2\eta\mathbb{E}\|\mathbf{x}^{K-1} - \bar{\mathbf{x}}^K\| \cdot \|\nabla f(\mathbf{x}^{K-1}) - \nabla f(\mathbf{x}^{K-1-j_{K-1}})\|.
\end{aligned} \tag{78}$$

Note that

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|\cdot\|\nabla f(\mathbf{x}^{K-j_K})-\nabla f(\mathbf{x}^K)\|\\
=&\sum_{t=1}^{K-1} q_t\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|\cdot\|\nabla f(\mathbf{x}^{K-t})-\nabla f(\mathbf{x}^K)\|+c_K\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|\cdot\|\nabla f(\mathbf{x}^0)-\nabla f(\mathbf{x}^K)\|\\
\le&\sum_{t=1}^{K-1} q_t\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|\left(\sum_{d=K-t}^{K-1}\|\nabla f(\mathbf{x}^d)-\nabla f(\mathbf{x}^{d+1})\|\right)+c_K\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|\left(\sum_{d=0}^{K-1}\|\nabla f(\mathbf{x}^d)-\nabla f(\mathbf{x}^{d+1})\|\right)\\
\le&L_r\sum_{t=1}^{K-1} q_t\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|\left(\sum_{d=K-t}^{K-1}\|\mathbf{x}^d-\mathbf{x}^{d+1}\|\right)+L_rc_K\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|\left(\sum_{d=0}^{K-1}\|\mathbf{x}^d-\mathbf{x}^{d+1}\|\right)\\
\overset{(77)}{\le}&L_r\sum_{t=1}^{K-1} q_t\sum_{d=K-t}^{K-1}\frac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|^2+L_rc_K\sum_{d=0}^{K-1}\frac{\rho^{\frac{K-1-d}{2}}}{\sqrt{m}}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|^2\\
=&\frac{L_r}{\sqrt{m}}\sum_{t=1}^{K-1} q_t\frac{\rho^{t/2}-1}{\rho^{1/2}-1}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|^2+\frac{L_rc_K}{\sqrt{m}}\frac{\rho^{K/2}-1}{\rho^{1/2}-1}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|^2\\
\le&\frac{L_r}{\sqrt{m}}\gamma_{\rho,1}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^K\|^2, \qquad (79)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|\cdot\|\nabla f(\mathbf{x}^{K-1})-\nabla f(\mathbf{x}^{K-1-j_{K-1}})\| \\
\leq&\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|\left(\sum_{d=K-1-j_{K-1}}^{K-2}\|\nabla f(\mathbf{x}^{d})-\nabla f(\mathbf{x}^{d+1})\|\right) \\
\leq&L_r\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|\left(\sum_{d=K-1-j_{K-1}}^{K-2}\|\mathbf{x}^{d}-\mathbf{x}^{d+1}\|\right) \\
\leq&\frac{L_r}{2\beta}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|^2+\frac{L_r\beta}{2}\mathbb{E}\left(\sum_{d=K-1-j_{K-1}}^{K-2}\|\mathbf{x}^{d}-\mathbf{x}^{d+1}\|\right)^2 \qquad \text{(for any } \beta>0\text{)} \\
=&\frac{L_r}{2\beta}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|^2+\frac{L_r\beta}{2}\sum_{t=1}^{K-2}q_t\mathbb{E}\left(\sum_{d=K-1-t}^{K-2}\|\mathbf{x}^{d}-\mathbf{x}^{d+1}\|\right)^2+\frac{L_r\beta}{2}c_{K-1}\mathbb{E}\left(\sum_{d=0}^{K-2}\|\mathbf{x}^{d}-\mathbf{x}^{d+1}\|\right)^2 \\
\leq&\frac{L_r}{2\beta}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|^2+\frac{L_r\beta}{2}\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+\frac{L_r\beta}{2}c_{K-1}(K-1)\sum_{d=0}^{K-2}\mathbb{E}\|\mathbf{x}^{d}-\mathbf{x}^{d+1}\|^2 \\
=&\frac{L_r}{2\beta}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|^2+\frac{L_r\beta}{2m}\sum_{t=1}^{K-2}q_t t\sum_{d=K-1-t}^{K-2}\mathbb{E}\|\mathbf{x}^{d}-\bar{\mathbf{x}}^{d+1}\|^2+\frac{L_r\beta}{2m}c_{K-1}(K-1)\sum_{d=0}^{K-2}\mathbb{E}\|\mathbf{x}^{d}-\bar{\mathbf{x}}^{d+1}\|^2 \\
\leq&\left(\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}\right)\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|^2 \\
=&\frac{L_r}{\sqrt{m}}\left(\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}}\right)^{1/2}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|^2 \\
\leq&\frac{L_r\gamma_{\rho,2}}{\sqrt{m}}\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|^2, \qquad (80)
\end{aligned}
$$

