Computational Optimization and Applications
https://doi.org/10.1007/s10589-020-00183-1

Accelerating incremental gradient optimization with curvature information

Hoi-To Wai[1] · Wei Shi[2] · César A. Uribe[4] · Angelia Nedić[3] · Anna Scaglione[3]

Received: 22 June 2019

Abstract
This paper studies an acceleration technique for incremental aggregated gradient (IAG) method through the use of *curvature* information for solving strongly convex finite sum optimization problems. These optimization problems of interest arise in large-scale learning applications. Our technique utilizes a curvature-aided gradient tracking step to produce accurate gradient estimates incrementally using Hessian information. We propose and analyze two methods utilizing the new technique, the curvature-aided IAG (CIAG) method and the accelerated CIAG (A-CIAG) method, which are analogous to gradient method and Nesterov's accelerated gradient method, respectively. Setting κ to be the condition number of the objective function, we prove the R linear convergence rates of $1 - \frac{4c_0\kappa}{(\kappa+1)^2}$ for the CIAG method, and $1 - \sqrt{\frac{c_1}{2\kappa}}$ for the A-CIAG method, where $c_0, c_1 \leq 1$ are constants inversely proportional to the distance between the initial point and the optimal solution. When the initial iterate is close to the optimal solution, the R linear convergence rates match with the gradient and accelerated gradient method, albeit CIAG and A-CIAG operate in an incremental setting with strictly lower computation complexity. Numerical experiments confirm our findings. The source codes used for this paper can be found on http://github.com/hoitowai/ciag/.

Keywords Incremental aggregated gradient method · Curvature information · Nesterov accelerated gradient method

In memory of Dr. Wei Shi, a respected friend and talented scholar.

Preliminary version of this work has been presented at the 55th Annual Allerton Conference on Communication, Control, and Computing in October, 2017 [29]

✉ Hoi-To Wai
htwai@se.cuhk.edu.hk

Extended author information available on the last page of the article

Published online: 07 March 2020

1 Introduction

Consider a finite sum optimization problem with m component functions and a d-dimensional decision variable:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \; F(\boldsymbol{\theta}) := \textstyle\sum_{i=1}^m f_i(\boldsymbol{\theta}) \; . \tag{1}$$

Problem (1) is motivated by the empirical risk minimization model where we are learning a parameter $\boldsymbol{\theta}$ from a finite set of data. The component function $f_i(\boldsymbol{\theta})$ represents the statistical mismatch between $\boldsymbol{\theta}$ and the ith piece of data collected. The aim of problem (1) is to learn the optimal parameter, denoted by $\boldsymbol{\theta}^\star$, that fits with the m available data points [28].

We are interested in the setting when each of the component function $f_i(\boldsymbol{\theta})$ is twice continuously differentiable and the sum function $F(\boldsymbol{\theta})$ is strongly convex. Especially, this paper focuses on *large-scale* instances of (1) with $m \gg 1$. The difficulty with solving (1) lies in the overwhelming size of m, which prohibits us from even applying simple first order methods. For example, the *full gradient* (FG) requires the recursion: let $\gamma > 0$ be a step size,

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \gamma \textstyle\sum_{i=1}^m \nabla f_i(\boldsymbol{\theta}^k) \; , \tag{2}$$

that involves a computation cost of $\mathcal{O}(md)$ floating point operations (FLOPS) per iteration to compute the sum of m gradient vectors. As $m \gg 1$, this is undesirable from a practical standpoint. To this end, a popular yet powerful approach is to adopt the so-called *incremental* (or stochastic) methods where only one of the component functions, e.g., the i_kth one, $f_{i_k}(\boldsymbol{\theta})$, is explored at the kth iteration. Examples include the incremental gradient (IG) [4], incremental aggregated gradient (IAG) [5, 14, 27] methods when i_k is deterministic; and the stochastic gradient (SG) [22], stochastic average gradient (SAG) [24], SAGA [9], stochastic variance reduced gradient (SVRG) [30] methods when i_k is chosen randomly. Related work can be found in [16, 26] for the case with non-convex functions; in [15] for the cases of finite-sum and primal-dual optimization; in [18, 19] for the case with non-smooth functions. The interested readers are referred to [6, Section 4, 5] for a comprehensive overview on the topic.

A number of the above incremental methods achieve linear convergence via the variance reduction technique, e.g., [9, 24, 30]. In the worst case, these methods require $k = \mathcal{O}(m \log(1/\epsilon))$ iterations to guarantee that they compute an ϵ-optimal solution $\boldsymbol{\theta}^k$ satisfying $F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \le \epsilon$. In problem instances when $m \gg 1$, such rates may not be ideal. In fact, recent work [1, 2, 15] showed that the dependence on m is necessary for incremental methods whose updates are linear combinations of the first order information. To improve the convergence rate of incremental methods, a recent direction is to adopt ideas of second-order optimization. Examples include [12, 17, 23] which extended Newton and quasi-Newton methods to the incremental setting, resulting in NIM [23] and IQN [17]. These works demonstrated that at the expense of additional storage or computation cost, it is possible to achieve local but *superlinear* convergence.

We propose a *curvature-aided gradient tracking* technique for accelerating incremental gradient methods using the curvature (Hessian) information. By applying Taylor expansion on the component functions' gradients, we derive a new gradient estimator whose error depends on the *squares* of the optimality gaps. Based on this new gradient tracking technique, we propose two incremental gradient methods, called Curvature-aided Incremental Aggregated Gradient (CIAG) method and Accelerated CIAG (A-CIAG) method. The proposed methods require $\mathcal{O}(md)$ storage cost and $\mathcal{O}(d^2)$ computation cost per iteration. Furthermore, we characterize the *R*-linear convergence rates of the proposed methods. For the CIAG (resp. A-CIAG) method, we show that the rate is $1-\frac{4c_0\kappa}{(\kappa+1)^2}$ (resp. $1-\sqrt{\frac{c_1}{2\kappa}}$), where κ is the conditional number of the objective function of the summed objective function $F(\boldsymbol{\theta})$ in (1), c_0, c_1 are constants which can be set to 1 given that the initial iterate is close to the optimal solution [cf. Theorems 1 and 2]. In addition, we establish the *R*-linear convergence rates for two *non-linear* inequalities [cf. Propositions 5 and 6] which are of independent interest. In detail, our result reveals that the convergence of CIAG and A-CIAG methods depends on the initialization, and the trajectory of convergence can be divided into two phases — the initial phase that exhibits a slow but global linear rate; and the asymptotic phase that exhibits linear convergence of an accelerated rate.

A comparison of the proposed CIAG/A-CIAG methods to state-of-the-art methods is summarized in Table 1. Compared to first order methods such as SAG, we remark that the proposed methods (as well as SVRG2, IQN, NIM) require the Hessian of the objective function to be Lipschitz continuous. Though the storage requirement for the proposed methods is $\mathcal{O}(md+d^2)$ which is higher

Table 1 Comparison of state-of-the-art methods for the strongly convex problem (1)

	Storage	Comp.	# Iterations to ϵ-optimal solution
IAG [5]	$\mathcal{O}(md)$	$\mathcal{O}(d)$	$\mathcal{O}(m\,\kappa\log(1/\epsilon))$ [worst-case]
SAG [24]	$\mathcal{O}(md)$	$\mathcal{O}(d)$	$\mathcal{O}(\max\{\kappa,m\}\log(1/\epsilon))$ [expected]
SAGA [9]	$\mathcal{O}(md)$	$\mathcal{O}(d)$	$\mathcal{O}((\kappa+m)\log(1/\epsilon))$ [expected]
SVRG [30]	$\mathcal{O}(d)$	$\mathcal{O}(d)$	$\mathcal{O}((\kappa+m)\log(1/\epsilon))$ [expected]
AccSVRG [21]	$\mathcal{O}(d)$	$\mathcal{O}(d)$	$\mathcal{O}((m\sqrt{\kappa}+m)\log(1/\epsilon))$ [expected]
SVRG2 [11]	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}(\kappa\log(1/\epsilon)+m)$ [expected]
IQN [17]	$\mathcal{O}(md^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}(m)$, *i.e.,* super-linear [worst-case]
NIM [23]	$\mathcal{O}(md+d^2)$	$\mathcal{O}(d^3)$	$\mathcal{O}(m)$, *i.e.,* super-linear [worst-case]
CIAG (proposed)	$\mathcal{O}(md+d^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}(\kappa\log(1/\epsilon)+m)$ [worst-case]
A-CIAG (proposed)	$\mathcal{O}(md+d^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon)+m)$ [worst-case]

The second column is the memory required for the working variables. The third column is the *per iteration* computation complexity in FLOPS. The last column is the (expected or worst-case) number of iterations to reach an ϵ-optimal solution. SVRG, AccSVRG, SVRG2 require recomputing the full gradient at every epoch and a careful tuning of the epoch size. The constant κ is the condition number of $F(\boldsymbol{\theta})$ [see (3) and (4) for the definition]. The rates for the last four methods are *asymptotic*, *i.e.,* they hold only when the desired accuracy is small, e.g., $\epsilon \ll 1$. The shaded rows denote the proposed methods which achieve a good trade-off between per-iteration complexity and convergence speed than the $\mathcal{O}(md)$ requirement of the SAG methods when $d \gg 1$, the number of iterations required is significantly lower when the desired accuracy is high, $\epsilon \ll 1$. Overall, as seen from Table 1, the proposed methods are desired when d is small-to-moderate, and the number of component functions m is large.

