Asynchronous Coordinate Descent under More Realistic Assumptions

Tao Sun* Robert Hannah† Wotao Yin‡

July 18, 2017

Abstract

Asynchronous-parallel algorithms have the potential to vastly speed up algorithms by eliminating costly synchronization. However, our understanding of these algorithms is limited because the current convergence of asynchronous (block) coordinate descent algorithms are based on somewhat unrealistic assumptions. In particular, the age of the shared optimization variables being used to update a block is assumed to be independent of the block being updated. Also, it is assumed that the updates are applied to randomly chosen blocks. In this paper, we argue that these assumptions either fail to hold or will imply less efficient implementations.

We then prove the convergence of asynchronous-parallel block coordinate descent under more realistic assumptions, in particular, always without the independence assumption. The analysis permits both the deterministic (essentially) cyclic and random rules for block choices. Because a bound on the asynchronous delays may or may not be available, we establish convergence for both bounded delays and unbounded delays. The analysis also covers nonconvex, weakly convex, and strongly convex functions. We construct Lyapunov functions that directly model both objective progress and delays, so delays are not treated errors or noise. A continuous-time ODE is provided to explain the construction at a high level.

Mathematical Subject Classification 90C30, 90C26, 47N10

1 Introduction

In this paper, we consider the asynchronous-parallel block coordinate descent (async-BCD) algorithm for solving

$$\min_{x\in\mathbb{R}^N} f(x) = f(x_1, \ldots, x_N), \tag{1.1}$$

where f is a differentiable function whose gradient is L-Lipschitz continuous.

Async-BCD [13, 12, 15] has virtually the same implementation as regular BCD. The difference is that the threads doing the parallel computation will not wait for all others to finish and share their updates, but merely continue to update with the most recent updates available[1]. In traditional algorithms, latency, bandwidth limits, and unexpected drains on resources, that delay the update of even a single thread will cause the entire system to wait. By eliminating this costly idle time, asynchronous algorithms can be much faster than traditional ones.

*College of Science, National University of Defense Technology, Changsha, 410073, Hunan, China. Email: nudttaosun@gmail.com;nudtsuntao@163.com

†Department of Mathematics, UCLA, 601 Westwood Plz, Los Angeles, CA 90095, USA, Email: RobertHannah89@math.ucla.edu

‡Department of Mathematics, UCLA, 601 Westwood Plz, Los Angeles, CA 90095, USA, Email: wotaoyin@math.ucla.edu

[1] Also the step size needs to be modified to ensure convergence results hold. However in practice traditional step sizes appear to work, barring extreme circumstances.

In async-BCD, each agent continually updates the solution vector, one block at a time, leaving all other blocks unchanged. Each block update is a read-compute-update cycle. It begins with an agent reading x from shared memory or a parameter server and saving it in a local cache as $\hat{x}$. Then, the agent computes a block partial gradient $-\frac{\gamma_k}{L}\nabla_i f(\hat{x})$, where γ_k is a step size. The computing can start before the reading is completed. If $\nabla_i f(\hat{x})$ does not require all components of $\hat{x}$, only the required ones are read. The final step of the cycle depends on the parallel system setup. In a shared memory setup, the agent reads x_i again and writes $x_i - \frac{\gamma_k}{L}\nabla_i f(\hat{x})$ to x_i. In the server-worker setup, the (worker) agent can send $-\frac{\gamma_k}{L}\nabla_i f(\hat{x})$ or just $\nabla_i f(\hat{x})$ to the server and let the server update x_i. Other setups are possible, too. The iteration counter k increments upon the completion of any block update, and the updating block is denote as i_k.

Because the block updates are asynchronous, when a block update is completed, the $\hat{x}$ that is read and then used to compute this update can be outdated at the completion time.

The iteration of asyn-BCD is, therefore, modeled [13] as

$$x_{i_k}^{k+1} = x_{i_k}^k - \frac{\gamma_k}{L}\nabla_{i_k} f(\hat{x}^k), \tag{1.2}$$

and $x_j^{k+1} = x_j^k$ for all non-updating blocks $j \neq i_k$. The convergence behavior of this algorithm depends on the sequence of updated blocks i_k, the step size sequence γ_k, as well as the ages of $\hat{x}^k$ relative to x^k, which is also called *delays*. We define the *delay vector*

$$\vec{j}(k) = (j(k,1), j(k,2), \ldots, j(k,N)) \in \mathbb{Z}^N.$$

More precisely,

$$\hat{x}^k = (x_1^{k-j(k,1)}, x_2^{k-j(k,2)}, \ldots, x_N^{k-j(k,N)}). \tag{1.3}$$

The *k'th delay* (or *current delay*) is $j(k) = \max_{1\le i\le N}\{j(k,i)\}$.

1.1 Dependence between delays and blocks

In previous analyses [12, 13, 15, 9], it is assumed that the block index i_k and the delay $\vec{j}(k)$ were independent sequences. This simplifies proofs, for example, giving $\mathbb{E}_{i_k}(P_{i_k}\nabla f(\hat{x}^k)) = \frac{1}{N}\nabla f(\hat{x}^k)$ when i_k is chosen at random, where P_i denotes the projection to the ith block. Without independence, $\vec{j}(k)$ will depend on i_k, causing $\hat{x}^k$ to be different for each possible i_k and breaking the equality. The independence assumption is unrealistic in practice. Consider a problem where some blocks are more expensive to update than others for they are larger, bear more nonzero entries in the training set, suffer poorer data locality, or many other reasons. Blocks that take longer to update should have greater delays when they are updated because more other updates will have happened between the time that $\hat{x}$ is read and when the update is completed. For the same reason, when blocks are assigned agents, the updates by slower or busier agents will generally have greater delays.

Indeed this turns out to be the case. Experiments were performed on a cluster with 2 nodes, each with 16 threads running on an Intel Xeon CPU E5-2690 v2. The algorithm was applied to the logistic regression problem on the "news20" from LIBSVM, with 64 contiguous coordinate blocks of equal size. Over 2000 epochs, blocks 0, 1, and 15 have average delays of 351, 115, and 28, respectively.

Even on a problem with each block having the same number of nonzeros, and when the computing environment is homogeneous, this dependence persists. We assigned 20 threads to each core, with each thread assigned to a block of 40 coordinates with equal numbers of nonzeros. The mean delay varied from 29 to 50 over the threads. This may be due to the cluster scheduler or issues of data locality, which were hard to examine.

Clearly, there is strong dependence of the delays $\vec{j}(k)$ on the updated block i_k. Let us consider an ideal situation where all blocks are equally difficult, all agents are equally fast, the job scheduling is fair, and data are centrally stored with the same distance to each agent. But, $\vec{j}(k)$ and i_k still depend on the start times of the 1st through kth updates, so $\vec{j}(k)$ and i_k are still related. Therefore, it is necessary to relax the independence assumption when applying the theory to asynchronous solvers.

1.2 Stochastic and deterministic block rules

This paper considers two different *block rules*: *deterministic* and *stochastic*.

For the stochastic block rule, for each update a block is chosen from $\{1, 2, \ldots, N\}$ uniformly at random[2], for instance in [13, 12, 15]. For the deterministic rule, i_k is an arbitrary sequence that is assumed to be *essentially cyclic*. That is, there is an $N' \in \mathbb{N}$, $N' \geq N$, such that each block $i \in \{1, 2, \ldots, N\}$ is updated at least once in a window of N', that is,

$$\text{For each } t \in \mathbb{Z}^+, \exists \text{ integer } K(i,t) \in \{tN', tN'+1, \ldots, (1+t)N'-1\} \text{ such that } i_{K(i,t)} = i.$$

The essentially cyclic rule allows a cycle to go longer than N due to update delays. It encompasses different kinds of *cyclic* rules such as fixed ordering, random permutation, and greedy selection.

The stochastic block rule is easier to analyze because taking expectation will yield a good approximation to the full gradient. It ensures the every block is updated at the specified frequency. However, it can be expensive or even infeasible to implement for the following reasons.

In the shared memory setup, stochastic block rules require random data access, which is not only significantly slower than sequential data access but also cause frequent *cache misses* (waiting for data being fetching from slower cache or the main memory). The cyclic rules clearly avoid these issues.

In the server-worker setup where workers update randomly assigned blocks at each step, each worker must either store all the data or read the required data from the server at every step (in addition to reading x). This overhead is big. Permanently assigning blocks to agents is a more sound choice.

On the other hand, the analysis of cyclic rules generally has to consider the worst ordering and gives worse performance guarantees. In practice, cyclic rules often lead to good performance [7, 8, 3].

1.3 Bounded and unbounded delays

We consider different delay assumptions as well. *Bounded delay* is when $j(k) \leq \tau$ for some fixed $\tau \in \mathbb{Z}^+$ and all iterations k; while the *unbounded delay* allows $\sup_k\{j(k)\} = +\infty$. Bounded and unbounded delays can be further divided into deterministic and stochastic. Deterministic delays refer to the sequence of delay vectors $\vec{j}(0), \vec{j}(1), \vec{j}(2), \ldots$ that is arbitrary or follows an unknown distribution so is treated as arbitrary. Our stochastic delay results apply to distributions that decay faster than $O(k^{-3})$.

Deterministic unbounded delays apply to the case when async-BCD runs on unfamiliar hardware platforms. For convergence, we require a finite $\liminf_k\{j(k)\}$ and the current step size η^k to be chosen adaptive to the current delay $j(k)$, which must be measured.

Bounded delays and stochastic unbounded delays apply when the user can provide the bound and the distribution, respectively. The user can obtain them from previous experience or by running a pilot test. In return, a fixed step size works, and measuring the current delay is not needed.

