# Block-Coordinate Descent Methods with Adaptivity to Local Smoothness

Yinjun Wang, Wotao Yin

December 1, 2023

**Abstract**

Block-coordinate descent (BCD) methods are effective at solving large-scale optimization problems across diverse domains. Despite their prowess, exact minimization within BCD often entails computational challenges and lacks consistent convergence. This has led to the emergence of approximate minimization variants like block-coordinate gradient descent (BCGD) and block-coordinate proximal gradient (BCPG). However, a recurring challenge with these methods is the intricate task of stepsize tuning, which significantly impacts their empirical performance. In this paper, we introduce novel adaptive variants of BCGD and BCPG, dubbed A-BCGD-n and A-BCPG-n. These methods are inherently adaptive and ensuring fast convergence. Furthermore, they are underpinned by theoretical guarantees, provably achieving sublinear and linear convergence under standard assumptions. Our numerical evaluations substantiate these claims, positioning our proposed methods as potent solutions to address the prevalent challenges in block-coordinate optimization.

## 1 Introduction

Block-coordinate descent (BCD) methods have risen in popularity due to their efficacy in addressing high-dimensional optimization problems spanning domains including machine learning, compressed sensing, signal and image processing, and computational statistics. These methods iterate by minimizing the objective with respect to a chosen block while keeping the remaining blocks fixed. In other words, each BCD step solves a smaller-scale subproblem. Since exact minimization of a subproblem is still computationally expensive, non-exact minimization of a subproblem by, for example, performing gradient descent or proximal-gradient update on the chosen block is more popular. We call them block-coordinate gradient descent (BCGD) and block-coordinate proximal gradient (BCPG), respectively. These approaches have demonstrated superior performance over the standard (full coordinate) methods [16].

Nonetheless, the shift towards these gradient-based BCGD and BCPG methods brings its own set of challenges. One recurring challenge faced by practitioners is choosing their stepsize, which plays a pivotal role in the empirical performance of BCGD and BCPG. Adopting their theoretically largest stepsizes, even if they can be computed, often result in sluggish convergence in practice. Classic methods to improve over fixed stepsizes include linesearch and trust-region methods, which also come with the extra computational cost of objective function evaluations. Alternative strategies, such as Polyak's method [17] and Barzilai-Borwein method [1], come with their own sets of limited convergence guarantees. Recognizing these limitations, [13, 9] propose adaptive stepsize methods for (full coordinate) gradient descent and proximal gradient, respectively. In the $k$-th iteration, they compute an approximate to the local smoothness

$$L_k = \frac{\| \nabla f(x_k) - \nabla f(x_{k-1}) \|}{\| x_k - x_{k-1} \|}$$

and adjust stepsize $\alpha_k$ by:

$$\alpha_k = \min\left\{\frac{1}{2L_k},\ \alpha_{k-1}\sqrt{1+\frac{\alpha_{k-1}}{\alpha_{k-2}}}\right\}. \tag{1}$$

These methods exhibit fast and robust convergence.

However, applying these variants directly to block-coordinate iterations are inadequate, as using the smoothness information of all coordinates to update a selected block requires not only more expensive gradient computations but can also result in conservative stepsizes. Ideally, one would adjust the stepsize by approximating the local smoothness of merely the chosen block, as different blocks can have different smoothness values. Suppose blocks $\sigma(k-1)$ and $\sigma(k)$ are selected at the $k-1$-th and $k$-th iterations, and let

$$[x]_\sigma = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ x_\sigma \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{the } \sigma\text{-th block of } x$$

A direct extension of (1)

$$\begin{aligned} x_k &= x_{k-1} - \alpha_{k-1}[\nabla f(x_{k-1})]_{\sigma(k-1)} \qquad \text{(this previous update is included for reference)} \\ \alpha_k &= \min\left\{\frac{\left\| [x_k]_{\sigma(k-1)} - [x_{k-1}]_{\sigma(k-1)} \right\|}{2\left\| [\nabla f(x_k)]_{\sigma(k-1)} - [\nabla f(x_{k-1})]_{\sigma(k-1)} \right\|},\ \alpha_{k-1}\sqrt{1+\frac{\alpha_{k-1}}{\alpha_{k-2}}}\right\} \\ x_{k+1} &= x_k - \alpha_k[\nabla f(x_k)]_{\sigma(k)} \end{aligned}$$

falls short as the local smoothness of the $\sigma(k-1)$-th block affects the update of the $\sigma(k)$-th block. Consider an extreme-case example with three blocks of variables that represent length in different units: one in kilometers, another in meters, and the third in millimeters. All conditions equal, the smoothness of the last block is 1,000 times of the second and 1,000,000 times of the first. There is a clear need for adaptive variants that is robust across different scalings of the blocks.

In light of these challenges, this paper presents novel stepsize-adjusting rules in the presence of block-wise updates and introduces adaptive variants of BCGD and BCPG. We propose the A-BCGD and A-BCPG that are

1. **adaptive**: the proposed block-coordinate methods are free from manual stepsize tuning; they are designed to dynamically adjust their stepsize, adaptive to local smoothness of the blocks;

2. **provably convergent**: A-BCGD achieves ergodic $\mathcal{O}(\frac{1}{K})$ convergence under *locally* smooth and convex assumption and linear convergence under *locally* smooth and strongly convex assumption; A-BCPG achieves ergodic $\mathcal{O}(\frac{1}{K})$ convergence under the assumption that one objective function is *locally* smooth and strongly convex and the other term is closed, convex, proper, and proximable;

3. **stepsize-aggressive**: numerically, these methods are consistently fast.

The paper is organized as follows: Section 3.2 introduces A-BCGD and provides its proof of convergence. Section 3.3 presents A-BCPG and its proof of convergence. Numerical experiment results demonstrating their compelling performance are presented in Section 4. All proofs are detailed in Appendix.

# 2 Related Work in Literature

## 2.1 Adaptive first-order methods

The development and analysis of our methods follow the line-of-works, which adaptively choose and adjust stepsizes by combining the following estimates of local smoothness, monotonicity, and cocoercivity:

$$L_k = \frac{\| \nabla f(x_k) - \nabla f(x_{k-1}) \|}{\| x_k - x_{k-1} \|}$$
$$A_k = \frac{\langle \nabla f(x_k) - \nabla f(x_{k-1}), x_k - x_{k-1} \rangle}{\| x_k - x_{k-1} \|^2}$$
$$B_k = \frac{\| \nabla f(x_k) - \nabla f(x_{k-1}) \|^2}{\langle \nabla f(x_k) - \nabla f(x_{k-1}), x_k - x_{k-1} \rangle}.$$

The original paper [13] focuses on smooth convex optimization problems and proposes an adaptive variant of gradient descent. Then, [26] extends the analysis of [13] to handle two-term composite convex optimization problems with one smooth term and one nonsmooth term and proposes an adaptive primal-dual method. The work [9, 14] extends the analysis techniques and proposes an adaptive variant of the proximal gradient method; [9] additionally proposes an adaptive primal-dual method to solve three-term composite convex optimization problems with one smooth term and two nonsmooth terms. There are also further extensions of handling bilevel optimization [8], federated optimization [6] and second-order optimization [5]. Additionally, [12] proposes an adaptive golden ratio method for solving monotone variational inequalities.

Adaptive online methods have received a lot of attention in recent years and remain an active topic of research. These methods automate the stepsize with the introduction of additional hyperparameters. AdaGrad [3] pioneers this direction, followed by RMSProp [25], Adam [7], and other adaptive online methods [20] supporting neural network training. This direction of research is orthogonal to ours since their theories usually assume Lipschitzness of the objective function rather than of the gradients, and the corresponding objective is usually neither smooth nor convex. For a comprehensive comprehensive summary and comparison among adaptive online methods, see [20].

In this paper, we extend the analysis of adaptive gradient descent in [13] to handle that of A-BCGD and extend the analysis of adaptive proximal gradient in [9] to handle that of A-BCPG. Specifically, we create inequalities inspired by the construction of the inequality in Lemma 1 of [13] and that in Lemma 2.2 of [9].

## 2.2 Block-coordinate descent

BCD methods have enjoyed increasing popularity in recent years due to their effectiveness in solving high-dimensional optimization problems. They solve optimization problems by successively minimizing the objective function along each (block of) coordinate or coordinate hyperplane, which is ideal for parallel and distributed computing. The strong performance and parallelizability of BCD rely on solving subproblems that consist of fewer variables and have low complexities and low memory requirements.

For certain structured problems [16], using BCD saves much computation relative to the full coordinate update.

In practice, exactly minimizing the objective along the selected coordinates can sometimes be hard since its corresponding subproblem may be difficult to solve exactly. In addition, BCD with this update scheme may not converge for some non-smooth or non-convex problems. This deficiency motivates the introduction of alternative update schemes, such as proximal point and proximal linear, that give easier subproblems and ensure the convergence under weaker conditions. These update schemes may be interpreted as minimizing a surrogate function that upper bounds the original objective function when the stepsize is chosen appropriately.

There are various implementation approaches in choosing the block, including but not limited to cyclic selection rules![19, 24, 22], IID uniform random selection rules![18], independent and random but non-uniform selection rules![18], independent and random but non-uniform selection rules with probabilities inversely proportional to the coordinate-wise Lipschitz constants [18], random permutation selection rules that access the coordinates in a cyclic fashion but with the order shuffled every epoch [10, 4, 23, 27], and greedy selection rules [2, 11].

For BCD reviews, see [28, 21].

# 3 Algorithm Development

## 3.1 Notation

This paper focuses on the following composite convex optimization problem

$$\underset{x\in\mathbb{R}^p}{\text{minimize}} \quad h(x) \coloneqq f(x) + g(x), \tag{2}$$

where $f$ is locally smooth and $g$ is block-separable. Define

$$\nabla f(x) \triangleq \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_p}(x) \end{pmatrix} \in \mathbb{R}^p, \quad [\nabla f(x)]_i \triangleq \frac{\partial f}{\partial x_i}(x) \in \mathbb{R}, \quad [x]_i \triangleq x_i \in \mathbb{R},$$

where $i$ is an entry index. We also extend this notation to block index $\sigma \in \{1, 2, \ldots, s\}$:

$$[\nabla f(x)]_i \triangleq \frac{\partial f}{\partial x_i}(x) \in \mathbb{R}^{\frac{p}{s}}, \quad [x]_i \triangleq x_i \in \mathbb{R}^{\frac{p}{s}}.$$

Hereafter, $i$ is an entry index, $\sigma$ is a block index, $\mathcal{S}_\sigma$ is the set of entries in the $\sigma$-th block, and we write $[x]_\sigma = [x]_{\mathcal{S}_\sigma}$ for simplicity.

Given a function $f : \mathcal{X} \to \mathbb{R}$ and $L > 0$, we say that $f$ is $L$-smooth if its gradient $\nabla f$ is Lipschitz continuous: $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \forall x, y$. We say $f$ is locally smooth if $\nabla f$ is Lipschitz continuous on any compact subset $\mathcal{C}$, that is, $\forall \mathcal{C} \subset \mathcal{X}$, $\exists L_{\mathcal{C}} > 0$, i.e., $\|\nabla f(x) - \nabla f(y)\| \le L_{\mathcal{C}}\|x - y\|, \forall x, y \in \mathcal{C}$. There are many interesting locally-smooth functions that are not globally smooth: $x \mapsto \exp(x), \tan(x), x^p$ for $p > 2$ in $\mathbb{R}$, and $\log(x)$ in $\mathbb{R}_{++}$. More generally, any continuously differentiable function is locally smooth. We say $f$ is locally strongly convex if, for any compact subset $\mathcal{C}$: $\forall \mathcal{C} \subset \mathcal{X}$, there exists $\mu_{\mathcal{C}} > 0$ such that $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu_{\mathcal{C}}}{2}\|x - y\|^2, \forall x, y \in \mathcal{C}$. Finally, the closed convex hull generated by $\{x^*, x_0, x_1, \ldots\}$ is denoted as $\overline{\text{Conv}}(\{x^*, x_0, x_1, \ldots\})$.

### 3.2 A-BCGD-n

This subsection assumes $g \equiv 0$ and introduces our adaptive variant of BCGD, dubbed A-BCGD-n (Algorithm 1).