It is worth mentioning that recently in [11] the authors proposed a method that incorporates Hessian information to accelerate SVRG method, giving the SVRG2 method. The authors developed approximation techniques to reduce the per iteration complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(d)$. The best convergence rates achieved therein match that of the FG method and is the same as the CIAG method. However, we note that SVRG2 re-computes the full Hessian at the beginning of each epoch. This costs an additional $\mathcal{O}(md^2)$ FLOPS per epoch and the cost is negligible only when the epochs are long, e.g., when the epoch lengths are $\mathcal{O}(m)$.

Organization Section 2 studies incremental algorithms with *gradient tracking* and introduces the curvature-aided gradient tracking technique. Section 3 describes the proposed CIAG and A-CIAG methods, and discusses the implementation issues. Section 4 provides the main convergence results. Section 5 demonstrates the efficacy of the proposed methods via numerical experiments.

Notations We use $\mathbf{0}$ to denote an all-zero vector/matrix with suitable dimensions. Unless otherwise specified, we denote by $\| \cdot \|$ the Euclidean norm. A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if

$$f(\boldsymbol{\theta}') \leq f(\boldsymbol{\theta}) + \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + (L/2)\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 , \tag{3}$$

for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, and it is μ-strongly convex if

$$f(\boldsymbol{\theta}') \geq f(\boldsymbol{\theta}) + \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + (\mu/2)\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 , \tag{4}$$

for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$. Define $\kappa(f) := L/\mu$ as the condition number of f. Also, f has an L_H -Lipschitz continuous Hessian if

$$\|\nabla^2 f(\boldsymbol{\theta}') - \nabla^2 f(\boldsymbol{\theta})\| \leq L_H \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|, \; \forall \, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d , \tag{5}$$

where the norm on the left hand side is the matrix norm induced by Euclidean norm. For a non-negative scalar sequence $\{a^{(k)}\}_{k \geq 1}$, we say that it converges R-linearly at a rate ρ if the sequence satisfies $\lim_{k \to \infty} a^{(k+1)}/a^{(k)} = \rho$, where $0 \leq \rho < 1$. We use standard Bachmann-Landau notations for asymptotic quantities: $a^{(k)} = \mathcal{O}(f^{(k)})$ (resp. $a^{(k)} = \Omega(g^{(k)})$) implies that there exists $k_0 \in \mathbb{Z}_+$ and non-negative constant C_1 (resp. C_2) such that $a^{(k)} \leq C_1 f^{(k)}$ (resp. $a^{(k)} \geq C_2 g^{(k)}$) for all $k \geq k_0$.

2 Gradient tracking in incremental methods

We provide a high-level description of how to incorporate curvature information to accelerate incremental aggregated gradient method. To set up the notation, let $i_k \in \{1, ..., m\}$ be the component function index selected at the kth iteration, e.g., a simple rule is to use the cyclic rule as $i_k = (k \bmod m) + 1$, or we may choose $i_k \sim \mathcal{U}\{1, ..., m\}$. Let us define:

$$\tau_i^k := \max\{\tau \,:\, \tau \leq k,\ i_\tau = i\}\ , \tag{6}$$

i.e., τ_i^k is the iteration index where the ith component function is last accessed after the completion of the kth iteration. As we focus on analyzing the worst-case performance, we assume that $\tau_i^k \in [k - K + 1, k]$ for a constant $K = \mathcal{O}(m)$. For example, $K = m$ when i_k is chosen by the cyclic rule.

First-order approximation As described in (2), at the kth iteration the gradient method computes the *full gradient* as the complete sum $\sum_{i=1}^m \nabla f_i(\boldsymbol{\theta}^k)$. Such vector is unavailable in the incremental setting as only the access to the i_kth function is desired. A simple idea is to replace the full gradient by:

$$\boldsymbol{g}_{\mathsf{IAG}}^k := \sum_{i=1}^m \nabla f_i(\boldsymbol{\theta}^{\tau_i^k}) \approx \sum_{i=1}^m \nabla f_i(\boldsymbol{\theta}^k)\ , \tag{7}$$

or equivalently using the recursion:

$$\boldsymbol{g}_{\mathsf{IAG}}^k = \boldsymbol{g}_{\mathsf{IAG}}^{k-1} - \nabla f_{i_k}(\boldsymbol{\theta}^{\tau_{i_k}^{k-1}}) + \nabla f_{i_k}(\boldsymbol{\theta}^k)\ , \tag{8}$$

with $\boldsymbol{g}_{\mathsf{IAG}}^0 = \sum_{i=1}^m \nabla f_i(\boldsymbol{\theta}^0)$. The expression (7) describes the SAG and IAG methods, where their only differences lie in the selection strategy for i_k.

When the functions f_i are smooth, we have $\|\nabla f_i(\boldsymbol{\theta}^{\tau_i^k}) - \nabla f_i(\boldsymbol{\theta}^k)\| = \mathcal{O}(\|\boldsymbol{\theta}^{\tau_i^k} - \boldsymbol{\theta}^k\|)$. Furthermore, [14] shows that the error is bounded:

$$\|\boldsymbol{g}_{\mathsf{IAG}}^k - \sum_{i=1}^m \nabla f_i(\boldsymbol{\theta}^k)\| = \mathcal{O}\Big(\gamma m \max_{i=1,\dots,m} \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^{\tau_i^k}\|\Big). \tag{9}$$

Note that this error represents the 'variance' in gradient estimation. Eq. (9) implies that the tracking error decays to zero as long as $\boldsymbol{\theta}^k$ converges to $\boldsymbol{\theta}^\star$ [recall that $k - \tau_i^k \leq K < \infty$ and $K = \Theta(m)$]. However, the dependency on m of the right hand side in (9) is undesirable as it leads to the following estimates:

$$\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^\star\|^2 \leq \Big(1 - \mathcal{O}(\frac{\gamma}{\kappa(F)})\Big)\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star\|^2 + \mathcal{O}(\gamma^2 m^2 \max_{(k-2K)_{++} \leq \ell \leq k} \|\boldsymbol{\theta}^\ell - \boldsymbol{\theta}^\star\|^2)\ , \tag{10}$$

where $(x)_{++} := \max\{1, x\}$. As analyzed by [10, 14], due to the dependence of m^2 on the right hand side, the sequence of squared norm $\{\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star\|^2\}_{k\geq 1}$ converges only when $\gamma = \mathcal{O}(1/m)$, and finally this shows that $\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star\|^2$ converges linearly at rate $1 - \mathcal{O}((m^2\kappa)^{-1})$. Notice that a recent work [27] strengthened this rate to $1 - \mathcal{O}((m\kappa(F))^{-1})$. When the component function i_k is selected independently at random for each k, the rate may also be improved, see [24].

Second-order approximation The undesirable m-dependence of convergence rate with (7) is partly due to the crude approximation $\nabla f_i(\boldsymbol{\theta}^k) \approx \nabla f_i(\boldsymbol{\theta}^{\tau_i^k})$ used. As a natural idea to improve the approximation, we consider the Taylor expansion applied on the ith *gradient* vector itself:

$$\nabla f_i(\boldsymbol{\theta}^k) \approx \nabla f_i(\boldsymbol{\theta}^{\tau_i^k}) + \nabla^2 f_i(\boldsymbol{\theta}^{\tau_i^k})(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{\tau_i^k}) \; . \tag{11}$$

Compared to a first order approximation with $\nabla f_i(\boldsymbol{\theta}^k) \approx \nabla f_i(\boldsymbol{\theta}^{\tau_i^k})$, the approximation given by the right hand side in (11) has two important features. First, we observe that this term yields a *second-order approximation* to the gradient of the component function $f_i(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^k$, and thus leading to an improved approximation. Second, the approximate term on the right hand side does not require access to the ith function evaluated at $\boldsymbol{\theta}^k$, which allows one to derive a similar recursion rule as (8) to *incrementally* compute the approximation.