1.4 Contributions

The contributions are mainly convergence results for three kinds of delays: bounded, stochastic unbounded, deterministic unbounded, while allowing delays to depend on blocks. The results are provided for nonconvex, convex, and strongly convex functions with Lipschitz gradients. Sublinear rates and linear rates are provided, and, in terms of order of magnitude, they match their synchronous results. Due to space limitation, we restrict ourselves to Lipschitz differentiable functions and leave out nonsmooth proximable functions.

Like many analyses of algorithms, our proofs are built on the construction of Lyapunov functions. We provide a simple ODE-based (i.e., continuous time) construction for bounded delays. Once going discrete and considering the three different kinds of delays, the Lyapunov functions inevitably involve complicated coefficients. But, the ODE-based construction illustrates our construction principle and provides insights on how delays affect convergence.

[2]The distribution doesn't have to be uniform. We need only assume that every block has a nonzero probability of being updated. It is easy to adjust our analysis to this case.

Roughly speaking, if the delays are bounded, then convergence should be analyzed over a sliding window of consecutive iterations because all the delays and progress in each window are intimately related. If no uniform bound is known, then the window must extend to the very first iteration.This analysis does bring great news to the practitioner. Basically speaking, even when there is no known load balancing (thus the delays may be sensible to the blocks) or bound of the delays, they can still ensure convergence by our provided step sizes. This applies to both random and deterministic choices of blocks.

We do not treat asynchronicity as noise as some recent papers do[3]. In our setting, modelling asynchronicity in this way destroys valuable information, and leads to inequalities that are too blunt to obtain stronger results. Compared to noise, delays are much less harmful. This is why sublinear and linear rates can be established for weak and strong convex problems respectively, even when delays depend on the blocks and are potentially unbounded.

We understand from the practitioner's point of view: there is a need for information on how to select parameters. However proving convergence in a new setting where there are no comparable convergence results at all takes a lot of the space. This prevents us from obtaining best-possible constants in our convergence rates, which can be extremely time-consuming. The main message is that sublinear and linear convergence can be obtained under more realistic assumptions. Our results do provide step sizes when the delay bound is given, the delay distribution is known, or the current delay is measured. The convergence results are generally tight in terms of the order of the involved quantities, but the constants are perhaps not tight .

1.5 Related work

Our work extends the theory on asynchronous BCD algorithms such as [17, 13, 12]. However, their analysis relies the independence assumption and assume bounded delays. The bounded delay assumption was weakened by recent papers [9, 16], but independence and random blocks were still needed.

Recently [11] proposes (in the SGD setting) a novel "read after" sequence relabeling technique to create the independence. However, enforcing independence in this way creates other artificial implementation requirements that may waste computational resources: For instance, agents need to read *all* shared data before computing their update, even if not all of it is required to compute updates, which can be extremely expensive for sparse data. It is also necessary to recompute certain parameters to prevent a biased update estimator, instead of caching and cheaply updating. Our analysis does not require these kinds of implementation fixes because it does not rely on any kind of unbiased update estimator. Also, our analysis also works for unbounded delays and deterministic block choices.

Related recent works also include [1, 2], which solve our problem with additional convex block-separable terms in the objective. In the first paper [1], independence between blocks and delays is avoided. However, they require a step size that diminishes at $1/k$ and that the sequence of iterate is bounded (which in general may not be true). The second paper [2] relaxes independence by using a different set of assumptions. In particular, their assumption D3 assumes that, regardless of the previous updates, there is a universally positive chance for every block to be updated in the next step. This Markov-type assumption relaxes the independence assumption but does not avoid it. In particular, it is not satisfied by Example 1 below.

In the convex case with a bounded delay τ, the step size in paper [13] is $O(\frac{1}{\tau^2/N})$. In their proofs, the Lyapunov function is based on $\|x^k - x^*\|_2^2$. Our analysis uses a Lyapunov function consisting of both the function value and the sequence history, where the latter vanishes when delays vanish. If the τ is much larger than the blocks of the problem, our result $O(\frac{1}{\tau})$ is better even under our much weaker conditions. The step size bound in [15, 9, 4] is $O(\frac{1}{\tau/\sqrt{N}})$, which is better than ours, but they need the independence assumption and the stochastic block rule.

Recently, [18] introduces an asynchronous method primal-dual for a problem similar to ours but having additional affine linear constraints. The analysis assumes bounded delays, random blocks, and independence.

[3]See, for example, (5.1) and (A.10) in [17], (3.5) in [14], and (14) and Lemma 4 in [6].

1.6 Notation

Let $x^* \in \arg\min f$. For the update in (1.2), we use the following notation:

$$\Delta^k := x^{k+1} - x^k \overset{(1.2)}{=} -\frac{\gamma_k}{L}\nabla_{i_k}, \qquad d^k := x^k - \hat{x}^k. \tag{1.4}$$

We also use the convention $\Delta^k := 0$ if $k < 0$. Let χ^k be the sigma algebra generated by $\{x^0, x^1, \ldots, x^k\}$. Let $\mathbb{E}_{\vec{j}(k)}$ denote the expectation over the value of $\vec{j}(k)$ (when it is a random variable). $\mathbb{E}$ denotes the total expectation.

2 Bounded delays

In this part, we present convergence results for the bounded delays. If the gradient of the function is L-Lipschitz (even if the function is nonconvex), we present the convergence for both the deterministic and stochastic block rule. If the function is convex, we can obtain a sublinear convergence rate. Further, if the function is restricted strongly convex, a linear convergence rate is obtained.

2.1 Continuous-time analysis

Let t be time in this subsection. Consider the ODE

$$\dot{x}(t) = -\eta\nabla f(\hat{x}(t)), \tag{2.1}$$

where $\eta > 0$. If we set $\hat{x}(t) \equiv x(t)$, this system describes a gradient flow, which monotonically decreases $f(x(t))$, and its discretization is the gradient descent iteration. Indeed, we have

$$\frac{d}{dt}f(x(t)) = \langle \nabla f(x(t)), \dot{x}(t)\rangle \overset{(2.1)}{=} -\frac{1}{\eta}\|\dot{x}(t)\|_2^2.$$

Instead, we allow delays (i.e., $\hat{x}(t) \neq x(t)$) and impose the bound $c > 0$ on the delays:

$$\|\hat{x}(t) - x(t)\|_2 \leq \int_{t-c}^{t} \|\dot{x}(s)\|_2 ds. \tag{2.2}$$

The delays introduce inexactness to the gradient flow $f(x(t))$. We lose monotonicity. Indeed,

$$\begin{aligned}\frac{d}{dt}f(x(t)) &= \langle \nabla f(x(t)), \dot{x}(t)\rangle = \langle \nabla f(\hat{x}(t)), \dot{x}(t)\rangle + \langle \nabla f(x(t)) - \nabla f(\hat{x}(t)), \dot{x}(t)\rangle \\ &\overset{a)}{\leq} -\frac{1}{\eta}\|\dot{x}(t)\|_2^2 + L\|x(t) - \hat{x}(t)\|_2 \cdot \|\dot{x}(t)\|_2 \overset{b)}{\leq} -\frac{1}{2\eta}\|\dot{x}(t)\|_2^2 + \frac{\eta c L^2}{2}\int_{t-c}^{t}\|\dot{x}(s)\|_2^2 ds,\end{aligned} \tag{2.3}$$

where a) is from (2.1) and Lipschitzness of ∇f and b) is from the Cauchy-Schwarz inequality $L\|x(t) - \hat{x}(t)\|_2 \cdot \|\dot{x}(t)\|_2 \leq \frac{\|\dot{x}(t)\|_2^2}{2\eta} + \frac{\eta L^2\|x(t)-\hat{x}(t)\|_2^2}{2}$ and, by (2.2), $\|x(t) - \hat{x}(t)\|_2^2 \leq c\int_{t-c}^{t}\|\dot{x}(s)\|_2^2 ds$. The inequalities are generally unavoidable. The integral term is due to the use of delayed gradient.

Therefore, we design an energy with both f and a weighted total kinetic term, where $\gamma > 0$ will be decided below:

$$\xi(t) = f(x(t)) + \gamma\int_{t-c}^{t}\big(s - (t-c)\big)\|\dot{x}(s)\|_2^2 ds. \tag{2.4}$$

$\xi(t)$ has the time derivative

$$\dot{\xi}(t) = \frac{d}{dt}f(x(t)) + \gamma c\|\dot{x}(t)\|_2^2 - \gamma\int_{t-c}^{t}\|\dot{x}(s)\|_2^2 ds.$$

By substituting the bound on $\frac{d}{dt}f(x(t))$ in (2.3), we get

$$\dot{\xi}(t) \leq -(\frac{1}{2\eta} - \gamma)\|\dot{x}(t)\|_2^2 - (\gamma - \frac{\eta c L^2}{2})\int_{t-c}^{t}\|\dot{x}(s)\|_2^2 ds. \tag{2.5}$$

As long as $\eta < \frac{1}{Lc}$, there exists $\gamma > 0$ such that $(\frac{1}{2\eta} - \gamma) > 0$ and $(\gamma - \frac{\eta c L^2}{2}) > 0$, so $\xi(t)$ is monotonically nonincreasing. Assume $\min f$ is finite. Since $\xi(t)$ is lower bounded by $\min f$, $\xi(t)$ must converge, subsequently yielding the convergence of $\dot{\xi} \to 0$, $\dot{x}(t) \to 0$ by (2.5), $\nabla f(\hat{x}(t)) \to 0$ by (2.1), and $\hat{x}(t) - x(t) \to 0$ by (2.2). The last two results further give $\nabla f(x(t)) \to 0$.