**Algorithm 1** A-BCGD-n

**Input:** $x_{-1} \in \mathbb{R}^p$, $\alpha_0 > 0$, $\theta_0^\sigma = +\infty$ for $\sigma = 1, 2, \ldots, s$, $N \geq 2$

$x_0 \leftarrow x_{-1} - \alpha_0 \nabla f(x_{-1})$

**for** $k = 0, 1, \ldots$ **do**

  Sample $\sigma(k) \sim \text{Uniform}\{1, 2, \ldots, s\}$

  **for** $n = 1, 2, \ldots, N$ **do**

    **if** $n = 1$ **then**

      $\alpha_{Nk+1}^{\sigma(k)} \leftarrow \min\Big\{\alpha_{Nk}^{\sigma(k)}\sqrt{1+\theta_{Nk}^{(k)}},\ \frac{\big\| [x_{Nk}]_{\sigma(k)} - [x_{Nk'+N-1}]_{\sigma(k)} \big\|}{2\big\| [\nabla f(x_{Nk})]_{\sigma(k)} - [\nabla f(x_{Nk'+N-1})]_{\sigma(k)} \big\|}\Big\}$ (see text for $k'$)

    **else**

      $\alpha_{Nk+n}^{\sigma(k)} \leftarrow \min\Big\{\alpha_{Nk+n-1}^{\sigma(k)}\sqrt{1+\theta_{Nk+n-1}^{(k)}},\ \frac{\big\| [x_{Nk+n-1}]_{\sigma(k)} - [x_{Nk+n-2}]_{\sigma(k)} \big\|}{2\big\| [\nabla f(x_{Nk+n-1})]_{\sigma(k)} - [\nabla f(x_{Nk+n-2})]_{\sigma(k)} \big\|}\Big\}$

    **end if**

    **for** $i \in \mathcal{S}_{\sigma(k)}$ **do**

      $[x_{Nk+n}]_i \leftarrow [x_{Nk+n-1}]_i - \alpha_{Nk+n}^{\sigma(k)} [\nabla f(x_{Nk+n-1})]_i$

    **end for**

    $\theta_{Nk+n}^{(k)} \leftarrow \frac{\alpha_{Nk+n}^{\sigma(k)}}{\alpha_{Nk+n-1}^{\sigma(k)}}$

  **end for**

  $\alpha_{N(k+1)}^{\sigma(j)} \leftarrow \alpha_{Nk}^{\sigma(j)}$ and $\theta_{N(k+1)}^{\sigma(j)} \leftarrow \theta_{Nk}^{\sigma(j)}$ for $j = 1, 2, \ldots, s$ except $k$

**end for**

Algorithm 1 first performs a full gradient step and then starts an outer loop, indexed by $k$, nesting an inner loop of $N$ iterations, indexed by $n$. For $k = 0, 1, 2, \ldots$, we sample a block $\sigma(k)$ from $\{1, 2, \ldots, s\}$ with equal probability. Then, the selected block is updated $N$ times in the following iterations:

$$
\begin{aligned}
[x_{Nk+1}]_{\sigma(k)} &\leftarrow [x_{Nk}]_{\sigma(k)} - \alpha_{Nk+1}^{\sigma(k)} [\nabla f(x_{Nk})]_{\sigma(k)} && \text{for } n = 1, \\
[x_{Nk+n}]_{\sigma(k)} &\leftarrow [x_{Nk+n-1}]_{\sigma(k)} - \alpha_{Nk+n}^{\sigma(k)} [\nabla f(x_{Nk+n-1})]_{\sigma(k)} && \text{for } n = 2, \ldots, N,
\end{aligned}
$$

where the computation of the step sizes $\alpha$ involve a self-update mechanism and a local smooth approximation $1/(2L)$. For $n = 2, \ldots, N$, it is straightforward to set the local smoothness of $f$ over the *selected* block as

$$
\frac{\big\| [\nabla f(x_{Nk+n})]_{\sigma(k)} - [\nabla f(x_{Nk+n-1})]_{\sigma(k)} \big\|}{\big\| [x_{Nk+n}]_{\sigma(k)} - [x_{Nk+n-1}]_{\sigma(k)} \big\|}.
$$

When $n = 1$, consider $k > 1$ and $\sigma(k-1) \neq \sigma(k)$, that is when we have just chosen a new block to update. Since the latest update of $[x]_{\sigma(k)}$ have occurred a while ago, either in the initial full gradient

step and in the last inner iteration of some outer iteration in the past, we introduce

$$k' = k'(\sigma(k)) \triangleq \begin{cases} \text{the index of the outer iteration when } \sigma(k) \text{ was previously sampled,} & \text{if it exists} \\ -1, & \text{otherwise.} \end{cases}$$

An example to illustrate the definitions of various letters is in the following table:

| current outer itr | $k, k^{(0)}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|
| selected block | $\sigma(k)$ | 2 | 3 | 1 | 3 | 2 | 2 | 3 |
| previous outer itr | $k', k^{(1)}$ | -1 | -1 | -1 | 1 | 0 | 4 | 3 |
| 2nd previous | $k^{(2)}$ | | | -1 | -1 | 0 | | 1 |
| 3rd previous | $k^{(3)}$ | | | | | -1 | | -1 |

At $n = 1$, local smoothness is approximated by

$$\frac{\left\| [\nabla f(x_{Nk})]_{\sigma(k)} - [\nabla f(x_{Nk'+N-1})]_{\sigma(k)} \right\|}{\left\| [x_{Nk}]_{\sigma(k)} - [x_{Nk'+N-1}]_{\sigma(k)} \right\|}, \tag{3}$$

where the iteration index $Nk' + N - 1$ is obviously $-1$, if $k' = -1$, and otherwise is the (cumulative inner) iteration index before the last update is applied to $[x]_{\sigma(k)}$. After the inner loop completes, the last line of the outer loop updates the indices of the step sizes and $\theta$'s of the non-selected blocks without changing their values.

Imagine a fully block-separable objective $f(x) = f_1([x]_1) + \cdots + f_s([x]_s)$. Then, Algorithm 1 reduces to applying the original method in [13] separately to $f_1([x]_1), \ldots, f_s([x]_s)$. Even though the blocks are updated in random orders, every block will pick up from its last status and resume without being affected.

When $f$ is generic, with the dependence between the blocks, the local smoothness of $f$ with respect to a block may also depends on all other blocks. In particular, $[\nabla f(x)]_\sigma$ can change when $[x]_\sigma$ is fixed and other blocks are updated. Therefore, when a new block is selected and $n = 1$, (3) is no longer block-$\sigma(k)$'s local-smoothness approximation. Nevertheless, the computed ratio may still be small, and when this happens, block $\sigma(k)$ will have a relatively large step. Furthermore, theoretical convergence is still ensured as we now present.

For the convenience of presentation, we let $k^{(0)} = k$ and $k^{(1)} = k'$ and define $k^{(2)}, \ldots, k^{(m_k)}$ as the indices of the 2nd, ..., $m_k$-th previous outer iterations when the current block $\sigma(k)$ was selected for update, where $m_k$ is how many times (including the current selection for outer iteration $k$) that block $\sigma(k)$ has been selected so far (thus we must have $k^{(m_k)} = -1$). Since their values depend more directly on $\sigma(k)$, we write their collection as

$$\mathcal{K}_{\sigma(k)} = \left\{ k^{(0)}, k^{(1)}, \ldots, k^{(m_k)} \right\},$$

When analyzing the current outer-iteration $k$, we need to refer to the recent outer-iterations that did not select $\sigma(k)$ to update, for which we use $k_\sigma$, for $\sigma \in \{1, 2, \ldots, s\} \setminus \{\sigma(k)\}$, as the index of outer iteration when block $\sigma$ was most recently updated.

Theorem 3.1 below states an ergodic $\mathcal{O}(\frac{1}{K})$ convergence rate under convexity, and Theorem 3.2 states a linear convergence rate under locally strong convexity.

**Theorem 3.1.** *Consider Algorithm 1 under the assumption that $f\colon \mathbb{R}^p \to \mathbb{R}$ is locally smooth and convex. Let $x^*$ be any solution of (2) and $f^* = f(x^*)$. Then we have $\mathbb{E}\big[\nabla f(x_i)\big] \to 0$ and for all $K$,*

$$\mathbb{E}\big[f(\widetilde{x}_{NK+N}) - f^*\big] \le \frac{\frac{1}{2s}\mathbb{E}\left[\| x_0 - x^* \|^2\right] + \frac{1}{4s}\mathbb{E}\left[\| x_0 - x_{-1} \|^2\right] + \mathbb{E}\left[\alpha_1^\sigma \theta_1^\sigma (f(x_{-1}) - f^*)\right]}{\mathbb{E}\big[C_{NK+N}\big]} = \mathcal{O}\Big(\frac{1}{K}\Big),$$

*where*

$$C_{NK+N} = \alpha^{\sigma}_{NK+N}(1+\theta^{\sigma}_{NK+N}) + \sum_{i=NK}^{NK+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1}) - \alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})\dots$$
$$+ \sum_{m=1}^{m_K}\left((\alpha^{\sigma}_{Nk^{(m)}+N}(1+\theta^{\sigma}_{Nk^{(m)}+N}) - \alpha^{\sigma}_{Nk^{(m-1)}+1}\theta^{\sigma}_{Nk^{(m-1)}+1}) + \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1}) - \alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})\right),$$

*and*

$$\widetilde{x}_{NK+N} = \left(\alpha^{\sigma}_{NK+N}(1+\theta^{\sigma}_{NK+N})x_{NK+N-1} + \sum_{i=NK}^{NK+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1}) - \alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})x_i\dots\right.$$
$$+ \sum_{m=1}^{m_K}\left((\alpha^{\sigma}_{Nk^{(m)}+N}(1+\theta^{\sigma}_{Nk^{(m)}+N}) - \alpha^{\sigma}_{Nk^{(m-1)}+1}\theta^{\sigma}_{Nk^{(m-1)}+1})x_{Nk^{(m)}+N-1}\dots\right.$$
$$\left.\left. + \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1}) - \alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})x_i\right)\right) \Big/ C_{NK+N}.$$

**Theorem 3.2.** *Consider Algorithm 1 under the assumption that $f\colon \mathbb{R}^p \to \mathbb{R}$ is locally smooth and strongly convex. Then we have*

$$\mathbb{E}\big[\,\|\,x_{Nk+n} - x^*\,\|^2\,\big] \leq (1-\frac{1}{4\kappa})^{k-1}M,$$

*where the constants $\kappa \coloneqq \frac{L}{\mu} > 1$, $L$ and $\mu$ are local smoothness and local strong convexity constant on $\overline{Conv}\{x^*, x_0, \dots\}$, and*

$$M = \mathbb{E}\big[\,\|\,x_1 - x^*\,\|^2\,\big] + \frac{s}{2}\left(1+\frac{2\mu}{L}\right)\mathbb{E}\big[\,\|\,x_1 - x_0\,\|^2\,\big] + 2s\mathbb{E}\big[\alpha^{\sigma}_0(1+\theta^{\sigma}_0)(f(x_0)-f_*)\big].$$

### 3.3 A-BCPG-n

**Algorithm 2** A-BCPG-n

**Input:** $x_{-1} \in \mathbb{R}^p$, $\alpha_0 > 0$, $\theta_0^\sigma = +\infty$ for $\sigma = 1, 2, \ldots, s$, $N \geq 2$

$x_0 \leftarrow x_{-1} - \alpha_0 \nabla f(x_{-1})$

**for** $k = 0, 1, \ldots$ **do**

Sample $\sigma(k) \sim \text{Uniform}\{1, 2, \ldots, s\}$

**for** $n = 1, 2, \ldots, N$ **do**

**if** $n = 1$ **then**

$\alpha_{Nk+1}^{\sigma(k)} \leftarrow \min\left\{\alpha_{Nk}^{\sigma(k)}\sqrt{1+\theta_{Nk}^{(k)}},\ \frac{\left\| [x_{Nk}]_{\sigma(k)} - [x_{Nk'+N-1}]_{\sigma(k)} \right\|}{2\left\| [\nabla f(x_{Nk})]_{\sigma(k)} - [\nabla f(x_{Nk'+N-1})]_{\sigma(k)} \right\|}\right\}$

**else**

$\alpha_{Nk+n}^{\sigma(k)} \leftarrow \min\left\{\alpha_{Nk+n-1}^{\sigma(k)}\sqrt{1+\theta_{Nk+n-1}^{(k)}},\ \frac{\left\| [x_{Nk+n-1}]_{\sigma(k)} - [x_{Nk+n-2}]_{\sigma(k)} \right\|}{2\left\| [\nabla f(x_{Nk+n-1})]_{\sigma(k)} - [\nabla f(x_{Nk+n-2})]_{\sigma(k)} \right\|}\right\}$

**end if**

**for** $i \in \mathcal{S}_{\sigma(k)}$ **do**

$[x_{Nk+n}]_i \leftarrow \text{Prox}_{\alpha_{Nk+n}^{\sigma(k)} g}\left([x_{Nk+n-1}]_i - \alpha_{Nk+n}^{\sigma(k)}[\nabla f(x_{Nk+n-1})]_i\right)$

**end for**

$\theta_{Nk+n}^{(k)} \leftarrow \frac{\alpha_{Nk+n}^{\sigma(k)}}{\alpha_{Nk+n-1}^{\sigma(k)}}$

**end for**

$\alpha_{N(k+1)}^{\sigma(j)} \leftarrow \alpha_{Nk}^{\sigma(j)}$ and $\theta_{N(k+1)}^{\sigma(j)} \leftarrow \theta_{Nk}^{\sigma(j)}$ for $j = 1, 2, \ldots, s$ except $k$

**end for**

We now introduce adaptive variant of BCPG to handle two-term composite optimization problems. A-BCPG-n is presented in Algorithm 2, which is an extension of A-BCGD-n that can handle an additional non-smooth term $g$; when $g \equiv 0$, A-BCPG-n is reduced to A-BCGD-n. We show that Algorithm 2 achieves an ergodic $\mathcal{O}(\frac{1}{K})$ convergence rate under convexity (Theorem 3.3). Note that when $g \equiv 0$, Theorem 3.3 is reduced to Theorem 3.1. Although the form of the bounds are just the same, the analysis of Theorem 3.3 is not a straightforward generalization of Theorem 3.1. Thus, we present both of the theorems in the paper.

**Theorem 3.3.** *Consider Algorithm 2 under the assumption that $f\colon \mathbb{R}^p \to \mathbb{R}$ is locally smooth and convex; $g\colon \mathbb{R}^p \to \mathbb{R}$ is closed, convex, proper. Let $x^*$ be any solution of (2) and $h^* := f^* + g^* = f(x^*) + g(x^*)$. Then it holds that for all $K$,*

$$\mathbb{E}\left[h(\widetilde{x}_{NK+N}) - h^*\right] \leq \frac{\frac{1}{2s}\mathbb{E}\left[\left\| x_0 - x^* \right\|^2\right] + \frac{1}{4s}\mathbb{E}\left[\left\| x_0 - x_{-1} \right\|^2\right] + \mathbb{E}\left[\alpha_1^\sigma \theta_1^\sigma (h(x_{-1}) - h^*)\right]}{\mathbb{E}\left[C_{NK+N}\right]} = \mathcal{O}\left(\frac{1}{K}\right).$$

*where*

$$C_{NK+N} = \alpha_{NK+N}^\sigma (1 + \theta_{NK+N}^\sigma) + \sum_{i=NK}^{NK+N-2} (\alpha_{i+1}^\sigma (1 + \theta_{i+1}^\sigma) - \alpha_{i+2}^\sigma \theta_{i+2}^\sigma) \ldots$$

$$+\sum_{m=1}^{m_K}\left((\alpha^{\sigma}_{Nk^{(m)}+N}(1+\theta^{\sigma}_{Nk^{(m)}+N})-\alpha^{\sigma}_{Nk^{(m-1)}+1}\theta^{\sigma}_{Nk^{(m-1)}+1})+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})\right),$$

*and*

$$\widetilde{x}_{NK+N}=\Bigg(\alpha^{\sigma}_{NK+N}(1+\theta^{\sigma}_{NK+N})x_{NK+N-1}+\sum_{i=NK}^{NK+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})x_i\ldots$$
$$+\sum_{m=1}^{m_K}\Bigg((\alpha^{\sigma}_{Nk^{(m)}+N}(1+\theta^{\sigma}_{Nk^{(m)}+N})-\alpha^{\sigma}_{Nk^{(m-1)}+1}\theta^{\sigma}_{Nk^{(m-1)}+1})x_{Nk^{(m)}+N-1}\cdots$$
$$+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})x_i\Bigg)\Bigg)\Bigg/C_{NK+N}.$$

# 4 Numerical Experiments

In the experiments, we consider

$$\text{Logistic Regression: } \min_x \frac{1}{m}\sum_{i=1}^{m}\log\left(1+\exp\left(-(a_i^Tx)b_i\right)\right),$$
$$\text{Sparse Logistic Regression: } \min_x \frac{1}{m}\sum_{i=1}^{m}\log\left(1+\exp\left(-(a_i^Tx)b_i\right)\right)+\lambda\left\|x\right\|_1,$$

where $(a_i,b_i)\in\mathbb{R}^p\times\{-1,1\}$, and report results on standard machine learning datasets `ijcnn`, `a9a`, `mushrooms`. Throughout the experiments, the number of block is set to be 4 for `ijcnn`, and 5 for `a9a` and `mushrooms`; and the criterion is set to be $f(x_k)-f^*<$ `1e-6`, where $f^*$ is computed via gradient descent. For each experiment, all the algorithms are initialized at the same position drawn from standard normal distribution.