where in the last equality, we have let $\beta=\frac{\sqrt{m}}{\left(\sum_{t=1}^{K-2}q_t t\frac{\rho^{t}-1}{1-\rho^{-1}}+c_{K-1}(K-1)\frac{\rho^{K-1}-1}{1-\rho^{-1}}\right)^{1/2}}$.

Substituting (79) and (80) into (78) gives

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

and thus

$$
\mathbb{E}\|\mathbf{x}^{K-1}-\bar{\mathbf{x}}^{K}\|^2\leq\left(1-\frac{4+2\eta L_r(1+\gamma_{\rho,1}+\gamma_{\rho,2})}{\sqrt{m}}\right)^{-1}\mathbb{E}\|\mathbf{x}^{K}-\bar{\mathbf{x}}^{K+1}\|^2\overset{(24)}{\leq}\rho\mathbb{E}\|\mathbf{x}^{K}-\bar{\mathbf{x}}^{K+1}\|^2.
$$

Therefore, by induction, it follows that (25) holds for all k, and we complete the proof. ■

A.9 Proof of Theorem 15

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

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

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

$$r_{i_k}(\mathbf{x}_{i_k}) \geq r_{i_k}(\mathbf{x}_{i_k}^{k+1}) - \left\langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}) + \frac{1}{\eta}(\mathbf{x}_{i_k}^{k+1} - \mathbf{x}_{i_k}^k), \mathbf{x}_{i_k} - \mathbf{x}_{i_k}^{k+1} \right\rangle. \tag{81}$$

Then we have

$$\begin{aligned}
&\left\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^{k+1})\right\|^2 \\
\leq &\left\|\mathbf{x}^{k+1} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 \\
= &\left\|\mathbf{x}^{k} + U_{i_k}(\mathbf{x}^{k+1} - \mathbf{x}^k) - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 \\
= &\left\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 + \|U_{i_k}(\mathbf{x}^{k+1} - \mathbf{x}^k)\|^2 + 2\langle \mathbf{x}^k - \mathcal{P}_{X^*}(\mathbf{x}^k), U_{i_k}(\mathbf{x}^{k+1} - \mathbf{x}^k)\rangle \\
= &\left\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 - \|\mathbf{x}_{i_k}^{k+1} - \mathbf{x}_{i_k}^k\|^2 + 2\langle \mathbf{x}_{i_k}^{k+1} - (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}, \mathbf{x}_{i_k}^{k+1} - \mathbf{x}_{i_k}^k\rangle \\
\overset{(81)}{\leq} &\left\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 - \|\mathbf{x}_{i_k}^{k+1} - \mathbf{x}_{i_k}^k\|^2 \\
&+ 2\eta\left(r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}) - r_{i_k}(\mathbf{x}_{i_k}^{k+1}) + \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\right) \\
= &\left\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 - \|\mathbf{x}_{i_k}^{k+1} - \mathbf{x}_{i_k}^k\|^2 \\
&+ 2\eta\left(r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}) - r_{i_k}(\mathbf{x}_{i_k}^{k+1}) + \langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}_{i_k}^{k}\rangle\right) \\
&+ 2\eta\left(\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\right) \\
\overset{(7)}{\leq} &\left\|\mathbf{x}^{k} - \mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2 - \|\mathbf{x}_{i_k}^{k+1} - \mathbf{x}_{i_k}^k\|^2 \\
&+ 2\eta\left(r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}) - r_{i_k}(\mathbf{x}_{i_k}^{k+1}) + \langle \nabla_{i_k} f(\mathbf{x}^{k-j_k}), (\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k} - \mathbf{x}_{i_k}^{k}\rangle\right) \\
&+ 2\eta\left(f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) + \frac{L_c}{2}\|\mathbf{x}_{i_k}^{k} - \mathbf{x}_{i_k}^{k+1}\|^2 + \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\right) \\
= &\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{82}$$