2.2 Discrete analysis

The analysis for our discrete iteration (1.2) is based on the following Lyapunov function:

$$\xi_k := f(x^k) + \frac{L}{2\varepsilon} \sum_{i=k-\tau}^{k-1} (i - (k-\tau) + 1) \|\Delta^i\|_2^2. \tag{2.6}$$

for $\varepsilon > 0$ to be determined later based on the step size and τ, the bound on the delays.

In the lemma below, we present a fundamental inequality, which states, regardless of which block i_k is updated and which $\hat{x}^k$ is used to compute the update in (1.2), there is a sufficient descent in our Lyapunov function for a proper step size $1/O(\tau)$.

Lemma 1 (sufficient descent for bounded delays)**.** ***Conditions:*** *Let f be a function (possibly nonconvex) with L-Lipschitz gradient and finite $\min f$. Let $(x^k)_{k\geq 0}$ be generated by the async-BCD algorithm* (1.2)*, and the delays be bounded by τ. Choose the step size*

$$\gamma_k \equiv \gamma = \frac{2c}{2\tau + 1}$$

for arbitrary fixed $0 < c < 1$. ***Result:*** *we can choose $\varepsilon > 0$ to obtain*

$$\xi_k - \xi_{k+1} \geq \frac{1}{2}(\frac{1}{\gamma} - \frac{1}{2} - \tau) L \cdot \|\Delta^k\|_2^2, \tag{2.7}$$

Consequently,

$$\lim_k \|\Delta^k\|_2 = 0, \tag{2.8}$$

$$\min_{1\leq i\leq k} \|\Delta^i\|_2 = o(1/\sqrt{k}). \tag{2.9}$$

The rate in (2.9) concerns the smallest $\|\Delta^i\|_2$ among $i = 1, \ldots, k$. We call it the *running best rate*. Although $\|\Delta^k\|_2$ is not monotonic, (2.8) and (2.9) indicate that $\|\Delta^k\|_2$ decays *overall faster* than $1/\sqrt{k}$.

Based on the lemma, we obtain a very general result for nonconvex problems:

Theorem 1. *Assume the conditions of Lemma 1, for f that may be nonconvex. Under the deterministic block rule, we have*

$$\lim_k \|\nabla f(x^k)\|_2 = 0, \quad \min_{1\leq i\leq k} \|\nabla f(x^k)\|_2 = o(1/\sqrt{k}). \tag{2.10}$$

This rate has the same order of magnitude as standard gradient descent.

2.3 Stochastic block rule

Under the stochastic block rule, an agent picks a block from $\{1, 2, ..., N\}$ uniformly randomly at the beginning of each update. For the kth completed update, the index of the chosen block is i_k. Our result in this subsection relies on the following assumption on the random variable i_k:

$$\mathbb{E}_{i_k}(\|\nabla_{i_k} f(x^{k-\tau})\|_2 \mid \chi^{k-\tau}) = \frac{1}{N} \sum_{i=1}^{N} \|\nabla_i f(x^{k-\tau})\|_2, \tag{2.11}$$

where $\chi^k = \sigma(x^0, x^1, \ldots, x^k, \vec{j}(0), \vec{j}(1), \ldots, \vec{j}(k))$, $k = 0, 1, \ldots$, is the filtration that represents the information that is accumulated as our algorithm runs. It is important to note that (2.11) uses $x^{k-\tau}$ instead of $\hat{x}^k$. The latter is *not* independent of i_k.

Condition (2.11) is a property about i_k rather than ∇f. It states that each of the N possible values of i_k occurs at probability $1/N$ given the information τ and more iterations older. *We can relax* (2.11) *to non-uniform distributions*; indeed, Theorem 2 below only needs that every block has a nonzero probability of being updated given $\chi^{k-\tau}$, that is,

$$\mathbb{E}(\|\nabla_{i_k} f(x^{k-\tau})\|_2 \mid \chi^{k-\tau}) \geq \frac{\varepsilon}{N}\sum_{i=1}^{N}\|\nabla_i f(x^{k-\tau})\|_2, \tag{2.12}$$

for some universal $\varepsilon > 0$. The uniform distribution in Assumption (2.11) is made for convenience and simplicity.

The above assumption is justified since this section assumes a finite delay bound τ and thus the history older than τ, though might still affect i_k, can no longer nullify the chance of i_k taking each of $\{1,\ldots,N\}$. On the other hand, making a similar assumption on $\mathbb{E}(\|\nabla_{i_k} f(x^{k-\tau})\|_2 \mid \chi^t)$, for any $t = k-1, k-2\ldots, k-\tau+1$, would be *unjustified*, as shown in the following example.

Example 1. Consider three different blocks and two identical agents. Assume blocks 1,2,3 take exactly 2,3,4 seconds to update by either agent. The maximal delay is $\tau = 4/2 = 2$. Assume both agents start their first jobs at nearly the same time. If the first completed update is $i_1 = 2$, by one of two agents, then i_2 must equal either 2 or 3; $i_2 = 1$ is impossible. This can be verified by enumerating all the possible combinations of the block choices made by the agents in their first two steps. In general, $i_k = 1$ is impossible when, before the $(k-1)$th completed update, the two agents start their new steps at nearly the same time and $i_{k-1} = 2$.

Next, we present a general result for a possibly nonconvex objective f.

Theorem 2. *Assume the conditions of Lemma 1.Under the stochastic block rule and assumption* (2.11)*, we have:*

$$\lim_k \mathbb{E}\|\nabla f(x^k)\|_2 = 0, \quad \min_{1\leq i\leq k}\mathbb{E}\|\nabla f(x^k)\|_2^2 = o(1/k). \tag{2.13}$$

2.3.1 Sublinear rate under convexity

When the function f is convex, we can obtain convergence rates, for which we need a slightly modified Lyapunov function

$$F_k := f(x^k) + \delta\cdot\sum_{i=k-\tau}^{k-1}(i-(k-\tau)+1)\|\Delta^i\|_2^2, \tag{2.14}$$

where $\delta := [1+\frac{\varepsilon}{2\tau}(\frac{1}{\gamma}-\frac{1}{2}-\tau)]\frac{L}{2\varepsilon}$[4]. We also define $\pi_k := \mathbb{E}(F_k - \min f)$, $\quad S(k,\tau) := \sum_{i=k-\tau}^{k-1}\delta\|\Delta^i\|_2^2$.

Lemma 2. *Assume the conditions of Lemma 1. Furthermore, let f be convex and use the stochastic block rule. Let $\overline{x^k}$ denote the projection of x^k to $\arg\min f$, assumed to exist, and let*

$$\beta := \max\{\tfrac{8NL^2}{\gamma^2}, (12N+2)L^2\tau+\delta\tau\}, \quad \alpha := \beta/[\tfrac{L}{4\tau}(\tfrac{1}{\gamma}-\tfrac{1}{2}-\tau)]. \tag{2.15}$$

Then we have:

$$(\pi_k)^2 \leq \alpha(\pi_k - \pi_{k+1})\cdot(\tau\mathbb{E}S(k,\tau) + \mathbb{E}\|x^k - \overline{x^k}\|_2^2). \tag{2.16}$$

When $\tau = 1$ (nearly no delay), we can obtain $\beta = O(NL^2/\gamma^2)$ and $\alpha = O(\beta\gamma/L) = O(NL/\gamma)$, which matches the result of standard BCD. Unfortunately, delays cause the complication of α.

We now present the sublinear convergence rate.

Theorem 3. *Assume the conditions of Lemma 1. Furthermore, let f be convex and coercive*[5]*, and use the stochastic block rule. Then we have:*

$$\mathbb{E}(f(x^k) - \min f) = O(1/k). \tag{2.17}$$

[4]Here, we assume $\tau \geq 1$.

[5]A function f is coercive if $\|x\| \to \infty$ means $f(x) \to \infty$.

2.3.2 Linear rate under convexity

We next consider when f is *ν-restricted strongly convex*[6] in addition to having L-Lipschitz gradient. That is, for $x \in \operatorname{dom}(f)$,

$$\langle \nabla f(x), x - \operatorname{Proj}_{\arg\min f}(x)\rangle \geq \nu \cdot \operatorname{dist}^2(x, \arg\min f). \tag{2.18}$$

Theorem 4. *Assume the conditions of Lemma 1. Furthermore, let f be ν-strongly convex, and use the stochastic block rule. Then we have:*

$$\mathbb{E}(f(x^k) - \min f) = O(c^k), \tag{2.19}$$

where $c := \frac{\alpha}{\min\{\nu,1\}} / (1 + \frac{\alpha}{\min\{\nu,1\}}) < 1$ *for α given in* (2.15).

3 Stochastic unbounded delay

In this part, the delay vector $\vec{j}(k)$ is an unbounded random variable, which allow extremely large delays in our algorithm. Under some mild restrictions on the distribution of $\vec{j}(k)$, we can still establish convergence. In light of our continuous-time analysis, we must develop a new bound for the last inner product in (2.3), which requires the tail distribution of $j(k)$ to decay sufficiently fast.

Specifically, we define fixed parameters p_j such that $p_j \geq \mathbb{P}(j(k) = j), \forall k$, $s_l = \sum_{j=l}^{+\infty} j p_j$, and $c_i := \sum_{l=i}^{+\infty} s_l$. Clearly, c_0 is larger than $c_1, c_2, \ldots$, and we need c_0 to be finite. Distributions with $p_j = \mathcal{O}(j^{-t})$, $t > 4$, and exponential-decay distributions satisfy this requirement.

Define the Lyapunov function G_k as

$$G_k := f(x^k) + \bar{\delta} \cdot \sum_{i=0}^{k-1} c_{k-1-i} \|\Delta^i\|_2^2, \tag{3.1}$$

where $\bar{\delta} := \frac{L}{2\varepsilon} + (\frac{1}{\gamma} - \frac{1}{2})\frac{L}{c_0} - \frac{L}{\sqrt{c_0}}$. To simplify the presentation, we define $R(k) := \sum_{i=0}^{k} c_{k-i}\mathbb{E}\|\Delta^i\|_2^2$.