**Comparison 1 (C1):** We first investigate how the choice of $N$ influences the performances of A-BCGD-n (A-BCPG-n). We set the initial stepsize to be $\alpha_0=$ `1e-6` to ensure that $x_1$ will be close enough to $x_0$ and likely will give a good estimate for the local smoothness. As illustrated in Figure 1 and 2, smaller $N$ achieves faster convergence across different instances. We suggest practitioners choose $N$ to be 2, if they do not want to conduct comparisons similar to **C1**.

**Comparison 2 (C2):** We then compare A-BCGD-n (A-BCPG-n) with BCGD (BCPG). BCGD (BCPG) adopts shuffled cyclic selection rule, i.e. given $\mathcal{S}$, a permutation of $\{1,2,\ldots,s\}$, the block is selected by

$$\sigma(k)=\mathcal{S}\left[\texttt{mod}(k,s)+1\right],\ \text{for } k=0,1,2,\ldots$$

where $\mathcal{S}[k]$ refers to the $k$-th element of $\mathcal{S}$. Such selection rule empirically shows faster convergence, compared with IID random selection, and cyclic selection rule. We set the stepsize of shuffled BCGD (BCPG) to be $\frac{2}{L}$, $\frac{5}{L}$, $\frac{10}{L}$, $\frac{25}{L}$, $\frac{50}{L}$, where $L$ is computed by

$$L=\frac{1}{4m}\left\|A\right\|^2,\ A=(a_1,a_2,\ldots,a_m)^T.$$

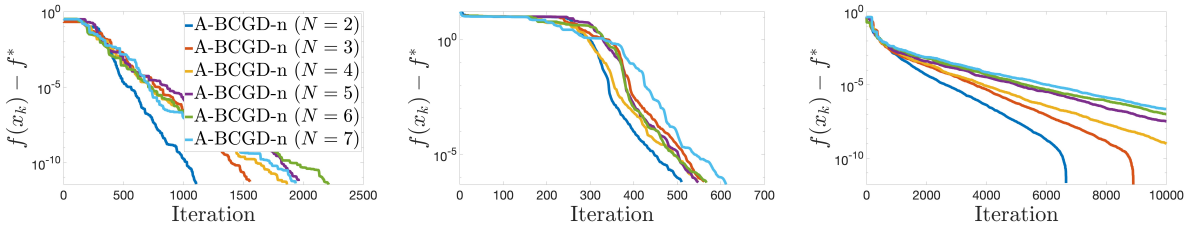


Figure 1: Results of logistic regression (C1). From left to right: (1) `ijcnn`; (2) `a9a`; (3) `mushrooms`.

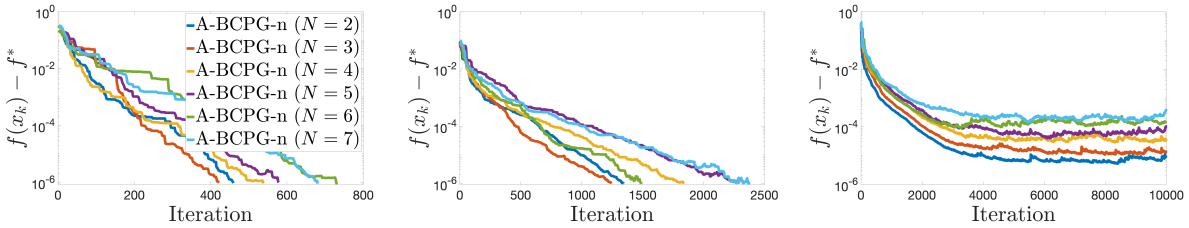


Figure 2: Results of sparse logistic regression (C1). From left to right: (1) `ijcnn`; (2) `a9a`; (3) `mushrooms`.

We set $N$ of A-BCGD-n (A-BCPG-n) according to the results of **C1**. Considering the path of block-coordinate methods may influence the convergence, we run each shuffled BCGD (BCPG) 10 times with different permutations of $\{1, 2, \ldots, s\}$, and report the curve of the fastest convergence. As illustrated in Figure 3 and 4, A-BCGD-n (A-BCPG-n) achieves much faster convergence speed, compared with the theoretical uppper bound $\frac{2}{L}$. It is possible that carefully tuned stepsize makes BCGD (BCPG) converge faster than A-BCGD-n (A-BCPG-n). However, such tuned stepsize is not robust across models and datasets, and thus, getting it requires a large amount of repeated runs.

**Comparison 3 (C3):** We finally compare A-BCGD-n (A-BCPG-n) with accelerated alternatives: (1) shuffled block-coordinate variant of Nesterov's accelerated method (Algorithm 3), which is extended from the method of full coordinate update in [15]; (2) shuffled BCGD (BCPG) with Armijo's linesearch (Algorithm 4), which is extended from the method of full coordinate update in Section 6 of [14]; (3) cyclic and shuffled heuristics of A-BCGD-n (A-BCPG-n). Guided by **C2**, we set the stepsize of Algorithm 3 to be $\frac{1}{L}$, $\frac{2}{L}$ and $\frac{5}{L}$. Instead of fixing $\alpha_{init}$ throughout the iterations, we consider an adaptive and faster variant of Armijo's linesearch strategy, i.e.

$$\begin{aligned} &\alpha_0 \text{ is initialized to be } \frac{500}{L} \\ &\alpha_k \text{ is initialized to be } 1.2\alpha_{k-1} \text{ for } k = 1, 2, \ldots \end{aligned}$$

and set $s$ to be 0.8. Such parameter choices are guided by the numerical experiments of [14]. Since repeated evaluations of $f$ and $\text{Prox}_g$ within linesearch incur additional computational costs, we count each backtracking step as an iteration. Different from **C2**, we only run each shuffled block-coordinate method once, and believe it would result in a more fair comparison. As illustrated in Figure 5, A-BCGD-n (A-BCPG-n) converges faster than Algorithm 4, and cyclic and shuffled heuristics of A-BCGD-n (A-BCPG-n) can even beat Algorithm 3.

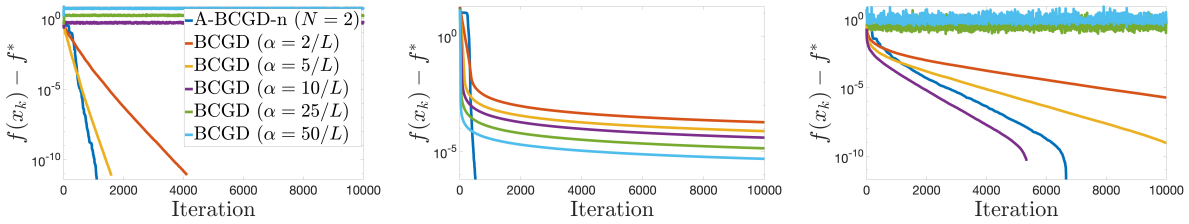


Figure 3: Results of logistic regression (C2). From left to right: (1) `ijcnn`; (2) `a9a`; (3) `mushrooms`.

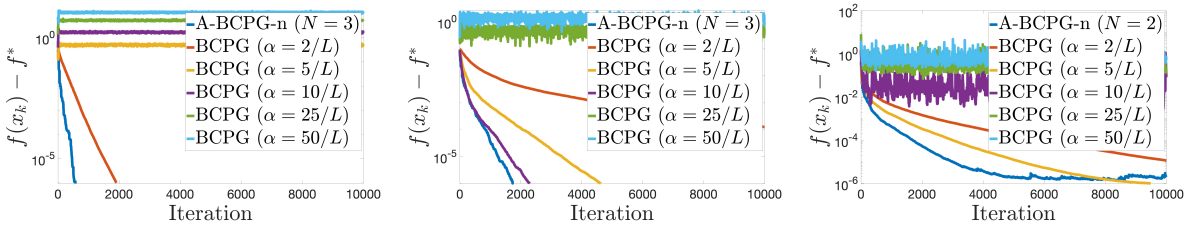


Figure 4: Results of sparse logistic regression (C2). From left to right: (1) `ijcnn`; (2) `a9a`; (3) `mushrooms`.

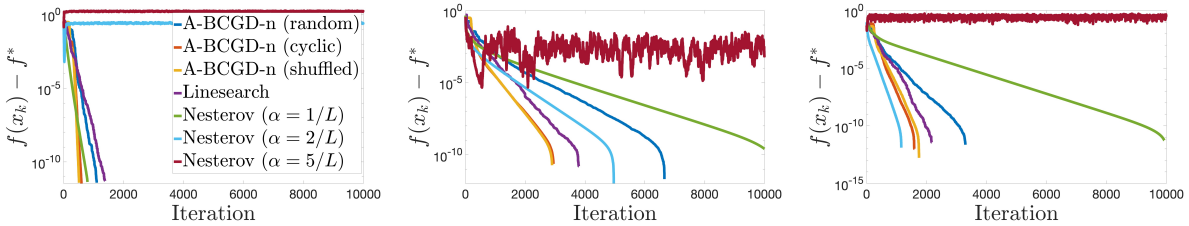


Figure 5: Results of logistic regression (C3). From left to right: (1) `ijcnn`; (2) `a9a`; (3) `mushrooms`.

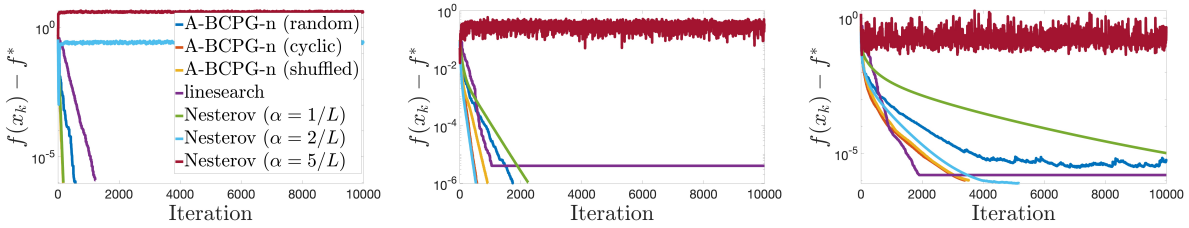


Figure 6: Results of sparse logistic regression (C3). From left to right: (1) `ijcnn`; (2) `a9a`; (3) `mushrooms`.

**Algorithm 3** Shuffled block-coordinate variant of Nesterov's accelerated method

**Input:** $x_0 = y_0 \in \mathbb{R}^p$, $\alpha > 0$

**for** $k = 0, 1, \ldots$ **do**

  $\sigma(k) \leftarrow \mathcal{S}\left[\texttt{mod}(k, s) + 1\right]$

  **for** $i \in \mathcal{S}_{\sigma(k)}$ **do**

    $[x_{k+1}]_i \leftarrow \text{Prox}_{\alpha g}\left([y_k]_i - \alpha[\nabla f(y_k)]_i\right)$

    $[y_{k+1}]_i \leftarrow [x_{k+1}]_i - \frac{k-1}{k+2}[x_{k+1} - x_k]_i$

  **end for**

**end for**

**Algorithm 4** Shuffled BCGD (BCPG) with Armijo's linesearch

**Input:** $x_0 = y_0 \in \mathbb{R}^p$

**for** $k = 0, 1, \ldots$ **do**

  $\sigma(k) \leftarrow \mathcal{S}\left[\texttt{mod}(k, s) + 1\right]$

  $\alpha_k \leftarrow \alpha_{init}$

  **for** $i \in \mathcal{S}_{\sigma(k)}$ **do**

    $[x_{k+1}]_i \leftarrow \text{Prox}_{\alpha_k g}\left([x_k]_i - \alpha_k[\nabla f(x_k)]_i\right)$

  **end for**

  **while** $f(x_{k+1}) > f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\alpha_k} \| x_{k+1} - x_k \|^2$ **do**

    $\alpha_k \leftarrow s\alpha_k$

    **for** $i \in \mathcal{S}_{\sigma(k)}$ **do**

      $[x_{k+1}]_i \leftarrow \text{Prox}_{\alpha_k g}\left([x_k]_i - \alpha_k[\nabla f(x_k)]_i\right)$

    **end for**

  **end while**

**end for**

# 5 Conclusion

In this work, we address a prevalent challenge in the domain BCD methods: the intricacies of stepsize tuning. By introducing adaptive variants, A-BCGD-n and A-BCPG-n, we showcase their adaptability and fast convergence capabilities. Supported by rigorous theoretical guarantees, these methods exhibit compelling performances. As we look to the future, we aim to develop provable guarantees for the cyclic and shuffled heuristics of A-BCGD-n and A-BCPG-n, especially considering their promising performances in numerical experiments. Additionally, we intend to address another intriguing challenge: establishing theoretical guarantees for the proposed methods in nonconvex settings.