To further evaluate the approximation quality given by (11), for each i we shall assume that the Hessian of f_i is Lipschitz continuous. We observe the following upper bound on the error:

Lemma 1 [20, Lemma 1.2.5] *Assume that* $\nabla^2 f_i(\boldsymbol{\theta})$ *is* $L_{H,i}$ *Lipschitz. Then*:

$$\begin{aligned} &\left\| \nabla f_i(\boldsymbol{\theta}) - \left(\nabla f_i(\boldsymbol{\theta}') + \nabla^2 f_i(\boldsymbol{\theta}')(\boldsymbol{\theta} - \boldsymbol{\theta}') \right) \right\| \\ &\quad \le (L_{H,i}/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2, \; \forall \, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d \; . \end{aligned} \tag{12}$$

From Lemma 1, we notice that the approximation error of the right hand side in (11) depends on the *squared difference* between $\boldsymbol{\theta}^k$ and $\boldsymbol{\theta}^{\tau_i^k}$. When $\boldsymbol{\theta}^k$ is close to $\boldsymbol{\theta}^{\tau_i^k}$, this error will be significantly smaller than what is obtained for the first order approximation methods in (7). We remark that similar approximation schemes are proposed in recent work [11, 31], which demonstrated the benefits of applying higher-order approximation in stochastic optimization.

We refer to this form of gradient approximation in (11) as the *curvature-aided gradient tracking technique*. In the remainder of this paper, we shall develop and analyze practical optimization methods using the technique.

3 Proposed CIAG and A-CIAG methods

We propose two incremental methods, CIAG and A-CIAG methods, which are built using the curvature-aided gradient tracking technique (11). Our first method, the CIAG method, performs the recursion:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \gamma \boldsymbol{g}^k_{\mathsf{CIAG}} \; , \tag{18}$$

where $\gamma > 0$ is a step size and the gradient surrogate is given by

$$\boldsymbol{g}_{\mathsf{CIAG}}^k := \sum_{i=1}^m \left(\nabla f_i(\boldsymbol{\theta}^{\tau_i^k}) - \nabla^2 f_i(\boldsymbol{\theta}^{\tau_i^k}) \boldsymbol{\theta}^{\tau_i^k} \right) + \left(\sum_{i=1}^m \nabla^2 f_i(\boldsymbol{\theta}^{\tau_i^k}) \right) \boldsymbol{\theta}^k , \tag{19}$$

where τ_i^k was defined in (6). The method can be interpreted as applying the curvature-aided gradient tracking [cf. (11)] to *each* individual component function and applying the FG update (2). On the other hand, the accelerated CIAG (A-CIAG) method follows the recursion:

$$\boldsymbol{\theta}_{ex}^k = \boldsymbol{\theta}^k + \alpha(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}) , \tag{20a}$$

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}_{ex}^k - \gamma \boldsymbol{g}_{\mathsf{ACIAG}}^k , \tag{20b}$$

where $\alpha \in [0, 1), \gamma > 0$ are predefined parameters and the gradient surrogate is

$$\boldsymbol{g}_{\mathsf{ACIAG}}^k := \textstyle\sum_{i=1}^m \left(\nabla f_i(\boldsymbol{\theta}_{ex}^{\tau_i^k}) - \nabla^2 f_i(\boldsymbol{\theta}_{ex}^{\tau_i^k}) \boldsymbol{\theta}_{ex}^{\tau_i^k} \right) + \left(\textstyle\sum_{i=1}^m \nabla^2 f_i(\boldsymbol{\theta}_{ex}^{\tau_i^k}) \right) \boldsymbol{\theta}_{ex}^k , \tag{21}$$

where τ_i^k was defined in (6). The term $\boldsymbol{\theta}_{ex}^k$ in (20a) is called the extrapolated iterate, which incorporates the 'inertia' from previous iterates into the updates. Similar to the CIAG method, the A-CIAG method applies curvature-aided gradient tracking to the gradient of each individual component function, evaluated at $\boldsymbol{\theta}_{ex}^k$. We remark that even though Hessians are used in the CIAG and A-CIAG methods, we do not attempt to compute any matrix inverses which is in contrast with Newton methods. The Hessians here are used to accelerate the gradient tracking.

The CIAG and A-CIAG methods are similar since the full gradients evaluated at $\boldsymbol{\theta}^k$ and $\boldsymbol{\theta}_{ex}^k$ are approximated by their respective curvature-aided gradient tracking approximates [cf. (19) and (21)]. Both methods can be implemented in a memory efficient incremental fashion. To see this, define

$$\boldsymbol{\theta}_i^k := \begin{cases} \boldsymbol{\theta}^{\tau_i^k}, & \text{for} \mathsf{CIAG}, \\ \boldsymbol{\theta}_{ex}^{\tau_i^k}, & \text{for} \mathsf{A-CIAG}, \end{cases} \quad \boldsymbol{H}^k : = \sum_{i=1}^m \nabla^2 f_i(\boldsymbol{\theta}_i^k), \tag{22}$$

and

$$\boldsymbol{b}^k := \textstyle\sum_{i=1}^m \left(\nabla f_i(\boldsymbol{\theta}_i^k) - \nabla^2 f_i(\boldsymbol{\theta}_i^k) \boldsymbol{\theta}_i^k \right). \tag{23}$$

Note that $\boldsymbol{b}^k$ and $\boldsymbol{H}^k$ are the accumulated staled iterates, gradients and Hessians. These can be computed incrementally through storing the staled iterates in memory. Moreover, we can compute the gradient surrogates as $\boldsymbol{g}_{\mathsf{CIAG}}^k = \boldsymbol{b}^k + \boldsymbol{H}^k \boldsymbol{\theta}^k$ and $\boldsymbol{g}_{\mathsf{ACIAG}}^k = \boldsymbol{b}^k + \boldsymbol{H}^k \boldsymbol{\theta}_{ex}^k$. A pseudo code for implementing CIAG and A-CIAG is provided in Algorithm 1.

Algorithm 1 CIAG and A-CIAG Method.

1: **Input**: Initial point $\boldsymbol{\theta}^1 \in \mathbb{R}^d$, step size parameter $\gamma > 0$, (for A-CIAG only) extrapolation constant α.
2: Initialize the vectors/matrices in the memory:

$$\boldsymbol{\theta}_i^1 = \boldsymbol{\theta}^1,\ \tau_i^1 = 0,\ \forall\ i,\ \boldsymbol{b}^0 = \mathbf{0},\ \boldsymbol{H}^0 = \mathbf{0}\ . \tag{13}$$

3: **for** $k = 1, 2, \ldots$ **do**
4: Select $i_k \in \{1, ..., m\}$, e.g., $i_k = (k \bmod m) + 1$, and the algorithm is given access to $f_{i_k}(\boldsymbol{\theta})$. Set the counter variables as $\tau_{i_k}^k = k$, $\tau_j^k = \tau_j^{k-1}$ for all $j \neq i_k$.
5: *(For* A-CIAG *only)* If $k = 1$, we take $\boldsymbol{\theta}_{ex}^1 = \boldsymbol{\theta}^1$; otherwise we compute the extrapolated variable as $\boldsymbol{\theta}_{ex}^k = \boldsymbol{\theta}^k + \alpha(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1})$.
6: Update the vector in the memory:

$$\boldsymbol{\theta}_{i_k}^k = \begin{cases} \boldsymbol{\theta}^k, & \text{for CIAG }, \\ \boldsymbol{\theta}_{ex}^k, & \text{for A-CIAG }, \end{cases} \tag{14}$$

and set $\boldsymbol{\theta}_j^k = \boldsymbol{\theta}_j^{k-1}$ for all $j \neq i_k$.
7: If $\tau_{i_k}^{k-1} = 0$, then update: (*self-initialization*)

$$\begin{aligned} \boldsymbol{b}^k &= \boldsymbol{b}^{k-1} + \nabla f_{i_k}(\boldsymbol{\theta}_{i_k}^k) - \nabla^2 f_{i_k}(\boldsymbol{\theta}_{i_k}^k)\boldsymbol{\theta}_{i_k}^k\ , \\ \boldsymbol{H}^k &= \boldsymbol{H}^{k-1} + \nabla^2 f_{i_k}(\boldsymbol{\theta}_{i_k}^k)\ . \end{aligned} \tag{15}$$

Otherwise, update:

$$\begin{aligned} \boldsymbol{b}^k &= \boldsymbol{b}^{k-1} - \nabla f_{i_k}(\boldsymbol{\theta}_{i_k}^{k-1}) + \nabla f_{i_k}(\boldsymbol{\theta}_{i_k}^k) \\ &\quad + \nabla^2 f_{i_k}(\boldsymbol{\theta}_{i_k}^{k-1})\boldsymbol{\theta}_{i_k}^{k-1} - \nabla^2 f_{i_k}(\boldsymbol{\theta}_{i_k}^k)\boldsymbol{\theta}_{i_k}^k\ , \\ \boldsymbol{H}^k &= \boldsymbol{H}^{k-1} - \nabla^2 f_{i_k}(\boldsymbol{\theta}_{i_k}^{k-1}) + \nabla^2 f_{i_k}(\boldsymbol{\theta}_{i_k}^k)\ . \end{aligned} \tag{16}$$

8: Compute the update:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}_{i_k}^k - \gamma\big(\boldsymbol{b}^k + \boldsymbol{H}^k\boldsymbol{\theta}_{i_k}^k\big)\ . \tag{17}$$

9: **end for**
10: **Return**: an approximate solution to (1), $\boldsymbol{\theta}^{k+1}$.