Note that

$$\begin{aligned}\mathbb{E}\langle\nabla_{i_k}f(\mathbf{x}^{k-j_k})-\nabla_{i_k}f(\mathbf{x}^k),\mathbf{x}^k_{i_k}-\mathbf{x}^{k+1}_{i_k}\rangle=&\frac{1}{m}\mathbb{E}\langle\nabla f(\mathbf{x}^{k-j_k})-\nabla f(\mathbf{x}^k),\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\rangle\\ \le&\frac{1}{m}\mathbb{E}\|\nabla f(\mathbf{x}^{k-j_k})-\nabla f(\mathbf{x}^k)\|\cdot\|\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\|\\ \overset{(80)}{\le}&\frac{L_r\gamma_{\rho,2}}{m\sqrt{m}}\mathbb{E}\|\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\|^2,\end{aligned}\tag{83}$$

and

$$\begin{aligned}&\mathbb{E}\left[f(\mathbf{x}^k)-f(\mathbf{x}^{k+1})+r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k})-r_{i_k}(\mathbf{x}^{k+1}_{i_k})\right]\\ =&\mathbb{E}\left[f(\mathbf{x}^k)-f(\mathbf{x}^{k+1})+r_{i_k}((\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k})-r_{i_k}(\mathbf{x}^k_{i_k})+R(\mathbf{x}^k)-R(\mathbf{x}^{k+1})\right]\\ =&\mathbb{E}[F(\mathbf{x}^k)-F(\mathbf{x}^{k+1})]+\frac{1}{m}\mathbb{E}[R(\mathcal{P}_{X^*}(\mathbf{x}^k))-R(\mathbf{x}^k)].\end{aligned}\tag{84}$$

In addition,

$$\begin{aligned}&\mathbb{E}\langle\nabla_{i_k}f(\mathbf{x}^{k-j_k}),(\mathcal{P}_{X^*}(\mathbf{x}^k))_{i_k}-\mathbf{x}^k_{i_k}\rangle\\ =&\frac{1}{m}\mathbb{E}\langle\nabla f(\mathbf{x}^{k-j_k}),\mathcal{P}_{X^*}(\mathbf{x}^k)-\mathbf{x}^k\rangle\\ =&\frac{1}{m}\mathbb{E}\left[\langle\nabla f(\mathbf{x}^{k-j_k}),\mathcal{P}_{X^*}(\mathbf{x}^k)-\mathbf{x}^{k-j_k}\rangle+\langle\nabla f(\mathbf{x}^k),\mathbf{x}^{k-j_k}-\mathbf{x}^k\rangle+\langle\nabla f(\mathbf{x}^{k-j_k})-\nabla f(\mathbf{x}^k),\mathbf{x}^{k-j_k}-\mathbf{x}^k\rangle\right]\\ \le&\frac{1}{m}\mathbb{E}\big[f(\mathcal{P}_{X^*}(\mathbf{x}^k))-f(\mathbf{x}^{k-j_k})+f(\mathbf{x}^{k-j_k})-f(\mathbf{x}^k)\big]+\frac{1}{m}\mathbb{E}\langle\nabla f(\mathbf{x}^{k-j_k})-\nabla f(\mathbf{x}^k),\mathbf{x}^{k-j_k}-\mathbf{x}^k\rangle\\ \le&\frac{1}{m}\mathbb{E}\big[f(\mathcal{P}_{X^*}(\mathbf{x}^k))-f(\mathbf{x}^k)\big]+\frac{L_f}{m}\mathbb{E}\|\mathbf{x}^{k-j_k}-\mathbf{x}^k\|^2\\ \overset{(71)}{\le}&\frac{1}{m}\mathbb{E}\big[f(\mathcal{P}_{X^*}(\mathbf{x}^k))-f(\mathbf{x}^k)\big]+\frac{L_f}{m^2}\left(\sum_{t=1}^{k-1}q_t t\sum_{d=k-t}^{k-1}\mathbb{E}\|\mathbf{x}^d-\bar{\mathbf{x}}^{d+1}\|^2+\sum_{t=k}^{\infty}q_t t\sum_{d=0}^{k-1}\mathbb{E}\|\mathbf{x}^d-\bar{\mathbf{x}}^{d+1}\|^2\right)\\ \overset{(25)}{\le}&\frac{1}{m}\mathbb{E}\big[f(\mathcal{P}_{X^*}(\mathbf{x}^k))-f(\mathbf{x}^k)\big]+\frac{L_f}{m^2}\left(\sum_{t=1}^{k-1}q_t t\sum_{d=k-t}^{k-1}\rho^{k-d}+\sum_{t=k}^{\infty}q_t t\sum_{d=0}^{k-1}\rho^{k-d}\right)\mathbb{E}\|\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\|^2\\ =&\frac{1}{m}\mathbb{E}\big[f(\mathcal{P}_{X^*}(\mathbf{x}^k))-f(\mathbf{x}^k)\big]+\frac{L_f\gamma^2_{\rho,2}}{m^2}\mathbb{E}\|\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\|^2.\end{aligned}\tag{85}$$