# References


[1] Jonathan Barzilai and Jonathan M Borwein. "Two-point step size gradient methods". In: *IMA journal of numerical analysis* 8.1 (1988), pp. 141–148.

[2] Inderjit Dhillon, Pradeep Ravikumar, and Ambuj Tewari. "Nearest neighbor based greedy coordinate descent". In: *Advances in Neural Information Processing Systems* 24 (2011).

[3] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7 (2011).

[4] Jeff Haochen and Suvrit Sra. "Random shuffling beats sgd after finite epochs". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2624–2633.

[5] Majid Jahani et al. "Doubly adaptive scaled algorithm for machine learning using second-order information". In: *arXiv preprint arXiv:2109.05198* (2021).

[6] Junhyung Lyle Kim et al. "Adaptive Federated Learning with Auto-Tuned Clients". In: *arXiv preprint arXiv:2306.11201* (2023).

[7] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[8] Puya Latafat et al. "AdaBiM: An adaptive proximal gradient method for structured convex bilevel optimization". In: *arXiv preprint arXiv:2305.03559* (2023).

[9] Puya Latafat et al. "Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient". In: *arXiv preprint arXiv:2301.04431* (2023).

[10] Ching-Pei Lee and Stephen J Wright. "Random permutations fix a worst case for cyclic coordinate descent". In: *IMA Journal of Numerical Analysis* 39.3 (2019), pp. 1246–1275.

[11] Zhening Li, André Uschmajew, and Shuzhong Zhang. "On convergence of the maximum block improvement method". In: *SIAM Journal on Optimization* 25.1 (2015), pp. 210–233.

[12] Yura Malitsky. "Golden ratio algorithms for variational inequalities". In: *Mathematical Programming* 184.1-2 (2020), pp. 383–410.

[13] Yura Malitsky and Konstantin Mishchenko. "Adaptive gradient descent without descent". In: *arXiv preprint arXiv:1910.09529* (2019).

[14] Yura Malitsky and Konstantin Mishchenko. "Adaptive Proximal Gradient Method for Convex Optimization". In: *arXiv preprint arXiv:2308.02261* (2023).

[15] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/k^2)$". In: *Soviet Mathematics Doklady* 27.2 (1983), pp. 372–376.

[16] Zhimin Peng et al. "Coordinate friendly structures, algorithms and applications". In: *arXiv preprint arXiv:1601.00863* (2016).

[17] Boris Teodorovich Polyak. "Minimization of unsmooth functionals". In: *USSR Computational Mathematics and Mathematical Physics* 9.3 (1969), pp. 14–29.

[18] Peter Richtárik and Martin Takáč. "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function". In: *Mathematical Programming* 144.1-2 (2014), pp. 1–38.

[19] Ankan Saha and Ambuj Tewari. "On the nonasymptotic convergence of cyclic coordinate descent methods". In: *SIAM Journal on Optimization* 23.1 (2013), pp. 576–601.

[20] Robin M Schmidt, Frank Schneider, and Philipp Hennig. "Descending through a crowded valley-benchmarking deep learning optimizers". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9367–9376.

[21] Hao-Jun Michael Shi et al. "A primer on coordinate descent algorithms". In: *arXiv preprint arXiv:1610.00040* (2016).

[22] Ruoyu Sun and Mingyi Hong. "Improved iteration complexity bounds of cyclic block coordinate descent for convex problems". In: *Advances in Neural Information Processing Systems* 28 (2015).

[23] Ruoyu Sun, Zhi-Quan Luo, and Yinyu Ye. "On the efficiency of random permutation for ADMM and coordinate descent". In: *Mathematics of Operations Research* 45.1 (2020), pp. 233–271.

[24] Ruoyu Sun and Yinyu Ye. "Worst-case complexity of cyclic coordinate descent: O (nˆ 2) O (n 2) gap with randomized version". In: *Mathematical Programming* 185 (2021), pp. 487–520.

[25] Tijmen Tieleman, Geoffrey Hinton, et al. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2 (2012), pp. 26–31.

[26] Maria-Luiza Vladarean, Yura Malitsky, and Volkan Cevher. *A first-order primal-dual method with adaptivity to local smoothness*. 2021. DOI: 10.48550/ARXIV.2110.15148. URL: https://arxiv.org/abs/2110.15148.

[27] Stephen Wright and Ching-pei Lee. "Analyzing random permutations for cyclic coordinate descent". In: *Mathematics of computation* 89.325 (2020), pp. 2217–2248.

[28] Stephen J Wright. "Coordinate descent algorithms". In: *Mathematical programming* 151.1 (2015), pp. 3–34.

# A Proofs

## A.1 Proof of Theorem 3.1

**Theorem 3.1.** *Consider Algorithm 1 under the assumption that $f\colon \mathbb{R}^p \to \mathbb{R}$ is locally smooth and convex. Let $x^*$ be any solution of (2) and $f^* = f(x^*)$. Then we have $\mathbb{E}\big[\nabla f(x_i)\big] \to 0$ and for all $K$,*

$$\mathbb{E}\big[f(\widetilde{x}_{NK+N}) - f^*\big] \leq \frac{\frac{1}{2s}\mathbb{E}\left[\| x_0 - x^* \|^2\right] + \frac{1}{4s}\mathbb{E}\left[\| x_0 - x_{-1} \|^2\right] + \mathbb{E}\left[\alpha_1^\sigma \theta_1^\sigma (f(x_{-1}) - f^*)\right]}{\mathbb{E}\big[C_{NK+N}\big]} = \mathcal{O}\Big(\frac{1}{K}\Big),$$

*where*

$$C_{NK+N} = \alpha^\sigma_{NK+N}(1+\theta^\sigma_{NK+N}) + \sum_{i=NK}^{NK+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}) \dots$$
$$+ \sum_{m=1}^{m_K} \Bigg( (\alpha^\sigma_{Nk^{(m)}+N}(1+\theta^\sigma_{Nk^{(m)}+N}) - \alpha^\sigma_{Nk^{(m-1)}+1}\theta^\sigma_{Nk^{(m-1)}+1}) + \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}) \Bigg),$$

*and*

$$\widetilde{x}_{NK+N} = \Bigg( \alpha^\sigma_{NK+N}(1+\theta^\sigma_{NK+N}) x_{NK+N-1} + \sum_{i=NK}^{NK+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}) x_i \dots$$
$$+ \sum_{m=1}^{m_K} \Bigg( (\alpha^\sigma_{Nk^{(m)}+N}(1+\theta^\sigma_{Nk^{(m)}+N}) - \alpha^\sigma_{Nk^{(m-1)}+1}\theta^\sigma_{Nk^{(m-1)}+1}) x_{Nk^{(m)}+N-1} \dots$$
$$+ \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}) x_i \Bigg) \Bigg) \Big/ C_{NK+N}.$$

*Proof.* For a fixed $k = 0, 1, \dots$ with the selected block $\sigma$, we have (7) derived in Lemma A.1:

$$\| [x_{Nk+N}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk+N}]_\sigma - [x_{Nk+N-1}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+N}(1+\theta^\sigma_{Nk+N})(f(x_{Nk+N-1}) - f^*)$$
$$+ 2 \sum_{i=Nk}^{Nk+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2})(f(x_i) - f^*)$$
$$\leq \| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+1}\theta^\sigma_{Nk+1}(f(x_{Nk'+N-1}) - f^*).$$

We then telescope the inequality (7) with all $k \in \mathcal{K}_{\sigma(K_\sigma)}$, and get the following inequality for each $\sigma$:

$$\| [x_{NK_\sigma+N}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{NK_\sigma+N}]_\sigma - [x_{NK_\sigma+N-1}]_\sigma \|^2 + 2\alpha^\sigma_{NK_\sigma+N}(1+\theta^\sigma_{NK_\sigma+N})(f(x_{NK_\sigma+N-1}) - f^*)$$
$$+ 2 \sum_{i=NK_\sigma}^{NK_\sigma+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2})(f(x_i) - f^*)$$
$$+ 2 \sum_{m=1}^{m_{K_\sigma}} \Big( (\alpha^\sigma_{Nk^{(m)}+N}(1+\theta^\sigma_{Nk^{(m)}+N}) - \alpha^\sigma_{Nk^{(m-1)}+1}\theta^\sigma_{Nk^{(m-1)}+1})(f(x_{Nk^{(m)}+N-1}) - f^*) \dots$$

$$+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})(f(x_i)-f^*)\Bigg)$$
$$\leq \left\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x^*]_\sigma \right\|^2 + \frac{1}{2}\left\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x_{Nk^{(m_{K_\sigma})}-1}]_\sigma \right\|^2$$
$$+ 2\alpha_{Nk^{(m_{K_\sigma})}+1}^{\sigma}\theta_{Nk^{(m_{K_\sigma})}+1}^{\sigma}(f(x_{Nk^{(m_{K_\sigma})}-1})-f^*). \tag{4}$$

Since the second, third, and fourth lines are always nonnegative by the definition of the adaptive stepsize, we have the sequence $\{x_k\}_k$ is bounded. Since $\nabla f$ is locally Lipschitz, it is Lipschitz continuous on bounded sets. Therefore, for the set $\mathcal{C} = \overline{\text{Conv}}\{x^*, x_0, \dots\}$, which is bounded as the convex hull of bounded points, there exists $L > 0$ such that

$$\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \quad \forall x, y \in \mathcal{C}.$$

We now prove the algorithm achieves an ergodic $\mathcal{O}(\frac{1}{K})$ convergence rate. Dropping the first two terms on the first row of (4), and multiplying $\frac{1}{2}$, we have

$$\alpha_{NK_\sigma+N}^{\sigma}(1+\theta_{NK_\sigma+N}^{\sigma})(f(x_{NK_\sigma+N-1})-f^*)+\sum_{i=NK_\sigma}^{NK_\sigma+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})(f(x_i)-f^*)$$
$$+\sum_{m=1}^{m_{K_\sigma}}\Bigg((\alpha_{Nk^{(m)}+N}^{\sigma}(1+\theta_{Nk^{(m)}+N}^{\sigma})-\alpha_{Nk^{(m-1)}+1}^{\sigma}\theta_{Nk^{(m-1)}+1}^{\sigma})(f(x_{Nk^{(m)}+N-1})-f^*)\dots$$
$$+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})(f(x_i)-f^*)\Bigg)$$
$$\leq \frac{1}{2}\left\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x^*]_\sigma \right\|^2 + \frac{1}{4}\left\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x_{Nk^{(m_{K_\sigma})}-1}]_\sigma \right\|^2$$
$$+ \alpha_{Nk^{(m_{K_\sigma})}+1}^{\sigma}\theta_{Nk^{(m_{K_\sigma})}+1}^{\sigma}(f(x_{Nk^{(m_{K_\sigma})}-1})-f^*). \tag{5}$$

Applying Jensen's inequality to the LHS of (5), we have

$$\frac{1}{2}\left\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x^*]_\sigma \right\|^2 + \frac{1}{4}\left\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x_{Nk^{(m_{K_\sigma})}-1}]_\sigma \right\|^2 + \alpha_{Nk^{(m_{K_\sigma})}+1}^{\sigma}\theta_{Nk^{(m_{K_\sigma})}+1}^{\sigma}(f(x_{Nk^{(m_{K_\sigma})}-1})-f^*)$$
$$\geq \text{LHS}$$
$$\geq C_{NK_\sigma+N}\big(f(\widetilde{x}_{NK_\sigma+N})-f_*\big)$$

where

$$C_{NK_\sigma+N} = \alpha_{NK_\sigma+N}^{\sigma}(1+\theta_{NK_\sigma+N}^{\sigma})+\sum_{i=NK_\sigma}^{NK_\sigma+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})\dots$$
$$+\sum_{m=1}^{m_{K_\sigma}}\Bigg((\alpha_{Nk^{(m)}+N}^{\sigma}(1+\theta_{Nk^{(m)}+N}^{\sigma})-\alpha_{Nk^{(m-1)}+1}^{\sigma}\theta_{Nk^{(m-1)}+1}^{\sigma})+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})\Bigg)$$
$$> \frac{1}{2L}+0+\left(\sum_{m=1}^{m_{K_\sigma}}1\right)\min_{m=1,\dots,m_{K_\sigma}}\Big\{(\alpha_{Nk^{(m)}+N}^{\sigma}(1+\theta_{Nk^{(m)}+N}^{\sigma})-\alpha_{Nk^{(m-1)}+1}^{\sigma}\theta_{Nk^{(m-1)}+1}^{\sigma})\dots$$

$$+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})\Big\}$$

$$>\Big(\sum_{m=1}^{m_{K_\sigma}}1\Big)\min_{m=1,\dots,m_{K_\sigma}}\Big\{(\alpha_{Nk^{(m)}+N}^{\sigma}(1+\theta_{Nk^{(m)}+N}^{\sigma})-\alpha_{Nk^{(m-1)}+1}^{\sigma}\theta_{Nk^{(m-1)}+1}^{\sigma})\dots$$

$$+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})\Big\}$$

and

$$\widetilde{x}_{NK_\sigma+N}=\Bigg(\alpha_{NK_\sigma+N}^{\sigma}(1+\theta_{NK_\sigma+N}^{\sigma})x_{NK_\sigma+N-1}+\sum_{i=NK_\sigma}^{NK_\sigma+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})x_i\dots$$
$$+\sum_{m=1}^{m_{K_\sigma}}\Bigg((\alpha_{Nk^{(m)}+N}^{\sigma}(1+\theta_{Nk^{(m)}+N}^{\sigma})-\alpha_{Nk^{(m-1)}+1}^{\sigma}\theta_{Nk^{(m-1)}+1}^{\sigma})x_{Nk^{(m)}+N-1}\dots$$
$$+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma})x_i\Bigg)\Bigg)\Bigg/C_{NK_\sigma+N}.$$