3.1 Computation and storage costs

We comment on the computation and storage costs of CIAG and A-CIAG. Note that (15), (16) and (17) in the algorithm require $\mathcal{O}(d^2)$ FLOPS and they are the dominant computation steps involved. The overall complexities for CIAG and A-CIAG are thus $\mathcal{O}(d^2)$ FLOPS per iteration. On the other hand, the algorithm requires storing m vectors of d-dimesion and a $d \times d$ matrix [cf. (13)], therefore the storage cost is $\mathcal{O}(md + d^2) = \mathcal{O}(md)$ if $m \geq d$. When d is small, the computation and storage cost of CIAG and A-CIAG are comparable to existing methods such as SAG, SVRG; meanwhile for large d, the computation cost will be undesirable for CIAG and A-CIAG.

When the component functions are the negative log-likelihood of a linear model, the storage complexity can be reduced to as low as $\mathcal{O}(m)$. Note linear models are common in machine learning problems. We write

$$f_i(\boldsymbol{\theta}) = g_i(\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle) + (\rho/2)\|\boldsymbol{\theta}\|^2\ , \tag{24}$$

where $\boldsymbol{x}_i$ represents the ith associated data, while $g_i : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable. Observe that

$$\nabla f_i(\boldsymbol{\theta}) = g_i'(\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)\boldsymbol{x}_i + \rho\,\boldsymbol{\theta},\ \nabla^2 f_i(\boldsymbol{\theta}) = g_i''(\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)\boldsymbol{x}_i(\boldsymbol{x}_i)^\top + \rho \boldsymbol{I}\,, \tag{25}$$

where $\boldsymbol{I}$ is the $d \times d$ identity matrix. Substituting the above into (16) gives:

$$\begin{aligned} \boldsymbol{b}^k = \boldsymbol{b}^{k-1} &+ (g_{i_k}'(\langle \boldsymbol{\theta}_{i_k}^k, \boldsymbol{x}_{i_k}\rangle) - g_{i_k}'(\langle \boldsymbol{\theta}_{i_k}^{k-1}, \boldsymbol{x}_{i_k}\rangle))\boldsymbol{x}_{i_k} \\ &+ g_{i_k}''(\langle \boldsymbol{\theta}_{i_k}^{k-1}, \boldsymbol{x}_{i_k}\rangle)\langle \boldsymbol{\theta}_{i_k}^{k-1}, \boldsymbol{x}_{i_k}\rangle \boldsymbol{x}_{i_k} - g_{i_k}''(\langle \boldsymbol{\theta}_{i_k}^{k}, \boldsymbol{x}_{i_k}\rangle)\langle \boldsymbol{\theta}_{i_k}^{k}, \boldsymbol{x}_{i_k}\rangle \boldsymbol{x}_{i_k} \end{aligned} \tag{26}$$

$$\boldsymbol{H}^k = \boldsymbol{H}^{k-1} + (g_{i_k}''(\langle \boldsymbol{\theta}_{i_k}^k, \boldsymbol{x}_{i_k}\rangle) - g_{i_k}''(\langle \boldsymbol{\theta}_{i_k}^{k-1}, \boldsymbol{x}_{i_k}\rangle))\boldsymbol{x}_{i_k}\boldsymbol{x}_{i_k}^\top\,. \tag{27}$$

Therefore, it suffices for CIAG and A-CIAG to keep $\{\langle \boldsymbol{x}_i, \boldsymbol{\theta}_i\rangle\}_{i=1}^m$ for implementing the incremental updates, leading to an $\mathcal{O}(d^2 + m)$ storage cost.

4 Convergence analysis

In this section, we demonstrate that the proposed methods converge globally and characterize their convergence rates. Let us state the assumptions.

Assumption 1 The delayed iteration indices τ_i^k [cf. (6)] satisfy $0 \le k - \tau_i^k \le K$ for all i, k and for some $K \ge 0$.

Assumption 2 The function $F(\boldsymbol{\theta})$ is μ-strongly convex and L-smooth with $L \ge \mu > 0$.

Assumption 3 For each i, the Hessian of the function $f_i(\boldsymbol{\theta})$ is $L_{H,i}$-Lipschitz continuous.

We define $L_H := \sum_{i=1}^m L_{H,i}$ as the Lipschitz constant for the Hessian of the sum function $F(\boldsymbol{\theta})$. The first assumption above can be satisfied when Line 4 of Algorithm 1 is implemented with either cyclic function selection, *i.e.,* $i_k = (k \bmod m) + 1$, or implemented with a random shuffling step at the beginning of every epoch [13]. The second and the last assumptions are standard and they can be satisfied by a number of functions relevant to machine learning applications, e.g., the logistic loss function.

Let us first present the convergence result for the CIAG method — taking the parameterization $\gamma = c/(\mu + L)$ for some $0 < c \le 2$. We have:

Theorem 1 *Let Assumptions 1, 2 and 3 hold. Consider the* CIAG *method with its optimality gap sequence defined as* $V^{(k)} := \|\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star\|^2$ *where* $\boldsymbol{\theta}^\star$ *is the optimal solution to* (1). *If the step size parameter*, c, *satisfies*:

$$c < \min\Big\{ 2,\ \frac{1}{K}\sqrt{\frac{\mu L(\mu+L)}{2L_H(L^2(V^{(1)})^{\frac{1}{2}} + 4L_H^2(V^{(1)})^{\frac{3}{2}})}}, \\ \left(\frac{1}{K^4} \frac{\mu L(\mu+L)^4}{2L_H^2(L^4 V^{(1)} + 16 L_H^4 (V^{(1)})^3)} \right)^{1/5} \Big\}, \tag{28}$$

then there exists $\delta \in [1 - 2\gamma \frac{\mu L}{L+\mu}, 1)$ *such that the sequence* $\{V^{(k)}\}_{k \geq 1}$ *converges linearly*,

$$V^{(k)} \leq \delta^{\lceil (k-1)/(2K+1) \rceil} V^{(1)},\ \forall\, k \geq 1\ . \tag{29}$$

Moreover, there exists an upper bound sequence $\{\bar{V}^{(k)}\}_{k \geq 1}$, *which satisfies* $\bar{V}^{(k)} \geq V^{(k)}$ *for all* $k \geq 1$ and $\bar{V}^{(1)} = V^{(1)}$ such that

$$\lim_{k \to \infty} \frac{\bar{V}^{(k+1)}}{\bar{V}^{(k)}} \leq 1 - 2\gamma \frac{\mu L}{L+\mu} = 1 - \mathcal{O}\left(\frac{c}{\kappa(F)} \right). \tag{30}$$

Next, for the A-CIAG method, we adopt the parameters:

$$\alpha = \frac{1 - \sqrt{\mu\gamma}}{1 + \sqrt{\mu\gamma}},\ \ \gamma = \frac{c}{L}\ , \tag{31}$$

such that the extarpolation factor and step size are controlled by some $0 < c \leq 1/2$. Our result is summarized as follows:

Theorem 2 *Let Assumptions* 1, 2 *and* 3 *hold. Consider the* A-CIAG *method with its optimality gap sequence defined as* $h^{(k)} := F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star)$. *If the step size parameter*, c, *satisfies*:

$$c < \min\left\{ \bar{c}_1, \bar{c}_2, \bar{c}_3, 1/2 \right\}\ , \tag{32}$$

where

$$\begin{aligned} \bar{c}_1 &:= \left(\frac{\sqrt{\mu}}{\sqrt{18}K^2 L_H} \frac{L^2}{\frac{20L^2}{\mu}(2h^{(1)})^{\frac{1}{2}} + \left(\frac{40L_H}{\mu}\right)^2 (2h^{(1)})^{\frac{3}{2}}} \right)^{\frac{1}{2}}, \\ \bar{c}_2 &:= \left(\frac{2\mu}{81K^4 L_H^2} \frac{L^4}{\left(\frac{20L^2}{\mu}\right)^2 (2\bar{h}^{(1)}) + \left(\frac{40L_H}{\mu}\right)^4 (2h^{(1)})^3} \right)^{\frac{1}{4}}, \\ \bar{c}_3 &:= \frac{L}{\sqrt{324}K^2 L_H \frac{(h^{(1)})^{\frac{1}{2}}}{\sqrt{\mu}} + 1296 K^4 L_H^2 \frac{h^{(1)}}{\mu^2} + \mu}\ , \end{aligned} \tag{33}$$

then the optimality gap $h^{(k)}$ *satisfies*

$$h^{(k)} \leq \delta^{\lceil \frac{k}{2K+1} \rceil}(2h^{(1)}),\ \forall\, k \geq 1\ , \tag{34}$$

for some $\delta < 1$; *moreover, there exists an upper bound sequence* $\{\bar{h}^{(k)}\}_{k \geq 1}$ *such that* $\bar{h}^{(k)} \geq h^{(k)}$ *for all* $k \geq 1$ *and* $\bar{h}^{(1)} = h^{(1)}$, *such that*

$$\lim_{k \to \infty} \frac{\bar{h}^{(k+1)}}{\bar{h}^{(k)}} = 1 - \sqrt{\mu\gamma} = 1 - \sqrt{\frac{c}{\kappa(F)}}\ . \tag{35}$$

The above theorems reveal that there are two phases of convergence for the CIAG and A-CIAG methods – one that converges at a *slow* linear rate [cf. (29) and (34)], and the asymptotic phase where the algorithms converge linearly at a *fast* rate comparable to the FG and AFG methods [cf. (30) and (35)]. Such behavior is similar to the linear and superlinear convergence phases in the Newton's method [3], and they can be anticipated as the CIAG and A-CIAG methods make use of the second order information.

Due to the strong convexity of $F(\boldsymbol{\theta})$ [cf. Assumption 2], the potential function used in the two theorems above are comparable, since

$$\frac{\mu}{2}V^{(k)} \leq h^{(k)} \leq \frac{L}{2}V^{(k)} \Longrightarrow h^{(k)} = \Theta(V^{(k)})\ . \tag{36}$$

Both theorems require the step size parameters be chosen according to the initial conditions [cf. (28) and (32)]. The allowable range of step size γ is inversely proportional to L_H and $V^{(1)}$. The former term L_H measures the 'quadratic-ness' of the component functions such that $L_H = 0$ if $f_i(\boldsymbol{\theta})$ are quadratic. The latter term $V^{(1)}$ is the initial optimality gap for the algorithms, which is originated from the use of curvature information. The favorable cases are when $L_H \approx 0$ or $V^{(1)} \approx 0$ such that we can take $c \approx 1$. The latter implies the following asymptotic convergence rates:

$$\begin{aligned} \mathsf{CIAG}:\ \lim_{k \to \infty} \frac{\bar{V}^{(k+1)}}{\bar{V}^{(k)}} &= 1 - \mathcal{O}\left(\frac{1}{\kappa(F)}\right), \\ \mathsf{A-CIAG}:\ \lim_{k \to \infty} \frac{\bar{h}^{(k+1)}}{\bar{h}^{(k)}} &= 1 - \sqrt{\frac{1/2}{\kappa(F)}}\ . \end{aligned} \tag{37}$$

They coincide with the convergence rates achieved by the FG and AFG methods, respectively. Since the per-iteration complexity of CIAG, A-CIAG are $\mathcal{O}(d^2)$, while the complexity of FG, AFG are $\mathcal{O}(md)$, the advantage of the proposed methods is significant when $m \gg d$. It is also interesting to compare the convergence criterion for the CIAG and A-CIAG methods. Most noticeably, the A-CIAG criterion (33) differs from the CIAG criterion (28) for the region specified by $\tilde{c}_3$ as the A-CIAG requires the step size parameter be chosen as $c = \mathcal{O}(1/K^4)$ for large K. Since $K = \Theta(m)$, the initial optimality gap has to be as small as $h^{(1)} = \mathcal{O}(1/m^4)$ to attain the fast rate as (37). In comparison, the CIAG method only requires $V^{(1)} = \mathcal{O}(1/m^2)$ to attain (37) due to the milder requirements in (28).

When $V^{(1)} \not\approx 0$ and $L_H \not\approx 0$, we can choose a small step size parameter c at the beginning of CIAG (or A-CIAG) method that satisfies (28) (or (33)). By (29) (or (34)), for some finite T_0, the method is guaranteed to find a solution with $V^{(T_0+1)} < V^{(1)}$ after T_0 iterations. As such, we can 're-start' the method using $\boldsymbol{\theta}^{(T_0+1)}$ as the initial point with a larger step size parameter c. The procedure can then be repeated until c is increased to $c = 2$ (or $c = \frac{1}{2}$ for A-CIAG). In this way, the ideal asymptotic convergence rates in (37) are achieved. Alternatively, we can employ other incremental optimization methods such as SAG [24] to 'warm-start' the CIAG or A-CIAG methods with an initial point satisfying $V^{(1)} \approx 0$. We remark that analyzing the exact complexity of a 're-starting' schedule is non-trivial which is beyond the scope of the current paper.

Nevertheless, from our numerical experiments in Sect. 5, we find that the practical step sizes can often be chosen more aggressively. For a wide range of problems, using a constant step size parameter c, the CIAG and A-CIAG methods converge quickly in terms of the number of iterations used and CPU time, without the aforementioned re-starting or warm-starting techniques.

4.1 Proofs of Theorem 1 and 2

The analysis for CIAG and A-CIAG methods consists of three steps.

- First, we carry out a perturbation analysis on the FG and AFG methods. Effectively, we view the CIAG (resp. A-CIAG) method as a perturbed FG (resp. AFG) method with inexact gradients, where the errors are defined as:

$$\boldsymbol{e}^k_{\mathsf{CIAG}} := \boldsymbol{g}^k_{\mathsf{CIAG}} - \sum_{i=1}^m \nabla f_i(\boldsymbol{\theta}^k),\ \boldsymbol{e}^k_{\mathsf{ACIAG}} := \boldsymbol{g}^k_{\mathsf{ACIAG}} - \sum_{i=1}^m \nabla f_i(\boldsymbol{\theta}^k_{ex}) \,. \tag{38}$$

- Second, we analyze an upper bound on $\|\boldsymbol{e}^k_{\mathsf{CIAG}}\|$ or $\|\boldsymbol{e}^k_{\mathsf{ACIAG}}\|$ in terms of

$$V^{(k)} := \|\boldsymbol{\theta}^k - \boldsymbol{\theta}^\star\|^2,\ \ h^{(k)} := F(\boldsymbol{\theta}^k) - F(\boldsymbol{\theta}^\star) \,, \tag{39}$$

  where we shall use $V^{(k)}$, $h^{(k)}$ as the potential functions for CIAG method and A-CIAG method, respectively. Here, we exploited Assumption 3 on the Lipschitz continuity of Hessian for the component functions.
- Third, using the bounds derived in the previous steps we study a nonlinear inequality system to derive the convergence criterion of $V^{(k)}$ and $h^{(k)}$. The nonlinear inequality exhibits the desired R-linear convergence when $k \to \infty$.

Overall, the key to our proof is to analyze the dynamics of optimality gap $V^{(k)}$ (or $h^{(k)}$) as a nonlinear system, whose convergence can be guaranteed by an appropriately chosen step size γ and the asymptotic convergence rate will only depend on the linear terms in the optimality gaps.

In the following, we assume that the CIAG, ACIAG methods are both initialized such that (23) holds. This can be done 'on-the-fly' with a self-initialization step in (15) and the analysis below will hold for all $k \geq K$, *i.e.,* after a complete pass through the dataset.

Step 1. The first step is to analyze the CIAG (resp. A-CIAG) method as a perturbed version of the FG (resp. AFG) method, which employs $\boldsymbol{g}^k_{\mathsf{CIAG}}$ (resp. $\boldsymbol{g}^k_{\mathsf{ACIAG}}$) as the gradient surrogate. We have:

Proposition 1 *Consider the* CIAG *method. Under Assumption 2, if $\gamma \leq 2/(\mu + L)$, we have that for all $k \geq 1$,*

$$V^{(k+1)} \leq \left(1 - 2\gamma \frac{\mu L}{\mu + L}\right) V^{(k)} + \gamma^2 \|\boldsymbol{e}^k_{\mathsf{CIAG}}\|^2 + 2\gamma \sqrt{V^{(k)}} \|\boldsymbol{e}^k_{\mathsf{CIAG}}\| \,. \tag{40}$$

The proof largely follows from [14, Section 3.3] and is omitted.