Lemma 3 (Sufficient descent for stochastic unbounded delays)**.** ***Conditions:*** *Let f be a function (which may be nonconvex) with L-Lipschitz gradient and finite* $\min f$. *Let delays be stochastic unbounded. Use step size $\gamma_k \equiv \gamma = \frac{2c}{2\sqrt{c_0}+1}$ for arbitrary fixed $0 < c < 1$.* ***Results:*** *we can set $\varepsilon > 0$ to ensures sufficient descent:*

$$\mathbb{E}[G_k - G_{k+1}] \geq \frac{L}{c_0}(\frac{1}{\gamma} - \frac{1}{2} - \sqrt{c_0})R(k). \tag{3.2}$$

And we have

$$\lim_k \mathbb{E}\|\Delta^k\|_2 = 0 \quad \textit{and} \quad \lim_k \mathbb{E}\|d^k\|_2 = 0. \tag{3.3}$$

For technical reasons, it appears to be difficult to obtain the rates of $\mathbb{E}\|\Delta^k\|_2$ and $\mathbb{E}\|d^k\|_2$.

3.1 Deterministic block rule

Theorem 5. *Let the conditions of Lemma 3 hold for f. Under the deterministic block rule, we have:*

$$\lim_k \mathbb{E}\|\nabla f(x^k)\|_2 = 0. \tag{3.4}$$

[6] A condition weaker than ν-strong convexity and useful for problems involving an underdetermined linear mapping Ax; see [10, 12].

3.2 Stochastic block rule

Recall that under the stochastic block rule, the block to update is selected uniformly at random from $\{1, 2, \ldots, N\}$. The previous assumption (2.11), which is made for bounded delays, need to be updated into the following assumption for unbounded delays:

$$\mathbb{E}_{i_k}(\|\nabla_{i_k} f(x^{k-j(k)})\|_2^2) = \frac{1}{N}\sum_{i=1}^{N} \|\nabla_i f(x^{k-j(k)})\|_2^2, \tag{3.5}$$

where $j(k)$ is a random variable on both sides. As argued below (2.11), the uniform distribution can be relaxed, but we use it for simplicity.

Theorem 6. *Let the conditions of Lemma 3 hold. Under the stochastic block rule and assumption* (3.5), *we have*

$$\lim_k \mathbb{E}\|\nabla f(x^k)\|_2 = 0. \tag{3.6}$$

3.2.1 Convergence rate

When f is convex, we can derive convergence rates for

$$\phi_k := \mathbb{E}(G_k - \min f).$$

Lemma 4. *Let the conditions of Lemma 3 hold, and let f be convex. Let $\overline{x^k}$ denote the projection of x^k to* $\arg\min f$. *Let $\overline{\beta} = \max\{\frac{8NL^2}{\gamma^2 c_0}, (12N+2)L^2 + \bar{\delta}\}$ and $\overline{\alpha} = \overline{\beta}/[\frac{L}{2}(\frac{1}{\gamma} - \frac{1}{2} - \sqrt{c_0})]$. Then we have*

$$(\phi_k)^2 \leq \bar{\alpha}(\phi_k - \phi_{k+1}) \cdot (\bar{\delta} R(k) + \mathbb{E}\|x^k - \overline{x^k}\|_2^2), \tag{3.7}$$

A sublinear convergence rate can be obtained if $\sup_k\{\mathbb{E}\|x^k - \overline{x^k}\|_2^2\} < +\infty$, which can be ensured by adding a projection to a large artificial box set that surely contains the solution. Here we only present a linear convergence result.

Theorem 7. *Let the conditions of Lemma 3 hold. In addition, let f be ν-restricted strongly convex and set step size $\gamma_k \equiv \gamma < \frac{2}{2\sqrt{c_0}+1}$, with $c = \frac{\bar{\alpha}\max\{1, \frac{1}{\nu}\}}{1+\bar{\alpha}\max\{1,\frac{1}{\nu}\}} < 1$. Then,*

$$\mathbb{E}(f(x^k) - \min f) = O(c^k). \tag{3.8}$$

4 Deterministic unbounded delays

In this part, we consider deterministic unbounded delays, which require delay-adaptive step sizes. Set positive sequence $(\epsilon_i)_{i\geq 0}$ (which can be optimized later given the delays) such that $\kappa_i := \sum_{j=i}^{+\infty} \epsilon_j$ obeys $\kappa_1 < +\infty$. Set $D_j := \frac{1}{2} + \frac{\kappa_1}{2} + \sum_{i=1}^{j} \frac{1}{2\epsilon_i}$. We use a new Lyapunov function:

$$H_k := f(x^k) + \tfrac{L}{2}\sum_{i=1}^{+\infty} \kappa_i \|\Delta^{k-i}\|_2^2. \tag{4.1}$$

For any $T \geq \liminf j(k)$, let Q_T be the subsequence of $\mathbb{N}$ where the current delay is less than T. The points $x^k,\ k \in Q_T$, have convergence guarantees. The remaining points are still computed, but because they are affected by potentially unbounded delays, their quality is not guaranteed.

Lemma 5 (sufficient descent for unbounded deterministic delays)**.** ***Conditions:*** *Let f be a function (which may be nonconvex) with L-Lipschitz gradient and finite* $\min f$. *The delays $j(k)$ are deterministic and obey* $\liminf j(k) < \infty$. *Use step size $\gamma_k = c/D_{j(k)}$ for arbitrary fixed* $0 < c < 1$. ***Results:*** *We have*

$$H_k - H_{k+1} \geq L(\tfrac{1}{\gamma_k} - D_{j(k)})\|\Delta^k\|_2^2 \tag{4.2}$$

and

$$\lim_{k} \|\Delta^k\|_2 = 0. \tag{4.3}$$

On any subsequence Q_T (for arbitrarily large T), we have:

$$\lim_{(k\in Q_T)\to\infty} \|d^k\|_2 = 0, \quad \lim_{(k\in Q_T)\to\infty} \|\nabla_{i_k} f(\hat{x}^k)\|_2 = 0,$$

To prove our next result, we need a new assumption: essentially cyclically semi-unbounded delay (ECSD), which is slightly stronger than the essentially cyclic assumption: in every window of N' steps, every index i is updated at least once with a delay less than B (at iteration $K(i,t)$). The number B just needs to exist and can be arbitrarily large. It does not affect the step size.

Theorem 8. *Let the conditions of Lemma 5 hold. For the deterministic index rule under the ECSD assumption, for $T \geq B$, we have:*

$$\lim_{(k\in Q_T)\to\infty} \|\nabla f(x^k)\|_2 = 0. \tag{4.4}$$

5 Conclusion

In summary, we have proven a selection of convergence results for async-BCD under bounded and unbounded delays, and stochastic and deterministic block choices. These results do not require the independence assumption that occurs in the vast majority of other work so far. These results were obtained with the use of Lyapunov function techniques, and treating delays directly, rather than modelling them as noise. Future work may involve obtaining a more exhaustive list of convergence results, sharper convergence rates, and an extension to asynchronous stochastic gradient descent-like algorithms.

References

[1] 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.

[2] Loris Cannelli, Francisco Facchinei, Vyacheslav Kungurtsev, and Gesualdo Scutari. Asynchronous parallel algorithms for nonconvex big-data optimization. Part II: Complexity and numerical results. *arXiv preprint arXiv:1701.04900*, 2017.

[3] Yat Tin Chow, Tianyu Wu, and Wotao Yin. Cyclic coordinate update algorithms for fixed-point problems: Analysis and applications. *SIAM Journal on Scientific Computing*, accepted, 2017.

[4] Damek Davis. The asynchronous palm algorithm for nonsmooth nonconvex problems. *arXiv preprint arXiv:1604.00526*, 2016.

[5] Damek Davis and Wotao Yin. Convergence rate analysis of several splitting schemes. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 115–163. Springer, 2016.

[6] Christopher M De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of hogwild-style algorithms. In *Advances in neural information processing systems*, pages 2674–2682, 2015.

[7] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[8] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

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

[10] Ming-Jun Lai and Wotao Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. *SIAM Journal on Imaging Sciences*, 6(2):1059–1091, 2013.

[11] Rémi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. Asaga: Asynchronous Parallel Saga. 2016.

[12] J. Liu and S. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. 25(1):351–376, 2015.

[13] Ji Liu, Stephen J. Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. 16(1):285–322, 2015.

[14] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *arXiv preprint arXiv:1507.06970*, 2015.

[15] 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.

[16] Zhimin Peng, Yangyang Xu, Ming Yan, and Wotao Yin. On the convergence of asynchronous parallel iteration with arbitrary delays. *arXiv preprint arXiv:1612:04425*, 2016.

[17] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 693–701. Curran Associates, Inc., 2011.

[18] Yangyang Xu. Asynchronous parallel primal-dual block update methods. *arXiv preprint arXiv:1705.06391*, 2017.

Appendix

Our analysis uses the following standard inequalities. For any $x^1, x^2, \dots, x^M \in \mathrm{R}^N$ and $\varepsilon > 0$, it holds that

$$\langle x^1, x^2\rangle \le \varepsilon \|x^1\|_2^2 + \frac{1}{\varepsilon}\|x^2\|_2^2 \tag{5.1}$$

$$\langle x^1, x^2\rangle \le \|x^1\|_2 \cdot \|x^2\|_2 \tag{5.2}$$

$$\|\sum_{i=1}^{M} x^i\|_2^2 \le M\big(\sum_{i=1}^{M}\|x^i\|_2^2\big) \tag{5.3}$$

$$\|d^k\|_2 \le \sum_{i=k-j(k)}^{k-1} \|\Delta^i\|_2 \tag{5.4}$$

The last inequality is derived from (1.4), where d^k is defined, using a telescoping sum and the triangle inequality.