Note that $\mathbb{E}\Big[\big(\sum_{m=1}^{m_{K_\sigma}}1\big)\Big]=\frac{K_\sigma}{s}$ implies $\mathbb{E}\big[C_{NK_\sigma+N}\big]=\Omega(K_\sigma)$ and for $\forall\sigma$

$$\mathbb{E}\Big[\text{RHS of (5)}\Big]=\frac{1}{2}\mathbb{E}\Big[\,\|\,[x_{Nk^{(m_{K_\sigma})}}]_\sigma-[x^*]_\sigma\,\|^2\Big]+\frac{1}{4}\mathbb{E}\Big[\,\Big\|\,[x_{Nk^{(m_{K_\sigma})}}]_\sigma-[x_{Nk^{(m_{K_\sigma})}-1}]_\sigma\,\Big\|^2\Big]$$
$$+\mathbb{E}\Big[\alpha_{Nk^{(m_{K_\sigma})}+1}^{\sigma}\theta_{Nk^{(m_{K_\sigma})}+1}^{\sigma}(f(x_{Nk^{(m_{K_\sigma})}-1})-f^*)\Big]$$
$$=\frac{1}{2s}\mathbb{E}\Big[\,\|\,x_{Nk^{(m_{K_\sigma})}}-x^*\,\|^2\Big]+\frac{1}{4s}\mathbb{E}\Big[\,\Big\|\,x_{Nk^{(m_{K_\sigma})}}-x_{Nk^{(m_{K_\sigma})}-1}\,\Big\|^2\Big]$$
$$+\mathbb{E}\Big[\alpha_{Nk^{(m_{K_\sigma})}+1}^{\sigma}\theta_{Nk^{(m_{K_\sigma})}+1}^{\sigma}(f(x_{Nk^{(m_{K_\sigma})}-1})-f^*)\Big]$$

is a constant. We then take full expectation on both sides, and get

$$\mathbb{E}\big[f(\widetilde{x}_{NK_\sigma+N})-f^*\big]$$
$$\leq\frac{\frac{1}{2s}\mathbb{E}\Big[\,\|\,x_{Nk^{(m_{K_\sigma})}}-x^*\,\|^2\Big]+\frac{1}{4s}\mathbb{E}\Big[\,\Big\|\,x_{Nk^{(m_{K_\sigma})}}-x_{Nk^{(m_{K_\sigma})}-1}\,\Big\|^2\Big]+\mathbb{E}\Big[\alpha_{Nk^{(m_{K_\sigma})}+1}^{\sigma}\theta_{Nk^{(m_{K_\sigma})}+1}^{\sigma}(f(x_{Nk^{(m_{K_\sigma})}-1})-f^*)\Big]}{\mathbb{E}\big[C_{NK_\sigma+N}\big]}$$
$$=\mathcal{O}\Big(\frac{1}{K_\sigma}\Big).$$

To simplify the representation of the convergence analysis, we consider $K_\sigma=K$ such that $\sigma(K)=\sigma(0)$ without lost of generality. We then obtain

$$\mathbb{E}\big[f(\widetilde{x}_{NK+N})-f^*\big]\leq\frac{\frac{1}{2s}\mathbb{E}\Big[\,\|\,x_0-x^*\,\|^2\Big]+\frac{1}{4s}\mathbb{E}\Big[\,\|\,x_0-x_{-1}\,\|^2\Big]+\mathbb{E}\left[\alpha_1^\sigma\theta_1^\sigma(f(x_{-1})-f^*)\right]}{\mathbb{E}\big[C_{NK+N}\big]}=\mathcal{O}\Big(\frac{1}{K}\Big).$$

We finally prove $\mathbb{E}\big[\nabla f(x_i)\big] \to 0$. Since $\nabla f$ is Lipschitz continous on $\mathcal{C}$, we have

$$\langle [\nabla f(x_{Nk})]_\sigma, [x^*]_\sigma - [x_{Nk}]_\sigma \rangle \leq f^* - f(x_{Nk}) - \frac{1}{2L} \| [\nabla f(x_{Nk})]_\sigma \|^2 . \tag{6}$$

Recall that the development of Lemma A.1 is delicately constructing upper bounds of $A$ and $B$ respectively for the following inequality,

$$\begin{aligned} \| [x_{Nk+1}]_\sigma - [x^*]_\sigma \|^2 &= \| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 + 2\langle [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma, [x_{Nk}]_\sigma - [x^*]_\sigma \rangle \\ &\quad + \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 \\ &= \| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 + 2\alpha^\sigma_{Nk+1} \underbrace{\langle [\nabla f(x_{Nk})]_\sigma, [x^*]_\sigma - [x_{Nk}]_\sigma \rangle}_{=A} \\ &\quad + \underbrace{\| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2}_{=B}, \end{aligned}$$

Originally in lemma A.1, $A$ is bounded by

$$\langle [\nabla f(x_{Nk})]_\sigma, [x^*]_\sigma - [x_{Nk}]_\sigma \rangle \leq f^* - f(x_{Nk}).$$

Here, we bound $A$ using (6), keep the upper bound of $B$ the same, and then get

$$\mathbb{E}\big[\frac{\alpha^\sigma_{NK+N}}{L} \| \nabla f(x_{NK+N-1}) \|^2 \big] + \sum_{m=0}^{m_K} \Big( \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-1} \mathbb{E}\big[\frac{\alpha^\sigma_{i+1}}{L} \| \nabla f(x_i) \|^2 \big] \Big) \leq \frac{1}{s}\mathbb{E}\big[ \| x_0 - x^* \|^2 \big] + \frac{1}{2}\mathbb{E}\big[ \| x_0 - x_{-1} \|^2 \big] + 2\mathbb{E}\big[\alpha^\sigma_1 \theta^\sigma_1 (f(x_{-1} - f^*)\big],$$

in which, compared with the original inequality of Lemma A.1, additional $\| \nabla f(\cdot) \|$ terms are introduced to LHS. As $\alpha^{\sigma(j)}_{i+1} \geq \frac{1}{2L}$, one has that $\mathbb{E}\big[\nabla f(x_i)\big] \to 0$. □

**Lemma A.1.** *Consider Algorithm 1 under the assumption that $f\colon \mathbb{R}^p \to \mathbb{R}$ is locally smooth and convex. Let $x^*$ be any solution of (2) and $f^* = f(x^*)$. Then for all $k = 0, 1, \ldots$ with the selected block $\sigma$, it holds*

$$\begin{aligned} &\| [x_{Nk+N}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk+N}]_\sigma - [x_{Nk+N-1}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+N}(1 + \theta^\sigma_{Nk+N})(f(x_{Nk+N-1}) - f^*) \\ &\quad + 2 \sum_{i=Nk}^{Nk+N-2} (\alpha^\sigma_{i+1}(1 + \theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2})(f(x_i) - f^*) \\ &\quad \leq \| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+1}\theta^\sigma_{Nk+1}(f(x_{Nk'+N-1}) - f^*). \end{aligned} \tag{7}$$

*Proof.* Note that

$$\begin{aligned} \| [x_{Nk+1}]_\sigma - [x^*]_\sigma \|^2 &= \| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 + 2\langle [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma, [x_{Nk}]_\sigma - [x^*]_\sigma \rangle \\ &\quad + \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 \\ &= \| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 + 2\alpha^\sigma_{Nk+1} \underbrace{\langle [\nabla f(x_{Nk})]_\sigma, [x^*]_\sigma - [x_{Nk}]_\sigma \rangle}_{=A} \\ &\quad + \underbrace{\| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2}_{=B}, \end{aligned} \tag{8}$$

We bound $A$ by convexity of $f$:

$$A = \langle [\nabla f(x_{Nk})]_\sigma, [x^*]_\sigma - [x_{Nk}]_\sigma \rangle \leq f^* - f(x_{Nk}). \tag{9}$$

To bound $B$, we adopt the analysis technique similar to that of Lemma 1 in [13]. We consider:

$$\begin{aligned}
\| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 &= 2 \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 - \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 \\
&= -2\alpha^\sigma_{Nk+1} \langle [\nabla f(x_{Nk})]_\sigma, [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \rangle - \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 \\
&= \underbrace{2\alpha^\sigma_{Nk+1} \langle [\nabla f(x_{Nk})]_\sigma - [\nabla f(x_{Nk'+N-1})]_\sigma, [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \rangle}_{=C} \\
&\quad + \underbrace{2\alpha^\sigma_{Nk+1} \langle [\nabla f(x_{Nk'+N-1})]_\sigma, [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \rangle}_{=D} - \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 .
\end{aligned} \tag{10}$$

We use the definition of the adaptive stepsize, Cauchy-Schwarz and Young's inequalities to bound $C$:

$$\begin{aligned}
C &= 2\alpha^\sigma_{Nk+1} \langle [\nabla f(x_{Nk})]_\sigma - [\nabla f(x_{Nk'+N-1})]_\sigma, [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \rangle \\
&\leq 2\alpha^\sigma_{Nk+1} \| [\nabla f(x_{Nk})]_\sigma - [\nabla f(x_{Nk'+N-1})]_\sigma \| \| [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \| \\
&\leq \| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \| \| [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \| \\
&= \frac{1}{2} \| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \|^2 .
\end{aligned} \tag{11}$$

We then bound $D$ by convexity of $f$,

$$\begin{aligned}
D &= 2\alpha^\sigma_{Nk+1} \langle [\nabla f(x_{Nk'+N-1})]_\sigma, [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \rangle \\
&= \frac{2\alpha^\sigma_{Nk+1}}{\alpha^\sigma_{Nk}} \langle [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma, [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \rangle \\
&= 2\alpha^\sigma_{Nk+1} \theta^\sigma_{Nk+1} \langle [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma, [\nabla f(x_{Nk})]_\sigma \rangle \\
&\leq 2\alpha^\sigma_{Nk+1} \theta^\sigma_{Nk+1} (f(x_{Nk'+N-1}) - f(x_{Nk})),
\end{aligned} \tag{12}$$

Plugging (11) and (12) in (10), we obtain

$$\begin{aligned}
B &= \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 \\
&\leq \frac{1}{2} \| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|^2 - \frac{1}{2} \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+1} \theta^\sigma_{Nk+1} (f(x_{Nk'+N-1}) - f(x_{Nk})).
\end{aligned}$$

We then deduce the desired inequality from (8):

$$\begin{aligned}
&\| [x_{Nk+1}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+1} (1 + \theta^\sigma_{Nk+1}) (f(x_{Nk}) - f^*) \\
&\leq \| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+1} \theta^\sigma_{Nk+1} (f(x_{Nk'+N-1}) - f^*).
\end{aligned} \tag{13}$$

Following the similar derivation above, we have for $n = 2, \ldots, N$,

$$\begin{aligned}
&\| [x_{Nk+n}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk+n}]_\sigma - [x_{Nk+n-1}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+n} (1 + \theta^\sigma_{Nk+n}) (f(x_{Nk+n-1}) - f^*) \\
&\leq \| [x_{Nk+n-1}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{2} \| [x_{Nk+n-1}]_\sigma - [x_{Nk+n-2}]_\sigma \|^2 + 2\alpha^\sigma_{Nk+n} \theta^\sigma_{Nk+n} (f(x_{Nk+n-2}) - f^*).
\end{aligned} \tag{14}$$

Telescoping these inequalities from $n = 1$ to $N$, we obtain (7), where $N \geq 2$ ensures that $Nk + N - 2 \geq Nk + N$. □

## A.2 Proof of Theorem 3.2

**Theorem 3.2.** *Consider Algorithm 1 under the assumption that $f\colon \mathbb{R}^p \to \mathbb{R}$ is locally smooth and strongly convex. Then we have*

$$\mathbb{E}\big[\,\|\,x_{Nk+n} - x^*\,\|^2\,\big] \le (1 - \frac{1}{4\kappa})^{k-1} M,$$

*where the constants $\kappa \coloneqq \frac{L}{\mu} > 1$, $L$ and $\mu$ are local smoothness and local strong convexity constant on $\overline{Conv}\{x^*, x_0, \dots\}$, and*

$$M = \mathbb{E}\big[\,\|\,x_1 - x^*\,\|^2\,\big] + \frac{s}{2}\left(1 + \frac{2\mu}{L}\right)\mathbb{E}\big[\,\|\,x_1 - x_0\,\|^2\,\big] + 2s\mathbb{E}\big[\alpha_0^\sigma(1+\theta_0^\sigma)(f(x_0) - f_*)\big].$$

*Proof.* For proof simplicity, we will use a more conservative bound $\alpha^\sigma_{Nk+n} \le \alpha^\sigma_{Nk+n-1}\sqrt{1 + \frac{\theta^\sigma_{Nk+n-1}}{2}}$, just similar to the analysis in [13]. Note that using $\alpha^\sigma_{Nk+n} \le \alpha^\sigma_{Nk+n-1}\sqrt{1 + \frac{\theta^\sigma_{Nk+n-1}}{2}}$ does not change the statement of Theorem 3.1. Hence, there exist $L > 0$ such that $\nabla f$ is $L$-Lipschitz continuous on $\mathcal{C}$. Because of $\mu$-strong convexity, we have $\|\,\nabla f(x_k) - \nabla f(x_{k-1})\,\| \ge \mu\,\|\,x_k - x_{k-1}\,\|$; thus $\alpha^\sigma_{Nk+n} \le \frac{1}{2\mu}$. We tighten some steps in the analysis to improve bound (9). By strong convexity,

$$\alpha^\sigma_{Nk+n}\langle[\nabla f(x_{Nk+n-1})]_\sigma, [x^*]_\sigma - [x_{Nk+n-1}]_\sigma\rangle \le \alpha^\sigma_{Nk+n}\big(f^* - f(x_{Nk+n-1})\big) - \frac{\mu\alpha^\sigma_{Nk+n}}{2}\|[x_{Nk+n-1}]_\sigma - [x^*]_\sigma\|^2.$$

By $L$-smoothness and bound $\alpha^\sigma_{Nk+n} \le \frac{1}{2\mu}$, we have

$$\begin{aligned}
\alpha^\sigma_{Nk+n}\langle[\nabla f(x_{Nk+n-1})]_\sigma, [x^*]_\sigma - [x_{Nk+n-1}]_\sigma\rangle &\le \alpha^\sigma_{Nk+n}\big(f^* - f(x_{Nk+n-1})\big) - \frac{\alpha^\sigma_{Nk+n}}{2L}\|[\nabla f(x_{Nk+n-1})]_\sigma\|^2 \\
&= \alpha^\sigma_{Nk+n}\big(f^* - f(x_{Nk+n-1})\big) - \frac{1}{2L\alpha^\sigma_{Nk+n}}\|[x_{Nk+n}]_\sigma - [x_{Nk+n-1}]_\sigma\|^2 \\
&\le \alpha^\sigma_{Nk+n}\big(f^* - f(x_{Nk+n-1})\big) - \frac{\mu}{L}\|[x_{Nk+n}]_\sigma - [x_{Nk+n-1}]_\sigma\|^2.
\end{aligned}$$