Next We show that the expectation of the objective F is monotonically decreasing, i.e., $\mathbb{E}[F(\mathbf{x}^{k+1})]\le\mathbb{E}[F(\mathbf{x}^k)]$. We have, by taking expectation on both sides of (63), that

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}^{k+1})]\le&\mathbb{E}[F(\mathbf{x}^k)]+\frac{1}{m}\left(\frac{L_c}{2}-\frac{1}{\eta}\right)\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2+\frac{1}{m}\mathbb{E}\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^{k-j_k}),\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\rangle\\ \overset{(83)}{\le}&\mathbb{E}[F(\mathbf{x}^k)]+\frac{1}{m}\left(\frac{L_c}{2}-\frac{1}{\eta}\right)\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2+\frac{L_r\gamma_{\rho,2}}{m\sqrt{m}}\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2\\ =&\mathbb{E}[F(\mathbf{x}^k)]+\frac{1}{m}\left(\frac{L_c}{2}-\frac{1}{\eta}+\frac{L_r\gamma_{\rho,2}}{\sqrt{m}}\right)\mathbb{E}\|\bar{\mathbf{x}}^{k+1}-\mathbf{x}^k\|^2\\ \overset{(26)}{\le}&\mathbb{E}[F(\mathbf{x}^k)].\end{aligned}$$

Substituting (83) through (85) into (82) yields

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}^{k+1}-\mathcal{P}_{X^*}(\mathbf{x}^{k+1})\right\|^2\\
\le&\mathbb{E}\left\|\mathbf{x}^{k}-\mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2-(1-\eta L_c)\mathbb{E}\|\mathbf{x}_{i_k}^{k+1}-\mathbf{x}_{i_k}^{k}\|^2+2\eta\left(\frac{1}{m}\mathbb{E}\big[f(\mathcal{P}_{X^*}(\mathbf{x}^k))-f(\mathbf{x}^k)\big]+\frac{L_f\gamma_{\rho,2}^2}{m^2}\mathbb{E}\|\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\|^2\right)\\
&+\frac{2\eta L_r\gamma_{\rho,2}}{m\sqrt{m}}\mathbb{E}\|\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\|^2+2\eta\left(\mathbb{E}[F(\mathbf{x}^k)-F(\mathbf{x}^{k+1})]+\frac{1}{m}\mathbb{E}[R(\mathcal{P}_{X^*}(\mathbf{x}^k))-R(\mathbf{x}^k)]\right)\\
=&\mathbb{E}\left\|\mathbf{x}^{k}-\mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2-\frac{1}{m}\left[1-\eta L_c-\frac{2\eta L_f\gamma_{\rho,2}^2}{m}-\frac{2\eta L_r\gamma_{\rho,2}}{\sqrt{m}}\right]\mathbb{E}\|\mathbf{x}^k-\bar{\mathbf{x}}^{k+1}\|^2\\
&+\frac{2\eta}{m}\mathbb{E}[F^*-F(\mathbf{x}^k)]+2\eta\mathbb{E}[F(\mathbf{x}^k)-F(\mathbf{x}^{k+1})]\\
\overset{(26)}{\le}&\mathbb{E}\left\|\mathbf{x}^{k}-\mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2+\frac{2\eta}{m}\mathbb{E}[F^*-F(\mathbf{x}^k)]+2\eta\mathbb{E}[F(\mathbf{x}^k)-F(\mathbf{x}^{k+1})],
\end{aligned}$$