Proof of Lemma 1

Note that $\Delta_i^k = \delta(i, i_k)\cdot \Delta_{i_k}^k$, where $\delta(i,i_k) = \begin{cases} 0, & i = i_k \\ 1, & \text{else} \end{cases}$. Recalling the algorithm (1.2), we have:

$$-\langle \Delta^k, \nabla f(\hat{x}^k)\rangle = -\langle \Delta_{i_k}^k, \nabla_{i_k} f(\hat{x}^k)\rangle = \tfrac{L}{\gamma}\|\Delta^k\|_2^2. \tag{5.5}$$

Since ∇f is L-Lipschitz,

$$f(x^{k+1}) \le f(x^k) + \langle \nabla f(x^k), \Delta^k\rangle + \frac{L}{2}\|\Delta^k\|_2^2. \tag{5.6}$$

Hence

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\overset{(5.5)(5.6)}{\le} \langle \nabla f(x^k) - \nabla f(\hat{x}^k), \Delta^k\rangle + (\tfrac{L}{2} - \tfrac{L}{\gamma})\|\Delta^k\|_2^2 \\ &\overset{a)}{\le} L\|d^k\|_2 \cdot \|\Delta^k\|_2 + (\tfrac{L}{2} - \tfrac{L}{\gamma})\|\Delta^k\|_2^2 \\ &\overset{(5.4)}{\le} L \sum_{d=k-\tau}^{k-1} \|\Delta^d\|_2 \cdot \|\Delta^k\|_2 + (\tfrac{L}{2} - \tfrac{L}{\gamma})\|\Delta^k\|_2^2 \\ &\overset{b)}{\le} \tfrac{L}{2\varepsilon} \sum_{i=k-\tau}^{k-1} \|\Delta^i\|_2^2 + \left[\tfrac{(\tau\varepsilon+1)L}{2} - \tfrac{L}{\gamma}\right]\|\Delta^k\|_2^2, \end{aligned} \tag{5.7}$$

where a) follows from (5.2) and the Lipschitz of ∇f, and c) is obtained by applying $a\cdot b \le \frac{1}{2\varepsilon}|a|^2 + \frac{1}{2\varepsilon}|b|^2$ to each term in the sum.

If $\gamma < \frac{2}{2\tau+1}$, we can choose $\varepsilon > 0$ such that $\varepsilon + \frac{1}{\varepsilon} = 1 + \frac{1}{\tau}(\frac{1}{\gamma} - \frac{1}{2})$. Then, it can be verified by direct calculation and substitutions that we have:

$$\begin{aligned} \xi_k - \xi_{k+1} &\overset{(2.6)}{=} f(x^k) - f(x^{k+1}) + \tfrac{L}{2\varepsilon}\sum_{i=k-\tau}^{k-1}(i-(k-\tau)+1)\|\Delta^i\|_2^2 \\ &- \tfrac{L}{2\varepsilon}\sum_{i=k+1-\tau}^{k-1}(i-(k-\tau))\|\Delta^i\|_2^2 - \tfrac{L}{2\varepsilon}\tau\|\Delta^k\|_2^2 \\ &\overset{c)}{=} f(x^k) - f(x^{k+1}) + \tfrac{L}{2\varepsilon}\sum_{i=k-\tau}^{k-1}\|\Delta^i\|_2^2 - \tfrac{L}{2\varepsilon}\tau\|\Delta^k\|_2^2 \overset{(5.7)}{\ge} \tfrac{1}{2}(\tfrac{1}{\gamma} - \tfrac{1}{2} - \tau)L\cdot\|\Delta^k\|_2^2, \end{aligned} \tag{5.8}$$

where c) follows from $(i-(k-\tau)+1)\|\Delta^i\|_2^2-(i-(k-\tau))\|\Delta^i\|_2^2 = \|\Delta^i\|_2^2$. Therefore we have $\|\Delta^k\|_2^2 \in \ell^1$ by using a telescoping sum[7]. This immediately implies (2.8), and (2.9) follows from [Lemma 3, [5]].

Proof of Theorem 1

Let $t = t(k) = \lfloor k/N' \rfloor$. Recall $K(i,t)$ is defined at Sec. 1.1. Notice we have:

$$\begin{aligned} \|\nabla_i f(x^k)\|_2 &\overset{a)}{\leq} \|\nabla_i f(\hat{x}^{K(i,t)})\|_2 + \|\nabla_i f(x^k) - \nabla_i f(\hat{x}^k)\|_2 + \|\nabla_i f(\hat{x}^k) - \nabla_i f(\hat{x}^{K(i,t)})\|_2 \\ &\overset{b)}{\leq} \|\nabla_i f(\hat{x}^{K(i,t)})\|_2 + L\|d^k\|_2 + L \sum_{j=K(i,t)}^{k-1} \|\hat{x}^{j+1} - \hat{x}^j\|_2, \end{aligned} \tag{5.9}$$

where a) is by the triangle inequality and b) by Lipschitz of ∇f and then applying the triangle inequality to the expansion of $\|\hat{x}^k - \hat{x}^{K(i,t)}\|$. We now bound each of the right-hand terms.

From Lemma 1 and by (5.4), we have

$$\lim_k \|d^k\|_2 \leq \lim_k \sum_{i=k-\tau}^{k-1} \|\Delta^i\|_2 = 0. \tag{5.10}$$

By the triangle inequality, we can derive

$$\|\hat{x}^{k+1} - \hat{x}^k\|_2 \leq \|d^k\|_2 + \|d^{k+1}\|_2 + \|\Delta^k\|_2. \tag{5.11}$$

Taking the limitation,

$$\lim_k \|\hat{x}^{k+1} - \hat{x}^k\|_2 = 0. \tag{5.12}$$

Now notice:

$$\|\nabla_i f(\hat{x}^{K(i,t)})\|_2 = \|\nabla_{i_{K(i,t)}} f(\hat{x}^{K(i,t)})\|_2 = \frac{L}{\gamma}\|d^{K(i,t)}\|_2. \tag{5.13}$$

Since, as $k \to \infty$, $K(i,t) \to \infty$ and $\|d^k\|_2 \to 0$, this last term converges to 0 and the limit result is proven. The running best rate is obtained through the following argument: since $\|\Delta^k\|_2$ is square summable (by Lemma 1), so are $\|d^k\|_2$ by (5.4), $\|\hat{x}^{k+1} - \hat{x}^k\|_2$ by (5.11), and $\|\nabla_i f(\hat{x}^{K(i,t)})\|_2$ (in $t = \Theta(k)$) by (5.13). Hence, $\|\nabla_i f(x^k)\|_2$ is square summable. With $\lim_k \|\nabla f(x^k)\|_2 = 0$, we obtain the running best rate again from [Lemma 3, [5]].

Proof of Theorem 2

Taking the expectation on both sides of (2.11) and multiplying N yield

$$N\mathbb{E}\|\nabla_{i_k} f(x^{k-\tau})\|_2 = \sum_{i=1}^{N} \mathbb{E}\|\nabla_i f(x^{k-\tau})\|_2. \tag{5.14}$$

By $\|\cdot\|_2 \leq \|\cdot\|_1$, we get:

$$\mathbb{E}\|\nabla f(x^{k-\tau})\|_2 \leq \sum_{i=1}^{N} \mathbb{E}\|\nabla_i f(x^{k-\tau})\|_2 \overset{(5.14)}{=} N\mathbb{E}\|\nabla_{i_k} f(x^{k-\tau})\|_2. \tag{5.15}$$

[7]We say a sequence a^k is in ℓ^1 if $\sum_{k=1}^{\infty} |a^k| < \infty$.

In the next part, we prove $\mathbb{E}\|\nabla_{i_k} f(x^{k-\tau})\|_2 \to 0$. From (2.7), we can see that $(\|\Delta^k\|_2)_{k\geq 0}$ is bounded. The dominated convergence theorem implies:

$$\lim_k \mathbb{E}\|\Delta^k\|_2 = 0. \tag{5.16}$$

By (5.4), we have:

$$\lim_k \mathbb{E}(\|d^k\|_2) = 0. \tag{5.17}$$

Hence,

$$\lim_k \mathbb{E}\|\nabla_{i_k} f(\hat{x}^k)\|_2 \overset{(1.2)}{=} \frac{L}{\gamma} \lim_k \mathbb{E}\|\Delta^k\|_2 = 0. \tag{5.18}$$

The triangle inequality and L-Lipschitz continuity yield

$$\begin{aligned} \mathbb{E}\|\nabla_{i_k} f(x^{k-\tau})\|_2 &\leq \mathbb{E}\|\nabla_{i_k} f(\hat{x}^k)\|_2 + \mathbb{E}\|\nabla_{i_k} f(x^k) - \nabla_{i_k} f(\hat{x}^k)\|_2 \\ &\quad + \mathbb{E}\|\nabla_{i_k} f(x^k) - \nabla_{i_k} f(x^{k-\tau})\|_2 \\ &\leq \mathbb{E}\|\nabla_{i_k} f(\hat{x}^k)\|_2 + L \cdot \mathbb{E}\|d^k\|_2 + L \sum_{i=k-\tau}^{k-1} \mathbb{E}\|\Delta^i\|_2. \end{aligned} \tag{5.19}$$

Applying (5.16), (5.17), and (5.18) to (5.19) yields

$$\lim_k \mathbb{E}\|\nabla_{i_k} f(x^{k-\tau})\|_2 = 0. \tag{5.20}$$

With (5.15), (5.20) yields

$$\lim_k \mathbb{E}\|\nabla f(x^{k-\tau})\|_2 = 0, \tag{5.21}$$

which is equivalent to

$$\lim_k \mathbb{E}\|\nabla f(x^k)\|_2 = 0. \tag{5.22}$$

Following a proof similar to that of Theorem 1, $\mathbb{E}\|\nabla f(x^k)\|_2^2$ is summable and thus has the running best rate.