Proposition 2 *Consider the* A-CIAG *method. Under Assumption 2, if $\gamma \leq 1/(2L)$, we have that for all $k \geq 1$,*

$$\begin{aligned} h^{(k+1)} \leq 2(1 - \sqrt{\mu\gamma})^k h^{(1)} + \sum_{\ell=1}^{k} (1 - \sqrt{\mu\gamma})^{k-\ell} \Bigg(\sqrt{2\gamma h^{(\ell)}} \|\boldsymbol{e}^\ell_{\mathsf{ACIAG}}\| \\ + \sqrt{\frac{9\gamma}{\mu}} \|\boldsymbol{e}^\ell_{\mathsf{ACIAG}}\|^2 - \frac{\mu}{4} \frac{1 - \mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}^\ell - \boldsymbol{\theta}^\ell_{ex}\|^2 \Bigg) . \end{aligned} \tag{41}$$

From the propositions, we observe that when the gradient errors vanish with $\boldsymbol{e}^\ell = \boldsymbol{0}$, setting $\gamma = 2/(\mu + L)$ (resp. $\gamma = 1/(2L)$) gives $V^{(k+1)} = \mathcal{O}((1 - 1/\kappa(F))^k)$ (resp. $h^{(k+1)} = \mathcal{O}((1 - \sqrt{1/2\kappa(F)})^k)$). In other words, the linear convergence rates for FG and AFG methods can be recovered.

Comparing the two propositions reveals the differences in error structure between the CIAG and A-CIAG methods. First, in (41) the A-CIAG's error is convolved with the linearly converging sequence $(1 - \sqrt{\mu\gamma})^\ell$, while in (40) the CIAG's error is simply additive; second, (41) consists of a negative term depending on $\|\boldsymbol{\theta}^\ell - \boldsymbol{\theta}^\ell_{ex}\|^2$. This negative term revealed through our refined analysis of the error dynamics of A-CIAG [see (81)] and the resultant proposition is an improvement over [25]. Significantly, this negative term is an important key for establishing the fast convergence of A-CIAG.

Step 2. Our next step relates the gradient errors $\boldsymbol{e}^{\ell}_{\mathsf{CIAG}}$, $\boldsymbol{e}^{\ell}_{\mathsf{ACIAG}}$ to the optimality gaps $V^{(\ell)}$, $h^{(\ell)}$ for the CIAG, A-CIAG methods, respectively. We obtain bounds for the errors as follows:

Proposition 3 *Consider the* CIAG *method. Under Assumptions 1, 2 and 3, we have that for all* $\ell \geq 1$,

$$\|\boldsymbol{e}^{\ell}_{\mathsf{CIAG}}\| \leq \gamma^2 K^2 L_H \left(L^2 \max_{(\ell-K)_{++} \leq q \leq \ell-1} V^{(q)} + 4L_H^2 \max_{(\ell-2K)_{++} \leq q \leq \ell-1} (V^{(q)})^2 \right). \tag{42}$$

Proposition 4 *Consider the* A-CIAG *method. Under Assumptions 1, 2 and 3, we have that for all* $\ell \geq 1$,

$$\begin{aligned} \|\boldsymbol{e}^{\ell}_{\mathsf{ACIAG}}\| \leq & \frac{3KL_H}{2} \sum_{j=(\ell-K)_{++}}^{\ell-1} \|\boldsymbol{\theta}^{j+1} - \boldsymbol{\theta}^{j+1}_{ex}\|^2 \\ & + \gamma^2 \frac{3K^2 L_H}{2} \frac{20L^2}{\mu} \max_{(\ell-K-1)_{++} \leq q \leq \ell-1} h^{(q)} \\ & + \gamma^2 \frac{3K^2 L_H}{2} \left(\frac{40 L_H}{\mu} \right)^2 \max_{(\ell-2K-1)_{++} \leq q \leq \ell-1} (h^{(q)})^2 \, . \end{aligned} \tag{43}$$

We observe that the upper bounds for $\|\boldsymbol{e}^{k}_{\mathsf{CIAG}}\|$ and $\|\boldsymbol{e}^{k}_{\mathsf{ACIAG}}\|$ obey similar structure in terms of their dependences on $V^{(q)}$ and $h^{(q)}$. The upper bound for $\|\boldsymbol{e}^{\ell}_{\mathsf{ACIAG}}\|$ depends on $\{\|\boldsymbol{\theta}^{j+1} - \boldsymbol{\theta}^{j+1}_{ex}\|^2\}_{j=(\ell-K)_{++}}^{\ell-1}$ which is the difference between the extrapolated variables and the main variables.

Step 3. We omit the constants which are obtained from the propositions given in the previous steps. Instead, we focus on the main steps in the analysis and relegate the exact analysis to Appendices B and F.

For the CIAG method, simply substituting (42) into (40) yields:

$$\begin{aligned} V^{(k+1)} \leq & \left(1 - 2\gamma \frac{\mu L}{\mu + L}\right) V^{(k)} + \mathcal{O}(\gamma^3) \times \\ & \max_{(k-2K)_{++} \leq q \leq k} \left((V^{(q)})^{\frac{3}{2}} + (V^{(q)})^2 + (V^{(q)})^{\frac{5}{2}} + (V^{(q)})^4 \right) , \end{aligned} \tag{44}$$

where the exact form of the system will be shown in (63). Define

$$V^{(k)}_{\max} := \max_{(k-2K)_{++} \leq q \leq k} V^{(q)} \tag{45}$$

and using the fact that, when $V^{(k)}$ is small, the second part in the right hand side of (44) can be bounded by its lowest order term $\mathcal{O}(\gamma^3)(V^{(k)}_{\max})^{\frac{3}{2}}$, we have:

$$V^{(k+1)} \leq \left(1 - 2\gamma \frac{\mu L}{\mu + L}\right) V^{(k)} + \mathcal{O}(\gamma^3)(V^{(k)}_{\max})^{\frac{3}{2}} \, . \tag{46}$$

On the other hand, for the A-CIAG method, substituting (43) into (41) and rearranging terms show that the $(k+1)$th optimality gap is bounded as:

$$
\begin{aligned}
h^{(k+1)} &\leq 2(1-\sqrt{\mu\gamma})^k h^{(1)} + \sum_{\ell=1}^{k} (1-\sqrt{\mu\gamma})^{k-\ell} \times \\
&\Big\{ \mathcal{O}(\gamma^{\frac{5}{2}}) \max_{(\ell-2K-1)_{++} \leq q \leq \ell} \Big((h^{(q)})^{\frac{3}{2}} + (h^{(q)})^2 + (h^{(q)})^{\frac{5}{2}} + (h^{(q)})^4 \Big) \\
&\quad + \Big(\max_{\ell \leq q \leq \min\{\ell+K,k\}} \mathcal{O}(\sqrt{\gamma h^{(q)}}) - \frac{\mu}{4} \frac{1-\mu\gamma}{\sqrt{\mu\gamma}} \Big) \|\boldsymbol{\theta}^\ell - \boldsymbol{\theta}_{ex}^\ell\|^2 \Big\} ,
\end{aligned} \tag{47}
$$

whose exact form can be found in (96). We emphasize that the last term on the right hand side depends on the difference between $\boldsymbol{\theta}^\ell$ and $\boldsymbol{\theta}_{ex}^\ell$, which is unique for A-CIAG method due to the use of extrapolated iterates.

To finish the proof, we identify that in both (46) and (47), the potential functions for CIAG and A-CIAG methods, *i.e.,* $V^{(k+1)}$ and $h^{(k+1)}$, are upper bounded by — a constant factor (< 1) multiplied by the previous potential function's value; and a *high-order* term that depends on the delayed version of the potential function's value. With a sufficiently small step size, the effects from the later term vanishes as $k \to \infty$ and the proposed methods converge linearly at the desired rates.

Particularly, in the case of CIAG, we consider the *non-linear* inequality:

$$
R^{(k+1)} \leq pR^{(k)} + \sum_{j=1}^{J} q_j \max_{k' \in [(k-M+1)_{++}, k]} (R^{(k')})^{\eta_j}, \ \forall\, k \geq 1 \ , \tag{48}
$$

where $0 \leq p < 1$, $q_j \geq 0$, $\eta_j > 1$ for all j with some $J, M < \infty$. We have:

Proposition 5 *Consider (48). For some $p \leq \delta < 1$, if*

$$
p + \textstyle\sum_{j=1}^{J} q_j (R^{(1)})^{\eta_j - 1} \leq \delta < 1 \ , \tag{49}
$$

then (a) $\{R^{(k)}\}_{k\geq 1}$ converges linearly as $R^{(k)} \leq \delta^{\lceil k/M \rceil} R^{(1)}$ for all $k \geq 1$; (b) there exists an upper bound sequence $\{\bar{R}^{(k)}\}_{k\geq 1}$ with $\bar{R}^{(k)} \geq R^{(k)}$ for all $k \geq 1$ and $\bar{R}^{(1)} = R^{(1)}$, that converges linearly at rate p asymptotically,

$$
\lim_{k\to\infty} \bar{R}^{(k+1)} / \bar{R}^{(k)} = p \ . \tag{50}
$$

Consequently, Theorem 1 can be proven by identifying that $R^{(k)} = V^{(k)}$ and substituting the appropriate constants in Proposition 5, see Appendix B.