These two bounds together give us

$$\begin{aligned}
\alpha^\sigma_{Nk+n}\langle[\nabla f(x_{Nk+n-1})]_\sigma, [x^*]_\sigma - [x_{Nk+n-1}]_\sigma\rangle \le \alpha^\sigma_{Nk+n}(f^* - f(x_{Nk+n-1})) &- \frac{\mu\alpha^\sigma_{Nk+n}}{4}\|[x_{Nk+n-1}]_\sigma - [x^*]_\sigma\|^2 \\
&- \frac{\mu}{2L}\|[x_{Nk+n}]_\sigma - [x_{Nk+n-1}]_\sigma\|^2.
\end{aligned}$$

Taking full expectation on both sides, we get

$$\begin{aligned}
\mathbb{E}\big[\alpha^\sigma_{Nk+n}\langle[\nabla f(x_{Nk+n-1})]_\sigma, [x^*]_\sigma - [x_{Nk+n-1}]_\sigma\rangle\big] \le \mathbb{E}\big[\alpha^\sigma_{Nk+n}(f^* - f(x_{Nk+n-1}))\big] &- \frac{\mu}{4s}\mathbb{E}\big[\alpha^\sigma_{Nk+n}\|x_{Nk+n-1} - x^*\|^2\big] \\
&- \frac{\mu}{2Ls}\mathbb{E}\big[\|x_{Nk+n} - x_{Nk+n-1}\|^2\big].
\end{aligned}$$

Recall (13) and (14) under full expectation:

$$\frac{1}{s}\mathbb{E}\big[\,\|\,x_{Nk+1} - x^*\,\|^2\,\big] + \frac{1}{2s}\mathbb{E}\big[\,\|\,x_{Nk+1} - x_{Nk}\,\|^2\,\big] + 2\mathbb{E}\big[\alpha^\sigma_{Nk+1}(1+\theta^\sigma_{Nk+1})(f(x_{Nk}) - f^*)\big]$$

$$\leq \frac{1}{s}\mathbb{E}\big[\,\|\,x_{Nk}-x^*\,\|^2\,\big]+\frac{1}{2s}\mathbb{E}\big[\,\|\,x_{Nk}-x_{Nk'+N-1}\,\|^2\,\big]+2\mathbb{E}\big[\alpha^\sigma_{Nk+1}\theta^\sigma_{Nk+1}(f(x_{Nk'+N-1})-f^*)\big],$$

and

$$\begin{aligned}&\frac{1}{s}\mathbb{E}\big[\,\|\,x_{Nk+n}-x^*\,\|^2\,\big]+\frac{1}{2s}\mathbb{E}\big[\,\|\,x_{Nk+n}-x_{Nk+n-1}\,\|^2\,\big]+2\mathbb{E}\big[\alpha^\sigma_{Nk+n}(1+\theta^\sigma_{Nk+n})(f(x_{Nk+n-1})-f^*)\big]\\ &\leq \frac{1}{s}\mathbb{E}\big[\,\|\,x_{Nk+n-1}-x^*\,\|^2\,\big]+\frac{1}{2s}\mathbb{E}\big[\,\|\,x_{Nk+n-1}-x_{Nk+n-2}\,\|^2\,\big]+2\mathbb{E}\big[\alpha^\sigma_{Nk+n}\theta^\sigma_{Nk+n}(f(x_{Nk+n-2})-f^*)\big],\end{aligned}$$

for $n=2,3,\ldots,N$. These can be modified as

$$\begin{aligned}&\frac{1}{s}\mathbb{E}\big[\,\|\,x_{Nk+1}-x^*\,\|^2\,\big]+\frac{1}{2s}\left(1+\frac{2\mu}{L}\right)\mathbb{E}\big[\,\|\,x_{Nk+1}-x_{Nk}\,\|^2\,\big]+2\mathbb{E}\big[\alpha^\sigma_{Nk+1}(1+\theta^\sigma_{Nk+1})(f(x_{Nk})-f_*)\big]\\ &\leq \frac{1}{s}\mathbb{E}\Big[\left(1-\frac{\alpha^\sigma_{Nk}\mu}{2}\right)\|\,x_{Nk}-x^*\,\|^2\,\Big]+\frac{1}{2}\mathbb{E}\big[\,\|\,x_{Nk}-x_{Nk'+N-1}\,\|^2\,\big]+2\mathbb{E}\big[\alpha^\sigma_{Nk+1}\theta^\sigma_{Nk+1}(f(x_{Nk'+N-1})-f_*)\big]\\ &\leq \frac{1}{s}\mathbb{E}\Big[\left(1-\frac{\alpha^\sigma_{Nk}\mu}{2}\right)\|\,x_{Nk}-x^*\,\|^2\,\Big]+\frac{1}{2}\mathbb{E}\big[\,\|\,x_{Nk}-x_{Nk'+N-1}\,\|^2\,\big]\\ &\quad+2\mathbb{E}\Big[\alpha^\sigma_{Nk}\left(1+\frac{\theta^\sigma_{Nk}}{2}\right)(f(x_{Nk'+N-1})-f_*)\Big],\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{s}\mathbb{E}\big[\,\|\,x_{Nk+n}-x^*\,\|^2\,\big]+\frac{1}{2s}\left(1+\frac{2\mu}{L}\right)\mathbb{E}\big[\,\|\,x_{Nk+n}-x_{Nk+n-1}\,\|^2\,\big]+2\mathbb{E}\big[\alpha^\sigma_{Nk+n}(1+\theta^\sigma_{Nk+n})(f(x_{Nk+n-1})-f_*)\big]\\ &\leq \frac{1}{s}\mathbb{E}\Big[\left(1-\frac{\alpha^\sigma_{Nk+n-1}\mu}{2}\right)\|\,x_{Nk+n-1}-x^*\,\|^2\,\Big]+\frac{1}{2}\mathbb{E}\big[\,\|\,x_{Nk+n-1}-x_{Nk+n-2}\,\|^2\,\big]+2\mathbb{E}\big[\alpha^\sigma_{Nk+n}\theta^\sigma_{Nk+n}(f(x_{Nk+n-2})-f_*)\big]\\ &\leq \frac{1}{s}\mathbb{E}\Big[\left(1-\frac{\alpha^\sigma_{Nk+n-1}\mu}{2}\right)\|\,x_{Nk+n-1}-x^*\,\|^2\,\Big]+\frac{1}{2}\mathbb{E}\big[\,\|\,x_{Nk+n-1}-x_{Nk+n-2}\,\|^2\,\big]\\ &\quad+2\mathbb{E}\big[\alpha^\sigma_{Nk+n-1}\left(1+\frac{\theta^\sigma_{Nk+n-1}}{2}\right)(f(x_{Nk+n-2})-f_*)\big]\end{aligned}$$

for $n=2,3,\ldots,N$. Under the new update we have contraction in every term: $1-\frac{\alpha^\sigma_{Nk+n-1}\mu}{2}$ in the first, $\frac{1}{1+\frac{2\mu}{L}}=1-\frac{2\mu}{L+2\mu}$ in the second and $\frac{1+\frac{\theta^\sigma_{Nk+n-1}}{2}}{1+\theta^\sigma_{Nk+n-1}}=1-\frac{\theta^\sigma_{Nk+n-1}}{2(1+\theta^\sigma_{Nk+n-1})}$ in the last one. Recall that $\alpha^\sigma_{Nk+n-1}\in\left[\frac{1}{2L},\frac{1}{2\mu}\right]$ for $k\geq 1$. Therefore, $\theta^\sigma_{Nk+n}=\frac{\alpha^\sigma_{Nk+n}}{\alpha^\sigma_{Nk+n-1}}\geq\frac{1}{\kappa}$ for any $k>1$, where $\kappa=\frac{L}{\mu}$. Since $\frac{x}{1+x}$ monotonically increases with respect to $x>0$, we have $\frac{\theta^\sigma_{Nk+n-1}}{2\left(1+\theta^\sigma_{Nk+n-1}\right)}\geq\frac{1}{2(\kappa+1)}$ when $k>2$. Since $\frac{\alpha^\sigma_{Nk+n}\mu}{2}\geq\frac{1}{4\kappa}$, $\frac{2\mu}{L+2\mu}=\frac{2}{\kappa+2}\geq\frac{1}{4\kappa}$, and $\frac{1}{2(\kappa+1)}\geq\frac{1}{4\kappa}$, we have

$$\mathbb{E}\big[\,\|\,x_{Nk+n}-x^*\,\|^2\,\big]\leq(1-\frac{1}{4\kappa})^{k-1}M,$$

where $M=\mathbb{E}\big[\,\|\,x_1-x^*\,\|^2\,\big]+\frac{1}{2}\left(1+\frac{2\mu}{L}\right)\mathbb{E}\big[\,\|\,x_1-x_0\,\|^2\,\big]+2s\mathbb{E}\big[\alpha^\sigma_0(1+\theta^\sigma_0)(f(x_0)-f_*)\big]$. □

## A.3 Proof of Theorem 3.3

**Theorem 3.3.** *Consider Algorithm 2 under the assumption that $f\colon\mathbb{R}^p\to\mathbb{R}$ is locally smooth and convex; $g\colon\mathbb{R}^p\to\mathbb{R}$ is closed, convex, proper. Let $x^*$ be any solution of (2) and $h^*:=f^*+g^*=f(x^*)+g(x^*)$.*

*Then it holds that for all $K$,*

$$\mathbb{E}\big[h(\widetilde{x}_{NK+N}) - h^*\big] \le \frac{\frac{1}{2s}\mathbb{E}\left[\| x_0 - x^* \|^2\right] + \frac{1}{4s}\mathbb{E}\left[\| x_0 - x_{-1} \|^2\right] + \mathbb{E}\left[\alpha_1^\sigma \theta_1^\sigma (h(x_{-1}) - h^*)\right]}{\mathbb{E}\big[C_{NK+N}\big]} = \mathcal{O}\Big(\frac{1}{K}\Big).$$

*where*

$$C_{NK+N} = \alpha^\sigma_{NK+N}(1+\theta^\sigma_{NK+N}) + \sum_{i=NK}^{NK+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}) \dots$$
$$+ \sum_{m=1}^{m_K} \Bigg( (\alpha^\sigma_{Nk^{(m)}+N}(1+\theta^\sigma_{Nk^{(m)}+N}) - \alpha^\sigma_{Nk^{(m-1)}+1}\theta^\sigma_{Nk^{(m-1)}+1}) + \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}) \Bigg),$$

*and*

$$\widetilde{x}_{NK+N} = \Bigg( \alpha^\sigma_{NK+N}(1+\theta^\sigma_{NK+N}) x_{NK+N-1} + \sum_{i=NK}^{NK+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}) x_i \dots$$
$$+ \sum_{m=1}^{m_K} \Bigg( (\alpha^\sigma_{Nk^{(m)}+N}(1+\theta^\sigma_{Nk^{(m)}+N}) - \alpha^\sigma_{Nk^{(m-1)}+1}\theta^\sigma_{Nk^{(m-1)}+1}) x_{Nk^{(m)}+N-1} \dots$$
$$+ \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}) x_i \Bigg) \Bigg) \Big/ C_{NK+N}.$$

*Proof.* For a fixed $k = 0, 1, \dots,$ with selected block $\sigma$, we have (17) derived in Lemma A.2:

$$\frac{1}{2}\| [x_{Nk+N}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{4}\| [x_{Nk+N}]_\sigma - [x_{Nk+N-1}]_\sigma \|^2 + \alpha^\sigma_{Nk+N}(1+\theta^\sigma_{Nk+N})(h(x_{Nk+N-1}) - h^*)$$
$$+ \sum_{i=Nk}^{Nk+N-2} \big(\frac{1}{4} + (\theta^\sigma_{i+2})^2 \alpha^\sigma_{i+1} A^\sigma_{i+1}\big(1 - \alpha^\sigma_{i+1} B^\sigma_{i+1}\big)\big) \| [x_{i+1}]_\sigma - [x_i]_\sigma \|^2 + \sum_{i=Nk}^{Nk+N-2} \big(\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2}\big)(h(x_i) - h^*)$$
$$\le \frac{1}{2}\| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{4}\| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|^2 + \alpha^\sigma_{Nk+1}(1+\theta^\sigma_{Nk+1})(h(x_{Nk'+N-1}) - h^*).$$

We then telescope the inequality (17) with all $k \in \mathcal{K}_{\sigma(K_\sigma)}$, and get the following inequality for each $\sigma$

$$\frac{1}{2}\| [x_{NK_\sigma+N}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{4}\| [x_{NK_\sigma+N}]_\sigma - [x_{NK_\sigma+N-1}]_\sigma \|^2 + \alpha^\sigma_{NK_\sigma+N}(1+\theta^\sigma_{NK_\sigma+N})(h(x_{NK_\sigma+N-1}) - h^*)$$
$$+ \sum_{i=NK_\sigma}^{NK_\sigma+N-2} \big(\frac{1}{4} + (\theta^\sigma_{i+2})^2 \alpha^\sigma_{i+1} A^\sigma_{i+1}\big(1 - \alpha^\sigma_{i+1} B^\sigma_{i+1}\big)\big) \| [x_{i+1}]_\sigma - [x_i]_\sigma \|^2$$
$$+ \sum_{m=1}^{m_{K_\sigma}} \Bigg( \big(\frac{1}{4} + (\theta^\sigma_{Nk^{(m-1)}+1})^2 \alpha^\sigma_{Nk^{(m)}+N} A^\sigma_{Nk^{(m)}+N}\big(1 - \alpha^\sigma_{Nk^{(m)}+N} B^\sigma_{Nk^{(m)}+N}\big)\big) \| [x_{Nk^{(m)}+N}]_\sigma - [x_{Nk^{(m)}+N-1}]_\sigma \|^2 \dots$$
$$+ \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2} \big(\frac{1}{4} + (\theta^\sigma_{i+2})^2 \alpha^\sigma_{i+1} A^\sigma_{i+1}\big(1 - \alpha^\sigma_{i+1} B^\sigma_{i+1}\big)\big) \| [x_{i+1}]_\sigma - [x_i]_\sigma \|^2 \Bigg)$$

$$
\begin{aligned}
&+ \sum_{i=NK_\sigma}^{NK_\sigma+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2})(h(x_i) - h^*) \\
&\quad + \sum_{m=1}^{m_{K_\sigma}} \Bigg( (\alpha^\sigma_{Nk^{(m)}+N}(1+\theta^\sigma_{Nk^{(m)}+N}) - \alpha^\sigma_{Nk^{(m-1)}+1}\theta^\sigma_{Nk^{(m-1)}+1})(h(x_{Nk^{(m)}+N-1}) - h^*) \dots \\
&\qquad + \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2})(h(x_i) - h^*) \Bigg) \\
&\qquad \le \frac{1}{2} \left\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x^*]_\sigma \right\|^2 + \frac{1}{4} \left\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x_{Nk^{(m_{K_\sigma})}-1}]_\sigma \right\|^2 \\
&\qquad\quad + \alpha^\sigma_{Nk^{(m_{K_\sigma})}+1}(1+\theta^\sigma_{Nk^{(m_{K_\sigma})}+1})(h(x_{Nk^{(m_{K_\sigma})}-1}) - h^*). \qquad (15)
\end{aligned}
$$