Proof of Lemma 2

The proof consists of two steps: in the first one, we prove

$$\pi_k - \pi_{k+1} \geq \frac{L}{4\tau}(\frac{1}{\gamma} - \frac{1}{2} - \tau) \cdot (\mathbb{E}S(k+1, \tau+1)), \tag{5.23}$$

while in the second one, we prove

$$\pi_k^2 \leq \beta \cdot (\mathbb{E}S(k+1, \tau+1)) \cdot (\mathbb{E}S(k, \tau) + \mathbb{E}\|x^k - \overline{x^k}\|_2^2). \tag{5.24}$$

Combining (5.23) and (5.24) gives us the claim in the lemma.

Proving (5.23): Since $\gamma < \frac{2}{2\tau+1}$, we can choose $\varepsilon > 0$ such that

$$\varepsilon + \frac{1}{\varepsilon} = 1 + \frac{1}{\tau}(\frac{1}{\gamma} - \frac{1}{2}) \tag{5.25}$$

Direct subtraction of F_k and F_{k+1} yields:

$$
\begin{aligned}
F_k - F_{k+1} &\overset{a)}{\geq} f(x^k) - f(x^{k+1}) + \delta \sum_{i=k-\tau}^{k-1} (i-(k-\tau)+1)\|\Delta^i\|_2^2 \\
&- \delta \sum_{i=k+1-\tau}^{k-1} (i-(k-\tau))\|\Delta^i\|_2^2 - \delta\tau\|\Delta^k\|_2 \\
&\overset{b)}{=} f(x^k) - f(x^{k+1}) + \delta S(k,\tau) - \delta\tau\|\Delta^k\|_2 \\
&\overset{c)}{\geq} (\delta - \tfrac{L}{2\varepsilon})S(k,\tau) + \left[\tfrac{L}{\gamma} - \tfrac{(\tau\varepsilon+1)L}{2} - \delta\tau\right]\|\Delta^k\|_2^2 \\
&\overset{d)}{=} \tfrac{L}{4\tau}(\tfrac{1}{\gamma} - \tfrac{1}{2} - \tau)\cdot S(k,\tau) + \tfrac{L}{4}(\tfrac{1}{\gamma} - \tfrac{1}{2} - \tau)\cdot\|\Delta^k\|_2^2 \\
&\overset{e)}{\geq} \tfrac{L}{4\tau}(\tfrac{1}{\gamma} - \tfrac{1}{2} - \tau)\cdot S(k,\tau) + \tfrac{L}{4\tau}(\tfrac{1}{\gamma} - \tfrac{1}{2} - \tau)\cdot\|\Delta^k\|_2^2 \\
&\overset{f)}{=} \tfrac{L}{4\tau}(\tfrac{1}{\gamma} - \tfrac{1}{2} - \tau)\cdot S(k+1,\tau+1),
\end{aligned}
\tag{5.26}
$$

where a) follows from the definition F_k, b) from the definition of $S(k,\tau)$, c) from (5.7), d) is a direct computation using (5.25), e) is due to $\tau \geq 1$, and f) is also a result of the definition of $S(k,\tau)$.

Proving (5.24): The convexity of f yields

$$
f(x^k) - f(\overline{x^k}) \leq \langle \nabla f(x^k), \overline{x^k} - x^k \rangle. \tag{5.27}
$$

Let

$$
a^k := \begin{pmatrix} \overline{x^k} - x^k \\ \sqrt{\delta\tau}\Delta^{k-1} \\ \vdots \\ \sqrt{\delta\tau}\Delta^{k-\tau} \end{pmatrix}, \quad b^k := \begin{pmatrix} \nabla f(x^k) \\ \sqrt{\delta\tau}\Delta^{k-1} \\ \vdots \\ \sqrt{\delta\tau}\Delta^{k-\tau} \end{pmatrix}. \tag{5.28}
$$

Using this and the definition of F_k (2.14), we have:

$$
F_k - \min f \leq \langle a^k, b^k \rangle \leq \|a^k\|_2 \|b^k\|_2. \tag{5.29}
$$

We bound $\mathbb{E}\|\nabla_{i_k} f(x^{k-\tau})\|_2^2$ as follows:

$$
\begin{aligned}
\mathbb{E}\|\nabla_{i_k} f(x^{k-\tau})\|_2^2 &\overset{a)}{\leq} \mathbb{E}\big(\|\nabla_{i_k} f(x^k)\|_2 + \|\nabla_{i_k} f(x^{k-\tau}) - \nabla_{i_k} f(x^k)\|_2\big)^2 \\
&\overset{b)}{\leq} 2\mathbb{E}\|\nabla_{i_k} f(x^k)\|_2^2 + 2L^2\tau \sum_{i=k-\tau}^{k-1} \mathbb{E}\|\Delta^i\|_2^2 \\
&\overset{c)}{\leq} 4\mathbb{E}\|\nabla_{i_k} f(\hat{x}^k)\|_2^2 + 4L^2\mathbb{E}\|d^k\|_2^2 + 2L^2\tau \sum_{i=k-\tau}^{k-1} \mathbb{E}\|\Delta^i\|_2^2 \\
&= \tfrac{4L^2}{\gamma^2}\mathbb{E}\|\Delta^k\|_2^2 + 6L^2\tau \sum_{i=k-\tau}^{k-1} \mathbb{E}\|\Delta^i\|_2^2,
\end{aligned}
\tag{5.30}
$$

where a) follows from the triangle inequality, b) from the Lipschitz of ∇f and (5.3), and c) from $\|\nabla_{i_k} f(x^k)\|_2^2 \leq 2\|\nabla_{i_k} f(\hat{x}^k)\|_2^2 + 2\|d^k\|_2^2$ and (5.4). We also have the bound

$$
\|\nabla f(x^k)\|_2^2 \leq 2\|\nabla f(x^{k-\tau})\|_2^2 + 2L^2\tau \sum_{i=k-\tau}^{k-1} \|\Delta^i\|_2^2, \tag{5.31}
$$

Hence, applying (5.14) to (5.30) yields

$$\mathbb{E}\|\nabla f(x^{k-\tau})\|_2^2 \le \tfrac{4NL^2}{\gamma^2}\mathbb{E}\|\Delta^k\|_2^2 + 6NL^2\tau \sum_{i=k-\tau}^{k-1}\mathbb{E}\|\Delta^i\|_2^2,$$

and further with (5.31),

$$\mathbb{E}\|\nabla f(x^k)\|_2^2 \le \tfrac{8NL^2}{\gamma^2}\mathbb{E}\|\Delta^k\|_2^2 + (12N+2)L^2\tau \sum_{i=k-\tau}^{k-1}\mathbb{E}\|\Delta^i\|_2^2. \tag{5.32}$$

Finally we obtain (5.24) from

$$\begin{aligned}
\pi_k^2 =& [\mathbb{E}(F_k - \min f)]^2 \overset{(5.29)}{\le} \mathbb{E}(\|a^k\|_2\|b^k\|_2)^2 \le \mathbb{E}(\|a^k\|_2^2)\cdot\mathbb{E}(\|b^k\|_2^2) \\
&\overset{a)}{\le} (\tau\mathbb{E}S(k,\tau) + \mathbb{E}\|\nabla f(x^k)\|_2^2) \times (\tau\mathbb{E}S(k,\tau) + \mathbb{E}\|x^k - \overline{x^k}\|_2^2) \\
&\overset{b)}{\le} \beta\mathbb{E}S(k+1,\tau+1)\cdot(\tau\mathbb{E}S(k,\tau) + \mathbb{E}\|x^k - \overline{x^k}\|_2^2),
\end{aligned} \tag{5.33}$$

where a) follows from the definitions of a^k, b^k and b) from (5.32) and the definition of $S(k,\tau)$.

Proof of Theorem 3

With (5.26), we can see that $f(x^k) \le F_k \le F_0$. Since f is coercive, the sequence $(x^k)_{k\ge 0}$ is bounded. Hence, we have $\sup_k\{\|x^k - \overline{x^k}\|_2\} < +\infty$. Hence, there exists $R > 0$ such that

$$\alpha(\sum_{i=k-\tau}^{k-1}\tau\delta\mathbb{E}\|\Delta^i\|_2^2 + \mathbb{E}\|x^k - \overline{x^k}\|_2^2) \le \tfrac{1}{R}. \tag{5.34}$$

for all k. Using Lemma 2, we have

$$\pi_k - \pi_{k+1} \ge R\pi_k^2. \tag{5.35}$$

Using (5.26), we can see that $\pi_k \ge \pi_{k+1}$ for all k. Thus, we have

$$\pi_k - \pi_{k+1} \ge R\pi_{k+1}\pi_k \tag{5.36}$$

$$\implies \tfrac{1}{\pi_{k+1}} - \tfrac{1}{\pi_k} \ge R. \tag{5.37}$$

Therefore, using a telescoping sum, we can deduce that:

$$\pi_{k+1} \le \tfrac{1}{kR + \frac{1}{\pi_0}}. \tag{5.38}$$

Noting $\mathbb{E}(f(x^k) - \min f) \le \pi_k$, we have proven the result.