and thus

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}^{k+1}-\mathcal{P}_{X^*}(\mathbf{x}^{k+1})\right\|^2+2\eta\mathbb{E}[F(\mathbf{x}^{k+1})-F^*]\\
\le&\mathbb{E}\left\|\mathbf{x}^{k}-\mathcal{P}_{X^*}(\mathbf{x}^{k})\right\|^2+2\eta\mathbb{E}[F(\mathbf{x}^{k})-F^*]-\frac{2\eta}{m}\mathbb{E}[F(\mathbf{x}^k)-F^*] \qquad (86)\\
\le&\left\|\mathbf{x}^{0}-\mathcal{P}_{X^*}(\mathbf{x}^{0})\right\|^2+2\eta\mathbb{E}[F(\mathbf{x}^{0})-F^*]-\frac{2\eta}{m}\sum_{t=0}^{k}\mathbb{E}[F(\mathbf{x}^t)-F^*]\\
\le&\left\|\mathbf{x}^{0}-\mathcal{P}_{X^*}(\mathbf{x}^{0})\right\|^2+2\eta\mathbb{E}[F(\mathbf{x}^{0})-F^*]-\frac{2\eta}{m}(k+1)\mathbb{E}[F(\mathbf{x}^{k+1})-F^*], \qquad (87)
\end{aligned}$$

where the last inequality is from the monotonicity of $\mathbb{E}[F(\mathbf{x}^k)]$. Hence, (27) follows.

When F is strongly convex with constant μ, we have

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

and thus from (86), it follows that

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

Therefore, (28) follows, and we complete the proof. ■

Appendix B. Non-identical distribution

In this section, we drop the identical distribution assumption but assume non-identical distributions for all iterations:

$$\text{Prob}(j_k = t) = q_t^k, \forall k, \text{ and } \forall 0 \leq t \leq k,$$

namely, the distribution of j_k depends on k and the delay $j_k \leq k$. We give the convergence proof of Algorithm 1 for smooth (possibly nonconvex) problems. For the nonsmooth case, the generalization is similar.

Theorem 20 *Let $\{\mathbf{x}^k\}_{k\geq 1}$ be the sequence generated from Algorithm 1. Assume there exists positive constants $q,\ Q,\ T_1,\ T_2,$ and T_3 such that*

$$1/2 < q \leq \sum_{t=0}^{\infty} q_t^{t+k} \leq Q, \forall k \geq 1, \tag{88a}$$

$$\sum_{t=1}^{\infty} q_t^t = T_1, \tag{88b}$$

$$\sum_{t=1}^{\infty} t q_t^{t+k} \leq T_2, \ \forall k \geq 1, \tag{88c}$$

$$\sum_{d=1}^{\infty} \sum_{s=d}^{d+k-1} q_s^{d+k} \leq T_3, \forall k \geq 0. \tag{88d}$$

Take the stepsize such that

$$\eta < \frac{(2q-1)/L_c}{q + \kappa T_2/\sqrt{m} + \kappa T_3 Q/\sqrt{m}}. \tag{89}$$

Then

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

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

Remark 21 *This result recovers the identical distribution case in Theorem 3. For the identical distribution case, we have $q = Q = 1$ and $T_1 = T_2 = T_3 = T$.*

The proof of this theorem is similar to that of Theorem 3 with more care taken to the superscript k on the probability q_t^k.