In the case of A-CIAG, it can be verified that under our choice of step size, the coefficient in front of $\|\boldsymbol{\theta}^\ell - \boldsymbol{\theta}_{ex}^\ell\|^2$ is always negative for all $\ell \geq 1$. Define

$$
h_{\max}^{(\ell)} := \max_{(\ell-2K-1)_{++} \leq q \leq \ell} h^{(q)} \ . \tag{51}
$$

When $h_{\max}^{(\ell)}$ is small, the terms inside the last bracket of the summation in (47) can be bounded by its lowest order term as $\mathcal{O}(\gamma^{\frac{5}{2}})(h_{\max}^{(\ell)})^{\frac{3}{2}}$. Therefore, we consider an upper bound sequence $\{\bar{h}^{(k)}\}_{k\geq 1}$ with $\bar{h}^{(k)} \geq h^{(k)}$ for all k and:

$$\bar{h}^{(k+1)} = 2(1-\sqrt{\mu\gamma})^k \bar{h}^{(1)} + \sum_{\ell=1}^{k} (1-\sqrt{\mu\gamma})^{k-\ell} \mathcal{O}(\gamma^{\frac{5}{2}})(\bar{h}_{\max}^{(\ell)})^{\frac{3}{2}} \ . \tag{52}$$

Subtracting $(1-\sqrt{\mu\gamma})\bar{h}^{(k)}$ from both sides gives:

$$\bar{h}^{(k+1)} = (1-\sqrt{\mu\gamma})\bar{h}^{(k)} + \mathcal{O}(\gamma^{\frac{5}{2}})(\bar{h}_{\max}^{(k)})^{\frac{3}{2}} \ , \tag{53}$$

which resembles (46) in the case of CIAG. Similar to the previous developments, we expect the system to converge linearly at the rate $1-\sqrt{\mu\gamma}$.

Formally, consider an abstracted form of (47) with the non-negative sequence $\{R^{(k)}\}_{k\geq 1}$ that satisfies:

$$\begin{aligned} R^{(k+1)} \leq p^k b R^{(1)} + \sum_{\ell=1}^{k} p^{k-\ell} \Bigg\{ & \sum_{j=1}^{J} s_j \max_{(\ell-M)_{++}\leq q\leq \ell} (R^{(q)})^{\eta_j} \\ & + \Big(\max_{\ell\leq q\leq k} f(R^{(q)}) - \bar{f} \Big) D^{(\ell)} \Bigg\}, \end{aligned} \tag{54}$$

for all $k \geq 1$, where $f(R^{(q)})$ is a non-decreasing function of $R^{(q)}$ and $\eta_j > 1$ for all j. The parameters $p, s_j, \bar{f}, b$ are all non-negative, we have $b \geq 1$ and $M < \infty$, and $\{D^{(\ell)}\}_{\ell\geq 1}$ is an arbitrary non-negative sequence. The above system converges linearly at a rate given by the constant factor $p < 1$:

Proposition 6 *Consider* (54). *Suppose that*

$$\bar{f} \geq f(bR^{(1)}) \quad \text{and} \quad \delta := p + \sum_{j=1}^{J} s_j (bR^{(1)})^{\eta_j - 1} < 1 \ . \tag{55}$$

Then, there exists an upper bound sequence $\{\bar{R}^{(k)}\}_{k\geq 1}$ *satisfying*

$$\begin{aligned} &\text{(i) } \bar{R}^{(k)} \geq R^{(k)}, \ \forall\, k \geq 1, \ \text{(ii) } \bar{R}^{(k+1)} \leq \delta^{\lceil k/M \rceil} (b\bar{R}^{(1)}), \ \forall\, k \geq 1 \\ &\text{and (iii) } \lim_{k\to\infty} \bar{R}^{(k+1)} / \bar{R}^{(k)} = p \ . \end{aligned} \tag{56}$$

Finally, Theorem 2 can be proven by identifying $R^{(k)} = h^{(k)}$ and substituting the appropriate constants, see Appendix F.

5 Numerical experiments

This section covers the performance of CIAG and A-CIAG using numerical experiments. We focus on the logistic regression problem for training linear classifiers. We are given m data tuples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the feature vector and $y_i \in \{\pm 1\}$ is the label. The ith component function is:

$$f_i(\boldsymbol{\theta}) = \frac{1}{2m} \|\boldsymbol{\theta}\|^2 + \log(1 + \exp(-y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)) \ . \tag{57}$$

This function has the form of a linear model in (24) and it also satisfies Assumptions 2 and 3. For instance, an upper bound to the gradient and Hessian smoothness of $F(\boldsymbol{\theta})$ and $f_i(\boldsymbol{\theta})$, respectively, can be evaluated as:

$$L = 1 + \frac{1}{4}\sum_{i=1}^{m} \|\boldsymbol{x}_i\|_2^2, \;\; L_{H,i} = \frac{1}{4}\|\boldsymbol{x}_i\|_2^2 \,. \tag{58}$$

5.1 Synthetic data

We adopt a simple random data model. First, we generate $\boldsymbol{\theta}_{\text{true}} \sim \mathcal{U}[-1,1]^d$ and the feature vector as $\boldsymbol{x}_i = [\tilde{\boldsymbol{x}}_i; 1]$ where $\tilde{\boldsymbol{x}}_i \sim \mathcal{U}[-1,1]^{d-1}$; then, the label is computed as $y_i = \text{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_{\text{true}} \rangle)$.

To set up the benchmark, the step sizes for NIM and IQN are $\gamma = 1$. For the IAG method, we set $\gamma = 50/(mL)$. For the CIAG and A-CIAG methods, we set $\gamma = 1/L$ and we set the extrapolation weight for A-CIAG as $\alpha = 0.95$. The above methods are implemented with deterministic, cyclic component function selection, *i.e.,* $i_k = (k \bmod m) + 1$. We also compare a few stochastic methods: for SAG and AccSVRG, we set $\gamma = 50/(mL)$; and the batch size is $B = 5$ for AccSVRG with an epoch length of m. For SVRG2, we set an epoch length of $0.1m$. The step sizes for NIM and SVRG2 will be specified later.

The evolution of the optimality gap against the number of effective passes through data are shown in Fig. 1 for different problem sizes. We defined the number of effective passes as the number of *iterations* (k) divided by m. From Fig. 1, we observe that both NIM and IQN methods have the fastest convergence, since both methods are shown to converge superlinearly. However, we note that the curvature aided methods, A-CIAG, CIAG and SVRG2, also demonstrate similar convergence speed in terms of the number of effective data passes used. Especially, the speed of the proposed A-CIAG almost matches that of the NIM method.

5.2 Real data

We empirically compare the algorithms on the datasets from `LibSVM` [8]. For this example, the algorithms are implemented in `C++` based on the source codes by [23] (available: http://github.com/arodomanov/nim_icml16_code) to demonstrate the fastest possible practical performance. We only compare the proposed CIAG, A-CIAG to IAG, NIM and SAG, where the first four algorithms employ a deterministic, cyclic component function selection and the last algorithm employs random component function selection. For the NIM method, we tested both of its *exact* and *inexact* setting, where the *inexact* setting is a double-loop method which uses a conjugate gradient method to tackle the Hessian inverse involved. We have used a mini-batch setting for all the tested methods such that each $f_i(\boldsymbol{\theta})$ is composed of $B = 5$ data tuples. The numerical experiments were

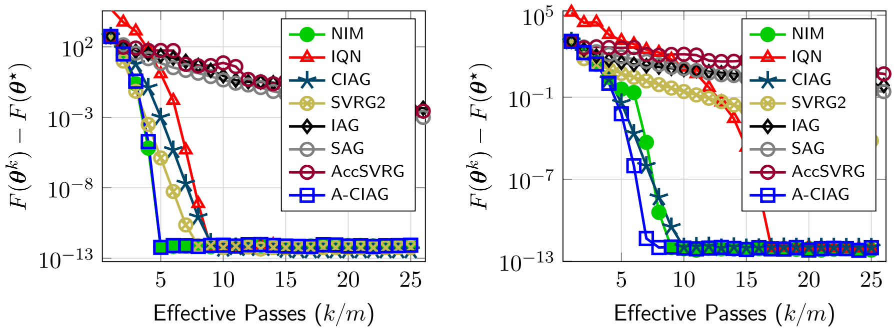

Fig. 1 The y-axis denotes the optimality gap plotted in log-scale and the x-axis shows the number of effective passes (defined as k/m). (Left) for the case with $m = 1000$, $d = 51$. (Right) for the case with $m = 2000$, $d = 501$. Here, for NIM and SVRG2 we use the step sizes $\gamma = 0.01$ and $\gamma = 0.1/L$, respectively

conducted on a Laptop computer with an Intel Core i7 2.8Ghz quad-core processor and 16 gigabytes of memory.