Proof of Theorem 4

We have

$$\mathbb{E}(f(x^k) - \min f) \ge \nu\mathbb{E}\|x^k - \overline{x^k}\|_2^2, \tag{5.39}$$

Hence recalling the definition from (2.14), we have

$$\mathbb{E}\pi_k \ge \nu\mathbb{E}\|x^k - \overline{x^k}\|_2^2 + \sum_{i=k-\tau}^{k-1}\delta\mathbb{E}\|\Delta^i\|_2^2 \ge \min\{\nu,1\}(\mathbb{E}\|x^k - \overline{x^k}\|_2^2 + S(k,\tau)).$$

Using this, the monotonicity of π^k, and Lemma 2 yields

$$\pi_k\pi_{k+1} \le (\pi_k)^2 \le \tfrac{\alpha}{\min\{\nu,1\}}(\pi_k - \pi_{k+1})\cdot\pi_k. \tag{5.40}$$

Rearranging this yields the result.

Proof of Lemma 3

The Lipschitz continuity of ∇f yields

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq \langle \nabla f(x^k), \Delta^k \rangle + \tfrac{L}{2} \|\Delta^k\|_2^2 \\ &\overset{a)}{=} \langle \nabla f(x^k) - \nabla f(\hat{x}^k), \Delta^k \rangle + (\tfrac{L}{2} - \tfrac{L}{\gamma}) \|\Delta^k\|_2^2 \\ &\leq L \|d^k\|_2 \cdot \|\Delta^k\|_2 + (\tfrac{L}{2} - \tfrac{L}{\gamma}) \|\Delta^k\|_2^2, \end{aligned} \tag{5.41}$$

where a) is from $-\frac{L}{\gamma}\|\Delta^k\|_2^2 = \langle \nabla f(\hat{x}^k), \Delta^k \rangle$. We bound the expectation of $\|d^k\|_2^2$ over the delay and using (5.3), we have:

$$\begin{aligned} \mathbb{E}_{\vec{j}(k)}\big(\|d^k\|_2^2 \mid \chi^k\big) &\leq \mathbb{E}_{\vec{j}(k)}\Big(\sum_{l=1}^{j(k)} j(k) \|\Delta^{k-l}\|_2^2 \mid \chi^k\Big) \\ &\leq \sum_{j=1}^{+\infty} j p_j \sum_{l=1}^{j} \|\Delta^{k-l}\|_2^2 \overset{b)}{=} \sum_{l=1}^{+\infty} \Big(\sum_{j=l}^{+\infty} j p_j\Big) \|\Delta^{k-l}\|_2^2 \overset{c)}{\leq} \sum_{i=0}^{k-1} c_{k-i} \|\Delta^i\|_2^2, \end{aligned} \tag{5.42}$$

where in b), we switched the order of summation in the double sum, and c) uses $\sum_{j=l}^{+\infty} j p_j \leq c_l$. Taking total expectation $\mathbb{E}(\cdot)$ on both sides of (5.42), we obtain

$$\mathbb{E}\|d^k\|_2^2 \leq \sum_{i=0}^{k-1} c_{k-i} \mathbb{E}\|\Delta^i\|_2^2 \overset{d)}{\leq} \sum_{i=0}^{k-1} c_{k-1-i} \mathbb{E}\|\Delta^i\|_2^2 = R(k-1), \tag{5.43}$$

where d) is by the fact $(c_i)_{i\geq 0}$ is descending. Hence:

$$\begin{aligned} \mathbb{E}[f(x^{k+1}) - f(x^k)] &\leq L \mathbb{E}\|d^k\|_2 \cdot \|\Delta^k\|_2 + (\tfrac{L}{2} - \tfrac{L}{\gamma}) \mathbb{E}\|\Delta^k\|_2^2 \\ &\leq \tfrac{L}{2\varepsilon} \mathbb{E}\|d^k\|_2^2 + \left[\tfrac{(\varepsilon+1)L}{2} - \tfrac{L}{\gamma}\right] \mathbb{E}\|\Delta^k\|_2^2 \\ &\leq \tfrac{L}{2\varepsilon} \sum_{l=1}^{+\infty} \Big(\sum_{j=l}^{+\infty} j p_j\Big) \mathbb{E}\|\Delta^{k-l}\|_2^2 + \left[\tfrac{(\varepsilon+1)L}{2} - \tfrac{L}{\gamma}\right] \mathbb{E}\|\Delta^k\|_2^2. \end{aligned} \tag{5.44}$$

Since $\gamma < \frac{2}{2\sqrt{c_0}+1}$, we can choose $\varepsilon > 0$ such that

$$\tfrac{1}{2}(\varepsilon + \tfrac{c_0}{\varepsilon}) = \tfrac{1}{\gamma} - \tfrac{1}{2}. \tag{5.45}$$

With such ε and (5.44), direct calculation using the definition of G^k yields (3.2). When $\gamma < \frac{2}{2\sqrt{c_0}+1}$, $\frac{L}{2}(\frac{1}{\gamma} - \frac{1}{2} - \sqrt{c_0}) > 0$. From (3.2), we can see $(R(k))_{k\geq 0}$ is summable (telescoping sum). Thus, we have $\lim_k R(k) = 0$. Then note (5.43) and

$$c_0 \mathbb{E}\|\Delta^k\|_2^2 \leq \sum_{i=0}^{k} c_{k-i} \mathbb{E}(\|\Delta^i\|_2^2) = R(k). \tag{5.46}$$

Hence then have

$$\lim_k \mathbb{E}(\|d^k\|_2^2) = 0, \quad \lim_k \mathbb{E}(\|\Delta^k\|_2^2) = 0. \tag{5.47}$$

Proof of Theorem 5

Let $t = t(k) = \lfloor k/N' \rfloor$. Recalling $K(i,t)$ is defined at Sec. 1.1, we have:

$$\begin{aligned} \|\nabla_i f(x^k)\|_2 &\overset{a)}{\leq} \|\nabla_i f(\hat{x}^{K(i,t)})\|_2 + \|\nabla_i f(x^{K(i,t)}) - \nabla_i f(\hat{x}^{K(i,t)})\|_2 + \|\nabla_i f(x^k) - \nabla_i f(x^{K(i,t)})\|_2 \\ &\overset{b)}{\leq} \|\nabla_i f(\hat{x}^{K(i,t)})\|_2 + L \|d^{K(i,t)}\|_2 + L \sum_{j=K(i,t)}^{k-1} \|\Delta^j\|_2, \end{aligned} \tag{5.48}$$

where a) is by the triangle inequality and b) by the Lipschitz of ∇f and then applying the triangle inequality to the expansion of $\|x^k - x^{K(i,t)}\|$. We now bound each of the right-hand terms.

Since, as $k \to \infty$, $K(i,t) \to \infty$. With the Cauchy-Schwarz inequality and (3.3), we have

$$\lim_k \mathbb{E}\|d^{K(i,t)}\|_2 \le \lim_k (\mathbb{E}\|d^{K(i,t)}\|_2^2)^{\frac{1}{2}} = 0. \tag{5.49}$$

By $\lim_j \mathbb{E}\|\Delta^j\|_2 \le \lim_j (\mathbb{E}\|\Delta^j\|_2^2)^{\frac{1}{2}} = 0$,

$$\lim_k L \sum_{j=K(i,t)}^{k-1} \mathbb{E}\|\Delta^j\|_2 = 0. \tag{5.50}$$

Now notice:

$$\|\nabla_i f(\hat{x}^{K(i,t)})\|_2 = \|\nabla_{i_{K(i,t)}} f(\hat{x}^{K(i,t)})\|_2 = \frac{L}{\gamma}\|d^{K(i,t)}\|_2. \tag{5.51}$$

Since $\mathbb{E}\|d^{K(i,t)}\|_2 \to 0$ as $K(i,t) \to \infty$, we have

$$\lim_k \mathbb{E}\|\nabla_i f(\hat{x}^{K(i,t)})\|_2 = 0. \tag{5.52}$$

Taking expectations on both sides of (5.48), and using (5.49), (5.50) and (5.52), we then prove the result.

Proof of Theorem 6

Recall $j(k)$ defined near (1.3). Similar to the bound of $\|d^k\|_2^2$ in (5.4), we have

$$\mathbb{E}_{\vec{j}(k)}\big(\|x^k - x^{k-j(k)}\|_2^2 \mid \chi^k\big) \le \sum_{i=0}^{k-1} s_{k-1-i}\|\Delta^i\|_2^2. \tag{5.53}$$

Taking total expectations of both sides yields

$$\mathbb{E}\|x^k - x^{k-j(k)}\|_2^2 \le \sum_{i=0}^{k-1} s_{k-1-i}\mathbb{E}\|\Delta^i\|_2^2. \tag{5.54}$$

We have

$$\begin{aligned}
\mathbb{E}\|\nabla_{i_k} f(x^{k-j(k)})\|_2^2 &\overset{a)}{\le} \mathbb{E}(\|\nabla_{i_k} f(x^k)\|_2 + \|\nabla_{i_k} f(x^{k-j(k)}) - \nabla_{i_k} f(x^k)\|_2)^2 \\
&\overset{b)}{\le} 2\mathbb{E}\|\nabla_{i_k} f(x^k)\|_2^2 + 2L^2\mathbb{E}\|x^k - x^{k-j(k)}\|_2^2 \\
&\overset{c)}{\le} 4\mathbb{E}\|\nabla_{i_k} f(\hat{x}^k)\|_2^2 + 4L^2\mathbb{E}\|d^k\|_2^2 + 2L^2\mathbb{E}\|x^k - x^{k-j(k)}\|_2^2 \\
&\overset{d)}{\le} \tfrac{4L^2}{\gamma^2}\mathbb{E}\|\Delta^k\|_2^2 + 6L^2\sum_{i=0}^{k-1} s_{k-1-i}\mathbb{E}\|\Delta^i\|_2^2,
\end{aligned} \tag{5.55}$$

where a) follows from the triangle inequality, b) from the Lipschitz of ∇f and (5.3), and c) from $\|\nabla_{i_k} f(x^k)\|_2^2 \le 2\|\nabla_{i_k} f(\hat{x})\|_2^2 + 2\|d^k\|_2^2$ and (5.4), and d) from (5.54). Taking total expectation of both sides of assumption (3.5) yields

$$\mathbb{E}\|\nabla_{i_k} f(x^{k-j(k)})\|_2^2 = \frac{\mathbb{E}\|\nabla f(x^{k-j(k)})\|_2^2}{N}. \tag{5.56}$$