Proof Similar to (38), we have

$$\begin{aligned}\mathbb{E}_{i_k,j_k} f(\mathbf{x}^{k+1}) \leq & f(\mathbf{x}^k) - \frac{\eta}{m}\sum_{t=0}^{k-1} q_t^k \langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-t}), \nabla f(\mathbf{x}^{k-t})\rangle - \frac{\eta}{m} q_k^k \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^0)\rangle \\ & - \left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right)\sum_{t=0}^{k-1} q_t^k \|\nabla f(\mathbf{x}^{k-t})\|^2 + \frac{\eta^2 L_c}{2m} q_k^k \|\nabla f(\mathbf{x}^0)\|^2,\end{aligned} \tag{90}$$

and similar to (41), we have

$$\begin{aligned}&\mathbb{E}[-\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^{k-t}),\nabla f(\mathbf{x}^{k-t})\rangle]\\ \le&\frac{\eta L_r}{2\sqrt{m}}\left(\sum_{d=k-t}^{k-1}\left(\sum_{r=0}^{d-1}q_r^d\mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2+q_d^d\|\nabla f(\mathbf{x}^0)\|^2\right)+t\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2\right).\end{aligned}\tag{91}$$

Substituting (91) into (90) and taking expectation gives

$$\begin{aligned}\mathbb{E}f(\mathbf{x}^{k+1})\le&\mathbb{E}f(\mathbf{x}^k)+\frac{\eta^2L_r}{2m\sqrt{m}}\sum_{t=1}^{k-1}tq_t^k\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2\\&+\frac{\eta^2L_r}{2m\sqrt{m}}\sum_{t=1}^{k-1}q_t^k\sum_{d=k-t}^{k-1}\left(\sum_{r=0}^{d-1}q_r^d\mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2+q_d^d\|\nabla f(\mathbf{x}^0)\|^2\right)\\&-\frac{\eta}{m}q_k^k\mathbb{E}\langle\nabla f(\mathbf{x}^k),\nabla f(\mathbf{x}^0)\rangle-\left(\frac{\eta}{m}-\frac{\eta^2L_c}{2m}\right)\sum_{t=0}^{k-1}q_t^k\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2+\frac{\eta^2L_c}{2m}q_k^k\mathbb{E}\|\nabla f(\mathbf{x}^0)\|^2.\end{aligned}\tag{92}$$

Note that

$$\sum_{t=1}^{k-1}q_t^k\sum_{d=k-t}^{k-1}q_d^d\|\nabla f(\mathbf{x}^0)\|^2=\sum_{d=1}^{k-1}\left(\sum_{t=k-d}^{k-1}q_t^k\right)q_d^d\|\nabla f(\mathbf{x}^0)\|^2=:\sum_{d=1}^{k-1}a_{k-d,k-1}^kq_d^d\|\nabla f(\mathbf{x}^0)\|^2,\tag{93}$$

and similar to (44)

$$\begin{aligned}\sum_{t=1}^{k-1}q_t^k\sum_{d=k-t}^{k-1}\sum_{r=0}^{d-1}q_r^d\mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2=&\sum_{d=1}^{k-1}a_{k-d,k-1}^k\sum_{r=0}^{d-1}q_r^d\mathbb{E}\|\nabla f(\mathbf{x}^{d-r})\|^2\\=&\sum_{d=1}^{k-1}a_{k-d,k-1}^k\sum_{r=1}^{d}q_{d-r}^d\mathbb{E}\|\nabla f(\mathbf{x}^r)\|^2\\=&\sum_{r=1}^{k-1}\left(\sum_{d=r}^{k-1}a_{k-d,k-1}^kq_{d-r}^d\right)\mathbb{E}\|\nabla f(\mathbf{x}^r)\|^2\\=&\sum_{t=1}^{k-1}\left(\sum_{d=1}^{t}a_{d,k-1}^kq_{t-d}^{k-d}\right)\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2.\end{aligned}\tag{94}$$

In addition,

$$-q_k^k\langle\nabla f(\mathbf{x}^k),\nabla f(\mathbf{x}^0)\rangle\le\frac{q_k^k}{2}\|\nabla f(\mathbf{x}^k)\|^2+\frac{q_k^k}{2}\|\nabla f(\mathbf{x}^0)\|^2,\tag{95}$$