Since

$$
\alpha^\sigma_{Nk+n} \le \sqrt{1+\theta^\sigma_{Nk+n-1}}\alpha^\sigma_{Nk+n-1} \Longrightarrow \alpha^\sigma_{Nk+n-1}(1+\theta^\sigma_{Nk+n-1}) - \alpha^\sigma_{Nk+n}\theta^\sigma_{Nk+n} \ge 0, \text{ for } n = 1, 2, \dots, N,
$$

$$
\begin{aligned}
\alpha^\sigma_{Nk+1} &\le \frac{\| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|}{2 \| [\nabla f(x_{Nk})]_\sigma - [\nabla f(x_{Nk'+N-1})]_\sigma \|} \\
&= \frac{1}{2}\sqrt{\frac{1}{A^\sigma_{Nk}B^\sigma_{Nk}}} \\
&\le \frac{1}{2}\sqrt{\frac{\alpha^\sigma_{Nk}B^\sigma_{Nk}}{A^\sigma_{Nk}B^\sigma_{Nk}\max\{\alpha^\sigma_{Nk}B^\sigma_{Nk} - 1,\ 0\}}} \\
&\Longrightarrow \frac{1}{4} + (\theta^\sigma_{Nk+1})^2\alpha^\sigma_{Nk}A^\sigma_{Nk}\big(1 - \alpha^\sigma_{Nk}B^\sigma_{Nk}\big) \ge 0,
\end{aligned}
$$

where $A^\sigma_{Nk}$ and $B^\sigma_{Nk}$ are defined by $[x_{Nk}]_\sigma$, $[x_{Nk'+N-1}]_\sigma$, $[\nabla f(x_{Nk})]_\sigma$, $[\nabla f(x_{Nk'+N-1})]_\sigma$, and for $n = 2, 3, \dots, N$,

$$
\begin{aligned}
\alpha^\sigma_{Nk+n} &\le \frac{\| [x_{Nk+n-1}]_\sigma - [x_{Nk+n-2}]_\sigma \|}{2 \| [\nabla f(x_{Nk+n-1})]_\sigma - [\nabla f(x_{Nk+n-2})]_\sigma \|} \\
&= \frac{1}{2}\sqrt{\frac{1}{A^\sigma_{Nk+n-1}B^\sigma_{Nk+n-1}}} \\
&\le \frac{1}{2}\sqrt{\frac{\alpha^\sigma_{Nk+n-1}B^\sigma_{Nk+n-1}}{A^\sigma_{Nk+n-1}B^\sigma_{Nk+n-1}\max\{\alpha^\sigma_{Nk+n-1}B^\sigma_{Nk+n-1} - 1,\ 0\}}} \\
&\Longrightarrow \frac{1}{4} + (\theta^\sigma_{Nk+n})^2\alpha^\sigma_{Nk+n-1}A^\sigma_{Nk+n-1}\big(1 - \alpha^\sigma_{Nk+n-1}B^\sigma_{Nk+n-1}\big) \ge 0,
\end{aligned}
$$

we have the second, third, fourth, fifth, sixth, seventh lines above are always nonnegative, we have the sequence $\{x_k\}_k$ is bounded. Since $\nabla f$ is locally Lipschitz, it is Lipschitz continuous on bounded sets. Therefore, for the set $\mathcal{C} = \overline{\text{Conv}}\{x^*, x_0, \dots\}$, which is bounded as the convex hull of bounded points, there exists $L > 0$ such that

$$
\| \nabla f(x) - \nabla f(y) \| \le L \| x - y \| \quad \forall x, y \in \mathcal{C}.
$$

We now prove the algorithm achieves an ergodic $\mathcal{O}(\frac{1}{K})$ convergence rate. Dropping the second, third, fourth rows and the first two terms on the first row of (15), taking expectation on both sides with respect to $\sigma$, we have

$$
\begin{aligned}
&\alpha^{\sigma}_{NK_\sigma+N}(1+\theta^{\sigma}_{NK_\sigma+N})(h(x_{NK_\sigma+N-1})-h^*)\\
&\quad+\sum_{i=NK_\sigma}^{NK_\sigma+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})(h(x_i)-h^*)\\
&\qquad+\sum_{m=1}^{m_{K_\sigma}}\Bigg((\alpha^{\sigma}_{Nk^{(m)}+N}(1+\theta^{\sigma}_{Nk^{(m)}+N})-\alpha^{\sigma}_{Nk^{(m-1)}+1}\theta^{\sigma}_{Nk^{(m-1)}+1})(h(x_{Nk^{(m)}+N-1})-h^*)\dots\\
&\qquad\quad+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})(h(x_i)-h^*)\Bigg)\\
&\qquad\quad\le\frac{1}{2}\left\|x_{Nk^{(m_{K_\sigma})}}-x^*\right\|^2+\frac{1}{4}\left\|x_{Nk^{(m_{K_\sigma})}}-x_{Nk^{(m_{K_\sigma})}-1}\right\|^2\\
&\qquad\qquad+\alpha^{\sigma}_{Nk^{(m_{K_\sigma})}+1}(1+\theta^{\sigma}_{Nk^{(m_{K_\sigma})}+1})(h(x_{Nk^{(m_{K_\sigma})}-1})-h^*)
\end{aligned}
\tag{16}
$$

Applying Jensen's inequality to the LHS of (16), we have

$$
\begin{aligned}
&\frac{1}{2}\left\|[x_{Nk^{(m_{K_\sigma})}}]_\sigma-[x^*]_\sigma\right\|^2+\frac{1}{4}\left\|[x_{Nk^{(m_{K_\sigma})}}]_\sigma-[x_{Nk^{(m_{K_\sigma})}-1}]_\sigma\right\|^2+\alpha^{\sigma}_{Nk^{(m_{K_\sigma})}+1}\theta^{\sigma}_{Nk^{(m_{K_\sigma})}+1}(h(x_{Nk^{(m_{K_\sigma})}-1})-h^*)\\
&\quad\ge\text{LHS}\\
&\quad\ge C_{NK_\sigma+N}\big(h(\widetilde{x}_{NK_\sigma+N})-h_*\big)
\end{aligned}
$$

where

$$
\begin{aligned}
C_{NK_\sigma+N}&=\alpha^{\sigma}_{NK_\sigma+N}(1+\theta^{\sigma}_{NK_\sigma+N})+\sum_{i=NK_\sigma}^{NK_\sigma+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})\dots\\
&\qquad+\sum_{m=1}^{m_{K_\sigma}}\left((\alpha^{\sigma}_{Nk^{(m)}+N}(1+\theta^{\sigma}_{Nk^{(m)}+N})-\alpha^{\sigma}_{Nk^{(m-1)}+1}\theta^{\sigma}_{Nk^{(m-1)}+1})+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})\right)\\
&>\frac{1}{2L}+0+\left(\sum_{m=1}^{m_{K_\sigma}}1\right)\min_{m=1,\dots,m_{K_\sigma}}\Bigg\{(\alpha^{\sigma}_{Nk^{(m)}+N}(1+\theta^{\sigma}_{Nk^{(m)}+N})-\alpha^{\sigma}_{Nk^{(m-1)}+1}\theta^{\sigma}_{Nk^{(m-1)}+1})\dots\\
&\qquad+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})\Bigg\}\\
&>\left(\sum_{m=1}^{m_{K_\sigma}}1\right)\min_{m=1,\dots,m_{K_\sigma}}\Bigg\{(\alpha^{\sigma}_{Nk^{(m)}+N}(1+\theta^{\sigma}_{Nk^{(m)}+N})-\alpha^{\sigma}_{Nk^{(m-1)}+1}\theta^{\sigma}_{Nk^{(m-1)}+1})\dots\\
&\qquad+\sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2}(\alpha^{\sigma}_{i+1}(1+\theta^{\sigma}_{i+1})-\alpha^{\sigma}_{i+2}\theta^{\sigma}_{i+2})\Bigg\}
\end{aligned}
$$

and

$$\widetilde{x}_{NK_\sigma+N} = \Bigg( \alpha^\sigma_{NK_\sigma+N}(1+\theta^\sigma_{NK_\sigma+N})x_{NK_\sigma+N-1} + \sum_{i=NK_\sigma}^{NK_\sigma+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2})x_i \dots$$
$$+ \sum_{m=1}^{m_{K_\sigma}} \Bigg( (\alpha^\sigma_{Nk^{(m)}+N}(1+\theta^\sigma_{Nk^{(m)}+N}) - \alpha^\sigma_{Nk^{(m-1)}+1}\theta^\sigma_{Nk^{(m-1)}+1})x_{Nk^{(m)}+N-1} \cdots$$
$$+ \sum_{i=Nk^{(m)}}^{Nk^{(m)}+N-2} (\alpha^\sigma_{i+1}(1+\theta^\sigma_{i+1}) - \alpha^\sigma_{i+2}\theta^\sigma_{i+2})x_i \Bigg) \Bigg) \Big/ C_{NK_\sigma+N}.$$

Note that $\mathbb{E}\Big[\big(\sum_{m=1}^{m_{K_\sigma}} 1\big)\Big] = \frac{K_\sigma}{s}$ implies $\mathbb{E}\big[C_{NK_\sigma+N}\big] = \Omega(K_\sigma)$ and for $\forall\sigma$

$$\mathbb{E}\Big[\text{RHS of (5)}\Big] = \frac{1}{2}\mathbb{E}\Big[ \| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x^*]_\sigma \|^2 \Big] + \frac{1}{4}\mathbb{E}\Big[ \Big\| [x_{Nk^{(m_{K_\sigma})}}]_\sigma - [x_{Nk^{(m_{K_\sigma})}-1}]_\sigma \Big\|^2 \Big]$$
$$+ \mathbb{E}\Big[\alpha^\sigma_{Nk^{(m_{K_\sigma})}+1}\theta^\sigma_{Nk^{(m_{K_\sigma})}+1}(h(x_{Nk^{(m_{K_\sigma})}-1}) - h^*)\Big]$$
$$= \frac{1}{2s}\mathbb{E}\left[ \| x_{Nk^{(m_{K_\sigma})}} - x^* \|^2 \right] + \frac{1}{4s}\mathbb{E}\left[ \left\| x_{Nk^{(m_{K_\sigma})}} - x_{Nk^{(m_{K_\sigma})}-1} \right\|^2 \right]$$
$$+ \mathbb{E}\left[\alpha^\sigma_{Nk^{(m_{K_\sigma})}+1}\theta^\sigma_{Nk^{(m_{K_\sigma})}+1}(h(x_{Nk^{(m_{K_\sigma})}-1}) - h^*)\right]$$

is a constant. We then take full expectation on both sides, and get

$$\mathbb{E}\big[h(\widetilde{x}_{NK_\sigma+N}) - h^*\big]$$
$$\leq \frac{\frac{1}{2s}\mathbb{E}\left[ \| x_{Nk^{(m_{K_\sigma})}} - x^* \|^2 \right] + \frac{1}{4s}\mathbb{E}\left[ \left\| x_{Nk^{(m_{K_\sigma})}} - x_{Nk^{(m_{K_\sigma})}-1} \right\|^2 \right] + \mathbb{E}\left[\alpha^\sigma_{Nk^{(m_{K_\sigma})}+1}\theta^\sigma_{Nk^{(m_{K_\sigma})}+1}(h(x_{Nk^{(m_{K_\sigma})}-1}) - h^*)\right]}{\mathbb{E}\big[C_{NK_\sigma+N}\big]}$$
$$= \mathcal{O}\Big(\frac{1}{K_\sigma}\Big).$$

To simplify the representation of the convergence analysis, we consider $K_\sigma = K$ such that $\sigma(K) = \sigma(0)$ without lost of generality. We then obtain

$$\mathbb{E}\big[h(\widetilde{x}_{NK+N}) - h^*\big] \leq \frac{\frac{1}{2s}\mathbb{E}\left[ \| x_0 - x^* \|^2 \right] + \frac{1}{4s}\mathbb{E}\left[ \| x_0 - x_{-1} \|^2 \right] + \mathbb{E}\left[\alpha^\sigma_1\theta^\sigma_1(h(x_{-1}) - h^*)\right]}{\mathbb{E}\big[C_{NK+N}\big]} = \mathcal{O}\Big(\frac{1}{K}\Big).$$