By the triangle inequality,

$$\|\nabla f(x^k)\|_2^2 \le 2\|\nabla f(x^{k-j(k)})\|_2^2 + 2L^2 \sum_{i=0}^{k-1} s_{k-1-i}\mathbb{E}\|\Delta^i\|_2^2. \tag{5.57}$$

Hence, combining (5.56) and (5.57) produces

$$\mathbb{E}\|\nabla f(x^{k-j(k)})\|_2^2 \le \tfrac{4NL^2}{\gamma^2}\mathbb{E}\|\Delta^k\|_2^2 + 6NL^2 \sum_{i=0}^{k-1} s_{k-1-i}\mathbb{E}\|\Delta^i\|_2^2;$$

which is substituted into (5.57) to yield

$$\mathbb{E}\|\nabla f(x^k)\|_2^2 \le \tfrac{8NL^2}{\gamma^2}\mathbb{E}\|\Delta^k\|_2^2 + (12N+2)L^2 \sum_{i=0}^{k-1} s_{k-1-i}\mathbb{E}\|\Delta^i\|_2^2. \tag{5.58}$$

By $\sum_{i=0}^{k-1} s_{k-1-i} \le \sum_{i=0}^{k-1} c_{k-1-i}\mathbb{E}\|\Delta^i\|_2^2 = R(k-1)$ and (3.2),

$$\lim_k \mathbb{E}\|\nabla f(x^k)\|_2^2 = 0. \tag{5.59}$$

The proof is completed by applying the Cauchy-Schwarz inequality

$$\mathbb{E}\|\nabla f(x^k)\|_2 \le (\mathbb{E}\|\nabla f(x^k)\|_2^2)^{\frac{1}{2}}. \tag{5.60}$$

Proof of Lemma 4

This proof is very similar to Lemma 2 except that $R(k)$ plays the role of $S(k,\tau)$. Let

$$a^k = \begin{pmatrix} \overline{x^k} - x^k \\ \sqrt{c_0\bar{\delta}}\Delta^{k-1} \\ \vdots \\ \sqrt{c_k\bar{\delta}}\Delta^0 \end{pmatrix}, b^k = \begin{pmatrix} \nabla f(x^k) \\ \sqrt{c_0\bar{\delta}}\Delta^{k-1} \\ \vdots \\ \sqrt{c_k\bar{\delta}}\Delta^0 \end{pmatrix}. \tag{5.61}$$

Thus, we have

$$G_k - \min f \le \langle a^k, b^k\rangle \le \|a^k\|_2\|b^k\|_2. \tag{5.62}$$

By taking expectations, we get

$$\mathbb{E}(G_k - \min f) \le \mathbb{E}(\|a^k\|_2\|b^k\|_2) \le [\mathbb{E}\|a^k\|_2^2 \cdot \mathbb{E}\|b^k\|_2^2]^{1/2}. \tag{5.63}$$

By (5.58) and the definitions of $a^k, b^k, R(k)$, we get

$$\begin{aligned}
&[\mathbb{E}(G_k - \min f)]^2 \le \mathbb{E}(\|a^k\|_2^2) \cdot \mathbb{E}(\|b^k\|_2^2) \\
&\le (\bar{\delta}R(k) + \mathbb{E}\|\nabla f(x^k)\|_2^2) \times (\bar{\delta}R(k) + \mathbb{E}\|x^k - \overline{x^k}\|_2^2) \\
&\le \bar{\beta}R(k) \times (R(k) + \mathbb{E}\|x^k - \overline{x^k}\|_2^2).
\end{aligned} \tag{5.64}$$

Finally, from the definition of $\overline{\alpha}$ and Lemma 3, the theorem follows.

Proof of Theorem 7

We have

$$\mathbb{E}(f(x^k) - \min f) \geq \nu \mathbb{E}\|x^k - \overline{x^k}\|_2^2, \tag{5.65}$$

which also means that

$$\mathbb{E}(\bar{\delta} R(k) + \|x^k - \overline{x^k}\|_2^2) \leq \max\{1, \frac{1}{\nu}\}\phi_k. \tag{5.66}$$

Lemma 4 yields

$$(\phi_k)^2 \leq \bar{\alpha} \max\{1, \frac{1}{\nu}\}(\phi_k - \phi_{k+1}) \cdot (\phi_k) \tag{5.67}$$

Note that ϕ_k is decreasing, we obtain

$$\phi_{k+1} \leq \bar{\alpha} \max\{1, \frac{1}{\nu}\}(\phi_k - \phi_{k+1}). \tag{5.68}$$

Then, we have the result by rearrangement.

Proof of Lemma 5

$$\begin{aligned} f(x^{k+1}) &\overset{a)}{\leq} f(x^k) + L\|d^k\|_2 \cdot \|\Delta^k\|_2 + (\tfrac{L}{2} - \tfrac{L}{\gamma_k})\|\Delta^k\|_2^2 \\ &\overset{b)}{\leq} f(x^k) + L \sum_{l=1}^{j(k)} \|\Delta^{k-l}\|_2 \cdot \|\Delta^k\|_2 + (\tfrac{L}{2} - \tfrac{L}{\gamma_k})\|\Delta^k\|_2^2 \\ &\overset{c)}{\leq} f(x^k) + L \sum_{l=1}^{j(k)} (\tfrac{\epsilon_l}{2}\|\Delta^{k-l}\|_2^2 + \tfrac{1}{2\epsilon_l}\|\Delta^k\|_2^2) + (\tfrac{L}{2} - \tfrac{L}{\gamma_k})\|\Delta^k\|_2^2 \\ &= f(x^k) + \tfrac{L}{2} \sum_{l=1}^{j(k)} \epsilon_l \|\Delta^{k-l}\|_2^2 + \tfrac{L}{2} \sum_{l=1}^{j(k)} \tfrac{1}{\epsilon_l}\|\Delta^k\|_2 + (\tfrac{L}{2} - \tfrac{L}{\gamma_k})\|\Delta^k\|_2^2 \\ &\overset{d)}{\leq} f(x^k) + \tfrac{L}{2} \sum_{l=1}^{+\infty} \epsilon_l \|\Delta^{k-l}\|_2^2 + \tfrac{L}{2}(1 + \sum_{l=1}^{j(k)} \tfrac{1}{\epsilon_l} - \tfrac{2}{\gamma_k})\|\Delta^k\|_2^2. \end{aligned} \tag{5.69}$$

where a) follows from Lipschitz of ∇f and definitions of d^k, Δ^k, b) from the triangle inequality, c) from (5.2), and d) from $j(k) < \infty$. Then, a direct calculation yields (4.2). Hence (4.3) follows by summability: $\|\Delta^k\|_2^2 \in \ell^1$.

$$\lim_{k \in Q_T} \|d^k\|_2 \leq \sum_{l=k-T}^{k-1} \lim_{l} \|\Delta^l\|_2 = 0 \tag{5.70}$$

$$L(\frac{1}{\gamma_k} - D_{j(k)})\|\Delta^k\|_2^2 = \frac{c(1-c)}{LD_{j(k)}}\|\nabla_{i_k} f(\hat{x}^k)\|_2^2. \tag{5.71}$$

Therefore,

$$\frac{1}{D_T} \sum_{k \in Q_T} \|\nabla_{i_k} f(\hat{x}^k)\|_2^2 < \sum_k \frac{\|\nabla_{i_k} f(\hat{x}^k)\|_2^2}{D_{j(k)}} < +\infty. \tag{5.72}$$

Proof of Theorem 8

For any T and $k \in Q_T$, let $t = t(k) = \lfloor k/N' \rfloor$, and by the triangle inequality:

$$\begin{aligned} \|\nabla_i f(x^k)\| \le & \|\nabla_i f(x^{K(i,t)}) - \nabla_i f(x^k)\|_2 \\ & + \|\nabla_i f(\hat{x}^{K(i,t)}) - \nabla_i f(x^{K(i,t)})\|_2 + \|\nabla_i f(\hat{x}^{K(i,t)})\|_2. \end{aligned} \tag{5.73}$$

From Lemma 5, we have

$$\lim_k \|\nabla_i f(x^{K(i,t)}) - \nabla_i f(x^k)\|_2 \le \lim_k L \sum_{i=k-N'+1}^{k-1} \|\Delta^i\|_2 = 0. \tag{5.74}$$

Noting $K(i,t) \in Q_T$ by the ECSD assumption, we can derive

$$\lim_k \|\nabla_i f(\hat{x}^{K(i,t)}) - \nabla_i f(x^{K(i,t)})\|_2 \le \quad \lim_k L\|d^{K(i,t)}\|_2 = 0. \tag{5.75}$$

Now notice by Lemma 5:

$$\lim_k \|\nabla_i f(\hat{x}^{K(i,t)})\|_2 = \lim_{K(i,t)} \|\nabla_{i_{K(i,t)}} f(\hat{x}^{K(i,t)})\|_2 = 0.$$

Since $K(i,t) \to \infty$, this right term converges to 0 and the result is proven.