Substituting (93) through (95) into (92) and combining terms, we have

$$\begin{aligned}\mathbb{E}f(\mathbf{x}^{k+1}) \leq & \mathbb{E}f(\mathbf{x}^k) + \left(\frac{\eta^2 L_r}{2m\sqrt{m}}\sum_{d=1}^{k-1} a_{k-d,k-1}^k q_d^d + \frac{\eta}{2m}q_k^k + \frac{\eta^2 L_c}{2m}q_k^k\right)\|\nabla f(\mathbf{x}^0)\|^2 \\ & - \left(\left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right)q_0^k - \frac{\eta}{2m}q_k^k\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\ & - \sum_{t=1}^{k-1}\left(\left(\frac{\eta}{m} - \frac{\eta^2 L_c}{2m}\right)q_t^k - \frac{\eta^2 L_r}{2m\sqrt{m}}tq_t^k - \frac{\eta^2 L_r}{2m\sqrt{m}}\sum_{d=1}^{t} a_{d,k-1}^k q_{t-d}^{k-d}\right)\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 \\ =: & \mathbb{E}f(\mathbf{x}^k) + \gamma_k\|\nabla f(\mathbf{x}^0)\|^2 - \beta_k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \sum_{t=1}^{k-1} C_{t,k}\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2. \end{aligned} \tag{96}$$

Especially, when $k = 0$, we have

$$\begin{aligned}\mathbb{E}f(\mathbf{x}^1) \leq & f(\mathbf{x}^0) - \frac{\eta}{m}\|\nabla f(\mathbf{x}^0)\|^2 + \frac{\eta^2 L_c}{2m}\|\nabla f(\mathbf{x}^0)\|^2 \\ \leq & f(\mathbf{x}^0) + \frac{\eta}{2m}\|\nabla f(\mathbf{x}^0)\|^2 + \frac{\eta^2 L_c}{2m}\|\nabla f(\mathbf{x}^0)\|^2 \\ =: & f(\mathbf{x}^0) + \gamma_0\|\nabla f(\mathbf{x}^0)\|^2. \end{aligned}$$

Note that

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{t=1}^{k-1} C_{t,k}\mathbb{E}\|\nabla f(\mathbf{x}^{k-t})\|^2 = \sum_{k=1}^{K}\sum_{t=1}^{k-1} C_{k-t,k}\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 \\ = & \sum_{t=1}^{K-1}\sum_{k=t+1}^{K} C_{k-t,k}\mathbb{E}\|\nabla f(\mathbf{x}^t)\|^2 = \sum_{k=1}^{K-1}\sum_{t=k+1}^{K} C_{t-k,t}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2, \end{aligned} \tag{97}$$

Using (97) and summing (96) from $k = 0$ through K, we have

$$\begin{aligned}\mathbb{E}f(\mathbf{x}^{K+1}) \leq & f(\mathbf{x}^0) + \sum_{k=0}^{K}\gamma_k\|\nabla f(\mathbf{x}^0)\|^2 - \sum_{k=1}^{K}\beta_k\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \sum_{k=1}^{K-1}\sum_{t=k+1}^{K} C_{t-k,t}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\ = & f(\mathbf{x}^0) + \sum_{k=0}^{K}\gamma_k\|\nabla f(\mathbf{x}^0)\|^2 - \beta_K\mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 \\ & - \sum_{k=1}^{K-1}\left(\beta_k + \sum_{t=k+1}^{K} C_{t-k,t}\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2. \end{aligned} \tag{98}$$

Note that

$$\begin{aligned}&\sum_{k=0}^{K}\sum_{d=1}^{k-1} a_{k-d,k-1}^k q_d^d \leq \sum_{k=0}^{\infty}\sum_{d=1}^{k-1} a_{k-d,k-1}^k q_d^d = \sum_{d=1}^{\infty}\sum_{k=d+1}^{\infty} a_{k-d,k-1}^k q_d^d \\ = & \sum_{d=1}^{\infty}\sum_{k=1}^{\infty} a_{k,k+d-1}^{k+d} q_d^d = \sum_{t=1}^{\infty}\sum_{d=1}^{\infty} a_{d,d+t-1}^{d+t} q_t^t = \sum_{t=1}^{\infty}\sum_{d=1}^{\infty}\sum_{s=d}^{d+t-1} q_s^{d+t} q_t^t \leq T_3\sum_{t=1}^{\infty} q_t^t = T_3 T_1. \end{aligned}$$