□

**Lemma A.2.** *Consider Algorithm 2 under the assumption that $f\colon \mathbb{R}^p \to \mathbb{R}$ is locally smooth and convex; $g\colon \mathbb{R}^p \to \mathbb{R}$ is closed, convex, proper. Let $x^*$ be any solution of (2) and $h^* := f^* + g^* = f(x^*) + g(x^*)$. Then for all $k = 0, 1, \dots$ with the selected block $\sigma$, it holds*

$$\frac{1}{2} \| [x_{Nk+N}]_\sigma - [x^*]_\sigma \|^2 + \frac{1}{4} \| [x_{Nk+N}]_\sigma - [x_{Nk+N-1}]_\sigma \|^2 + \alpha^\sigma_{Nk+N}(1+\theta^\sigma_{Nk+N})(h(x_{Nk+N-1}) - h^*)$$

$$+\sum_{i=Nk}^{Nk+N-2}\big(\frac{1}{4}+(\theta_{i+2}^{\sigma})^2\alpha_{i+1}^{\sigma}A_{i+1}^{\sigma}\big(1-\alpha_{i+1}^{\sigma}B_{i+1}^{\sigma}\big)\big)\,\|\,[x_{i+1}]_\sigma-[x_i]_\sigma\,\|^2+\sum_{i=Nk}^{Nk+N-2}\big(\alpha_{i+1}^{\sigma}(1+\theta_{i+1}^{\sigma})-\alpha_{i+2}^{\sigma}\theta_{i+2}^{\sigma}\big)(h(x_i)-h^*)$$
$$\leq\frac{1}{2}\,\|\,[x_{Nk}]_\sigma-[x^*]_\sigma\,\|^2+\frac{1}{4}\,\|\,[x_{Nk}]_\sigma-[x_{Nk'+N-1}]_\sigma\,\|^2+\alpha_{Nk+1}^{\sigma}(1+\theta_{Nk+1}^{\sigma})[h(x_{Nk'+N-1})-h^*], \tag{17}$$

*where for* $n=1,2,\ldots,N-1$,

$$A_{Nk+n}^{\sigma}=\frac{\langle[\nabla f(x_{Nk+n})]_\sigma-[\nabla f(x_{Nk+n-1})]_\sigma,[x_{Nk+n}]_\sigma-[x_{Nk+n-1}]_\sigma\rangle}{\|\,[x_{Nk+n}]_\sigma-[x_{Nk+n-1}]_\sigma\,\|^2}$$
$$B_{Nk+n}^{\sigma}=\frac{\|\,[\nabla f(x_{Nk+n})]_\sigma-[\nabla f(x_{Nk+n-1)}]_\sigma\,\|^2}{\langle[\nabla f(x_{Nk+n})]_\sigma-[\nabla f(x_{Nk+n-1})]_\sigma,[x_{Nk+n}]_\sigma-[x_{Nk+n-1}]_\sigma\rangle},$$

*and*

$$A_{Nk}^{\sigma}=\frac{\langle[\nabla f(x_{Nk})]_\sigma-[\nabla f(x_{Nk'+N-1})]_\sigma,[x_{Nk}]_\sigma-[x_{Nk'+N-1}]_\sigma\rangle}{\|\,[x_{Nk}]_\sigma-[x_{Nk'+N-1}]_\sigma\,\|^2}$$
$$B_{Nk}^{\sigma}=\frac{\|\,[\nabla f(x_{Nk})]_\sigma-[\nabla f(x_{Nk'+N-1)}]_\sigma\,\|^2}{\langle[\nabla f(x_{Nk})]_\sigma-[\nabla f(x_{Nk'+N-1})]_\sigma,[x_{Nk}]_\sigma-[x_{Nk'+N-1}]_\sigma\rangle}.$$

*Proof.* Let $H_{Nk}(x)=x-\alpha_{Nk}^{\sigma}\nabla f(x)$. Note that

$$\frac{1}{\alpha_{Nk+1}^{\sigma}}\big([H_{Nk+1}(x_{Nk})]_\sigma-[x_{Nk+1}]_\sigma\big)=\frac{1}{\alpha_{Nk+1}^{\sigma}}\big([x_{Nk}]_\sigma-[x_{Nk+1}]_\sigma\big)-[\nabla f(x_{Nk})]_\sigma\in[\partial g(x_{Nk+1})]_\sigma,$$

By the convexity of $g$ and cosine rule, we have

$$\begin{aligned}0&\leq g^*-g(x_{Nk+1})+\langle[\nabla f(x_{Nk})]_\sigma,[x^*]_\sigma-[x_{Nk+1}]_\sigma\rangle-\frac{1}{\alpha_{Nk+1}^{\sigma}}\langle[x_{Nk}]_\sigma-[x_{Nk+1}]_\sigma,[x^*]_\sigma-[x_{Nk+1}]_\sigma\rangle\\&=g^*-g(x_{Nk+1})+\underbrace{\langle[\nabla f(x_{Nk})]_\sigma,[x^*]_\sigma-[x_{Nk+1}]_\sigma\rangle}_{A}+\frac{1}{2\alpha_{Nk+1}^{\sigma}}\,\|\,[x_{Nk}]_\sigma-[x^*]_\sigma\,\|^2\\&\quad-\frac{1}{2\alpha_{Nk+1}^{\sigma}}\,\|\,[x_{Nk+1}]_\sigma-[x^*]_\sigma\,\|^2-\frac{1}{2\alpha_{Nk+1}^{\sigma}}\,\|\,[x_{Nk+1}]_\sigma-[x_{Nk}]_\sigma\,\|^2.\end{aligned} \tag{18}$$

We bound $A$ as

$$\begin{aligned}A&=\langle[\nabla f(x_{Nk})]_\sigma,[x^*]_\sigma-[x_{Nk}]_\sigma\rangle+\langle[\nabla f(x_{Nk})]_\sigma,[x_{Nk}]_\sigma-[x_{Nk+1}]_\sigma\rangle\\&=\langle[\nabla f(x_{Nk})]_\sigma,[x^*]_\sigma-[x_{Nk}]_\sigma\rangle+\frac{1}{\alpha_{Nk}^{\sigma}}\langle[H_{Nk}(x_{Nk'+N-1})]_\sigma-[x_{Nk}]_\sigma,[x_{Nk+1}]_\sigma-[x_{Nk}]_\sigma\rangle\\&\quad+\frac{1}{\alpha_{Nk}^{\sigma}}\langle[H_{Nk}(x_{Nk'+N-1})]_\sigma-[H_{Nk}(x_{Nk})]_\sigma,[x_{Nk}]_{\sigma:m}-[x_{Nk+1}]_\sigma\rangle\\&=f^*-f(x_{Nk})+g(x_{Nk+1})-g(x_{Nk})+\underbrace{\frac{1}{\alpha_{Nk}^{\sigma}}\langle[H_{Nk}(x_{Nk'+N-1})]_\sigma-[H_{Nk}(x_{Nk})]_\sigma,[x_{Nk}]_\sigma-[x_{Nk+1}]_\sigma\rangle}_{B}.\end{aligned} \tag{19}$$

Since

$$\begin{aligned}\| [H_{Nk}(x_{Nk'+N-1})]_\sigma - [H_{Nk}(x_{Nk})]_\sigma \|^2 &= \| [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \|^2 + \left(\alpha^\sigma_{Nk}\right)^2 \| [\nabla f(x_{Nk'+N-1})]_\sigma - [\nabla f(x_{Nk})]_\sigma \|^2 \\ &\quad - 2\alpha^\sigma_{Nk}\langle [\nabla f(x_{Nk'+N-1})]_\sigma - [\nabla f(x_{Nk})]_\sigma, [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \rangle \\ &= \| [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \|^2 + \left(\alpha^\sigma_{Nk}\right)^2 A^\sigma_{Nk} B^\sigma_{Nk} \| [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \|^2 \\ &\quad - 2\alpha^\sigma_{Nk} A^\sigma_{Nk} \| [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \|^2 \\ &= \left(1 - \alpha^\sigma_{Nk} A^\sigma_{Nk}(2 - \alpha^\sigma_{Nk} B^\sigma_{Nk})\right) \| [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \|^2 ,\end{aligned}$$

we bound $B$ as

$$\begin{aligned}B &\le \frac{\varepsilon_{Nk+1}}{2\alpha^\sigma_{Nk}} \| [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \|^2 + \frac{1}{2\varepsilon_{Nk+1}\alpha^\sigma_{Nk}} \| [H_{Nk}(x_{Nk'+N-1})]_\sigma - [H_{Nk}(x_{Nk})]_\sigma \|^2 \\ &= \frac{\varepsilon_{Nk+1}}{2\alpha^\sigma_{Nk}} \| [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \|^2 + \frac{1 - \alpha^\sigma_{Nk} A^\sigma_{Nk}(2 - \alpha^\sigma_{Nk} B^\sigma_{Nk})}{2\varepsilon_{Nk+1}\alpha^\sigma_{Nk}} \| [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \|^2 \\ &= \frac{1}{4\alpha^\sigma_{Nk+1}} \| [x_{Nk}]_\sigma - [x_{Nk+1}]_\sigma \|^2 + \frac{\alpha^\sigma_{Nk+1}\left(1 - \alpha^\sigma_{Nk} A^\sigma_{Nk}(2 - \alpha^\sigma_{Nk} B^\sigma_{Nk})\right)}{\left(\alpha^\sigma_{Nk}\right)^2} \| [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \|^2 .\end{aligned} \tag{20}$$

where $\varepsilon_{Nk+1} > 0$ from the Young's inequality is set to be $\varepsilon_{Nk+1} = \frac{1}{2\theta^\sigma_{Nk+1}}$ in the third line. We then combine (18), (19) and (20) to get

$$\begin{aligned}0 \le h^* - h(x_{Nk}) &+ \frac{1}{2\alpha^\sigma_{Nk+1}} \| [x_{Nk}]_\sigma - [x^*]_\sigma \|^2 - \frac{1}{2\alpha^\sigma_{Nk+1}} \| [x_{Nk+1}]_\sigma - [x^*]_\sigma \|^2 \\ &- \frac{1}{4\alpha^\sigma_{Nk+1}} \| [x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma \|^2 + \frac{\alpha^\sigma_{Nk+1}\left(1 - \alpha^\sigma_{Nk} A^\sigma_{Nk}(2 - \alpha^\sigma_{Nk} B^\sigma_{Nk})\right)}{\left(\alpha^\sigma_{Nk}\right)^2} \| [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \|^2 .\end{aligned} \tag{21}$$

Note that

$$\frac{[x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma}{\alpha^\sigma_{Nk}} - \left([\nabla f(x_{Nk'+N-1})]_\sigma - [\nabla f(x_{Nk})]_\sigma\right) \in [\partial h(x_{Nk})]_\sigma.$$

By the convexity of $h$, we have

$$\begin{aligned}0 &\le h(x_{Nk'+N-1}) - h(x_{Nk}) - \langle \frac{[x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma}{\alpha^\sigma_{Nk}} - \left([\nabla f(x_{Nk'+N-1})]_\sigma - [\nabla f(x_{Nk})]_\sigma\right), [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \rangle \\ &= h(x_{Nk'+N-1}) - h(x_{Nk}) - \frac{1}{\alpha^\sigma_{Nk}} \| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|^2 \\ &\quad + \langle [\nabla f(x_{Nk'+N-1})]_\sigma - [\nabla f(x_{Nk})]_\sigma, [x_{Nk'+N-1}]_\sigma - [x_{Nk}]_\sigma \rangle \\ &= h(x_{Nk'+N-1}) - h(x_{Nk}) - \frac{1 - \alpha^\sigma_{Nk} A^\sigma_{Nk}}{\alpha^\sigma_{Nk}} \| [x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma \|^2 .\end{aligned} \tag{22}$$

Note that

$$h^* - h(x_{Nk}) + \theta^\sigma_{Nk+1}(h(x_{Nk'+N-1}) - h(x_{Nk})) = \theta^\sigma_{Nk+1}(h(x_{Nk'+N-1}) - h^*) - (1 + \theta^\sigma_{Nk+1})(h(x_{Nk}) - h^*).$$

We multiply (22) by $\theta^{\sigma}_{Nk+1}$, combine it with (21), and multiply $\alpha^{\sigma}_{Nk+1}$ on both sides to get

$$\frac{1}{2}\,\|\,[x_{Nk+1}]_\sigma - [x^*]_\sigma\,\|^2 + \frac{1}{4}\,\|\,[x_{Nk+1}]_\sigma - [x_{Nk}]_\sigma\,\|^2 + \alpha^{\sigma}_{Nk+1}(1+\theta^{\sigma}_{Nk+1})(h(x_{Nk}) - h^*)$$
$$\leq \frac{1}{2}\,\|\,[x_{Nk}]_\sigma - [x^*]_\sigma\,\|^2 + \left(\theta^{\sigma}_{Nk+1}\right)^2\alpha^{\sigma}_{Nk}A^{\sigma}_{Nk}\left(\alpha^{\sigma}_{Nk}B^{\sigma}_{Nk}-1\right)\|\,[x_{Nk}]_\sigma - [x_{Nk'+N-1}]_\sigma\,\|^2 + \alpha^{\sigma}_{Nk+1}\theta^{\sigma}_{Nk+1}(h(x_{Nk'+N-1}) - h^*),$$

Following the similar derivation above, we have for $n = 2, \ldots, N$

$$\frac{1}{2}\,\|\,[x_{Nk+n}]_\sigma - [x^*]_\sigma\,\|^2 + \frac{1}{4}\,\|\,[x_{Nk+n}]_\sigma - [x_{Nk+n-1}]_\sigma\,\|^2 + \alpha^{\sigma}_{Nk+n}(1+\theta^{\sigma}_{Nk+n})(h(x_{Nk+n-1}) - h^*)$$
$$\leq \frac{1}{2}\,\|\,[x_{Nk+n-1}]_\sigma - [x^*]_\sigma\,\|^2 + \left(\theta^{\sigma}_{Nk+n}\right)^2\alpha^{\sigma}_{Nk+n-1}A^{\sigma}_{Nk+n-1}\left(\alpha^{\sigma}_{Nk+n-1}B^{\sigma}_{Nk+n-1}-1\right)\|\,[x_{Nk+n-1}]_\sigma - [x_{Nk+n-2}]_\sigma\,\|^2$$
$$+ \alpha^{\sigma}_{Nk+n}\theta^{\sigma}_{Nk+n}(h(x_{Nk+n-2}) - h^*).$$

Telescoping these inequalities from $n = 1$ to $N$, we obtain (17), where $N \geq 2$ ensures that $Nk + N - 2 \geq Nk + N$. □

# B Experiments