Hence,

$$\sum_{k=0}^{K}\gamma_k \leq \frac{\eta^2 L_r}{2m\sqrt{m}}T_3T_1 + \left(\frac{\eta}{2m}+\frac{\eta^2 L_c}{2m}\right)\sum_{k=0}^{K} q_k^k \leq \frac{\eta^2 L_r}{2m\sqrt{m}}T_3T_1 + \left(\frac{\eta}{2m}+\frac{\eta^2 L_c}{2m}\right)T_1.$$

In addition,

$$\begin{aligned}&\sum_{t=1}^{\infty}\sum_{d=1}^{t} a_{d,t+k-1}^{t+k} q_{t-d}^{t+k-d} = \sum_{d=1}^{\infty}\sum_{t=d}^{\infty} a_{d,t+k-1}^{t+k} q_{t-d}^{t+k-d} = \sum_{d=1}^{\infty}\sum_{t=0}^{\infty} a_{d,d+t+k-1}^{d+t+k} q_t^{t+k}\\ =&\sum_{t=0}^{\infty}\sum_{d=1}^{\infty}\sum_{s=d}^{d+t+k-1} q_s^{d+t+k} q_t^{t+k} \leq T_3\sum_{t=0}^{\infty} q_t^{t+k} \leq T_3 Q,\end{aligned}$$

and thus, for $k \geq 1$,

$$\begin{aligned}&\beta_k + \sum_{t=k+1}^{\infty} C_{t-k}^k = \beta_k + \sum_{t=1}^{\infty} C_t^{t+k}\\ =&\left(\frac{\eta}{m}-\frac{\eta^2 L_c}{2m}\right)q_0^k - \frac{\eta}{2m}q_k^k + \sum_{t=1}^{\infty}\left(\frac{\eta}{m}-\frac{\eta^2 L_c}{2m}\right)q_t^{t+k} - \sum_{t=1}^{\infty}\frac{\eta^2 L_r}{2m\sqrt{m}}t q_t^{t+k} - \sum_{t=1}^{\infty}\frac{\eta^2 L_r}{2m\sqrt{m}}\sum_{d=1}^{t} a_{d,t+k-1}^{t+k} q_{t-d}^{t+k-d}\\ \geq&\left(\frac{\eta}{m}-\frac{\eta^2 L_c}{2m}\right)\sum_{t=0}^{\infty} q_t^{t+k} - \frac{\eta}{2m}q_k^k - \frac{\eta^2 L_r}{2m\sqrt{m}}T_2 - \frac{\eta^2 L_r}{2m\sqrt{m}}T_3 Q\\ \geq&\left(\frac{\eta}{m}-\frac{\eta^2 L_c}{2m}\right)q - \frac{\eta}{2m} - \frac{\eta^2 L_r}{2m\sqrt{m}}T_2 - \frac{\eta^2 L_r}{2m\sqrt{m}}T_3 Q\\ =&\frac{\eta}{m}\left(q-\frac{1}{2}\right) - \frac{\eta^2}{2m}\left(L_c q + \frac{L_r T_2}{\sqrt{m}} + \frac{L_r T_3 Q}{\sqrt{m}}\right).\end{aligned}$$

Therefore, letting $K \to \infty$ in (98) and using the lower boundedness of f, we have

$$\lim_{K\to\infty}\beta_K \mathbb{E}\|\nabla f(\mathbf{x}^K)\|^2 + \sum_{k=1}^{\infty}\left(\frac{\eta}{m}\left(q-\frac{1}{2}\right) - \frac{\eta^2}{2m}\left(L_c q + \frac{L_r T_2}{\sqrt{m}} + \frac{L_r T_3 Q}{\sqrt{m}}\right)\right)\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 < \infty.$$

Since $\eta < \frac{(2q-1)/L_c}{q+\kappa T_2/\sqrt{m}+\kappa T_3 Q/\sqrt{m}}$, we have $\mathbb{E}\|\nabla f(\mathbf{x}^k)\| \to 0$ as $k \to \infty$. Through the same arguments as those at the end of the proof of Theorem 3, one can show that any limit point is a critical point almost surely. This completes the proof. ■