An overview of the performance comparison can be found in Table 2, while Table 3 shows the algorithms' parameter settings used for different dataset. As seen, the A-CIAG method outperformed the benchmarks for many of the considered datasets in terms of the wall clock convergence time, and the number of effective passes required is comparable to the best available method, NIM. An exception is the experiment on the `mnist` dataset. We suspect that this is due to the potentially poor condition number with the `mnist` dataset, as we observe that the unaccelerated methods SAG, IAG and CIAG exhibit significantly slower convergence than for the other datasets.

To investigate the behavior of the algorithms, Fig. 2 shows the evolution of gradient's norm $\|\nabla F(\theta^k)\|$ against the number of effective passes and wall clock time on the tested datasets. As seen, the convergence speed of A-CIAG matches that of the Newton-based NIM, yet the wall clock time required is faster as it does not involve computing the Hessians' inverse. It is worthwhile pointing out that except for the datasets `covtype` and `alpha`, the SAG method achieves a solution accuracy of $\|\nabla F(\theta^k)\| \leq 10^{-6}$ in less wall clock time than the proposed methods. The benefits of our proposed methods are significant when the solution accuracy is high, as predicted by our theoretical results.

6 Conclusions

We proposed two new optimization methods utilizing curvature-aided gradient tracking for large-scale optimization via incremental data processing. The proposed methods, CIAG and A-CIAG, attain an ϵ-optimal solution with only $\mathcal{O}(\kappa(F)\log(1/\epsilon))$ and $\mathcal{O}(\sqrt{\kappa(F)}\log(1/\epsilon))$ *iterations*, respectively, for a small ϵ, and provided that the initial point is close to the optimal solution. Numerical

Table 2 Performance comparison

Dataset	A-CIAG	CIAG	NIM	SAG[a]
`mushrooms`	**5.22 pass**	43.5 pass	4.81 (4.92) pass	359.9 pass
($m = 8124, d = 112$)	**0.299 sec.**	2.509 sec.	1.01 (0.329) sec.	1.521 sec.
`a9a`	**3.6 pass**	52.2 pass	3.0 (3.2) pass	165.8 pass
($m = 32561, d = 123$)	**1.067 sec.**	15.26 sec.	3.38 (1.10) sec.	3.685 sec.
`SUSY`	**7.1 pass**	7.6 pass	6.2 (6.2) pass	52.3 pass
($m = 5 \times 10^6, d = 18$)	**24.88 sec.**	26.81 sec.	35.92 (29.82) sec.	99.00 sec.
`covtype`	**4.5 pass**	13.5 pass	4.0 (4.5) pass	101.9 pass
($m = 581012, d = 54$)	**5.888 sec.**	17.71 sec.	13.84 (7.06) sec.	32.33 sec.
`w8a`	**5.5 pass**	7.2 pass	5.3 (5.4) pass	251.01 pass
($m = 49749, d = 300$)	**12.73 sec.**	16.48 sec.	69.13 (13.82) sec.	23.20 sec.
`mnist`	4.3 pass	143.6 pass	3.8 (**3.8**) **pass**	$\geq 10^3$ pass
($m = 60000, d = 784$)	89.59 sec.	2801 sec.	755.2 (**86.94**) **sec.**	$\geq$ 392 sec.
`alpha`	**2.4 pass**	7.6 pass	2.3 (2.5) pass	80.5 pass
($m = 5 \times 10^5, d = 500$)	**149.5 sec.**	475.6 sec.	1111.6 (176.4) sec.	210.7 sec.

We show the number of effective passes (defined as k/m) and the wall clock time required to reach convergence with $\|\nabla F(\theta^k)\| \leq 10^{-10}$. For the NIM method, we tested both 'exact' and 'inexact' settings in Hessian inverse. Results inside the brackets $(\cdot)$ correspond to 'inexact' setting

Bold values highlight the best performance for each dataset

[a]Results for SAG are averaged over 10 trials

Table 3 Parameters used for A-CIAG, CIAG methods in the numerical experiments

	`mushrooms`	`a9a`	`SUSY`	`covtype`	`w8a`	`mnist`	`alpha`
A-CIAG	$\gamma = \frac{10^{-3}}{L}$ $\alpha = 0.99$	$\gamma = \frac{10^{-4}}{L}$ $\alpha = 0.99$	$\gamma = \frac{10^{-5}}{L}$ $\alpha = 0.99$	$\gamma = \frac{5\cdot 10^{-6}}{L}$ $\alpha = 0.99$	$\gamma = \frac{10^{-3}}{L}$ $\alpha = 0.99$	$\gamma = \frac{10^{-4}}{L}$ $\alpha = 0.99$	$\gamma = \frac{5\cdot 10^{-6}}{L}$ $\alpha = 0.99$
CIAG	$\gamma = \frac{10^{-3}}{L}$	$\gamma = \frac{2\cdot 10^{-4}}{L}$	$\gamma = \frac{10^{-5}}{L}$	$\gamma = \frac{5\cdot 10^{-6}}{L}$	$\gamma = \frac{10^{-3}}{L}$	$\gamma = \frac{10^{-4}}{L}$	$\gamma = \frac{10^{-5}}{L}$

experiments on real and synthetic data demonstrate the benefit of our algorithms. Future work includes designing and analyzing an optimal 're-starting' scheme to achieve the ideal convergence rate (37) when the initial point is not close to the optimal solution, and extending the methods to tackling non-smooth optimization problems.

Acknowledgements This work has been partially supported by the NSF Grant CCF-1717391 and CUHK Direct Grant #4055113.

A Proof of Proposition 3

Let us express the gradient error as

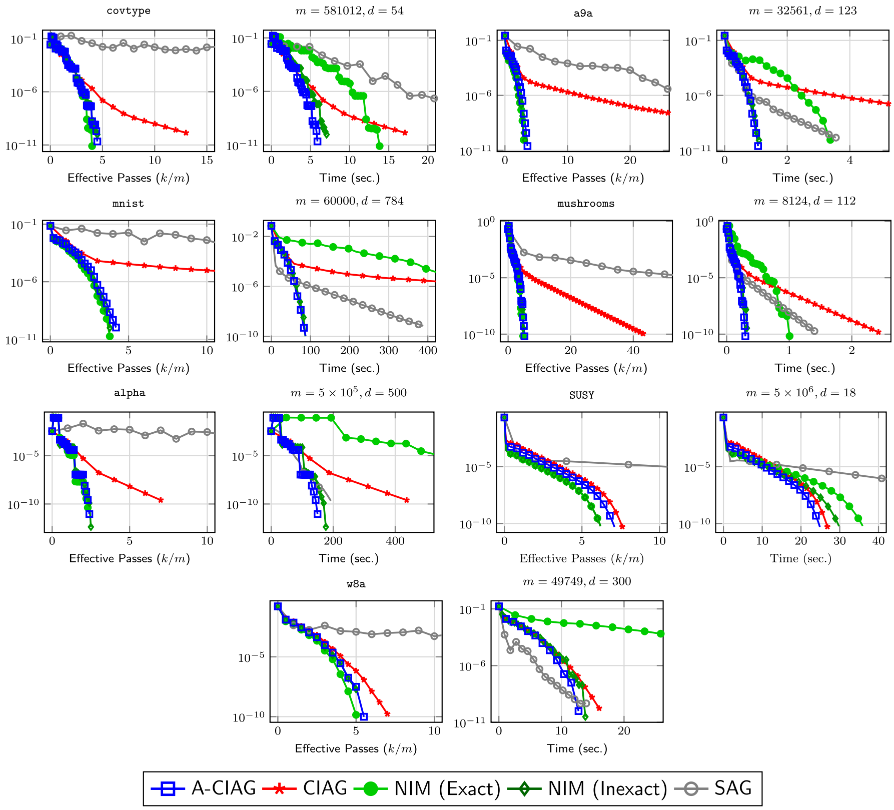

Fig. 2 Evolution of $\|\nabla F(\boldsymbol{\theta}^k)\|$ (y-axis) against number of effective passes and wall clock time on the datasets. The experiment settings are the same as in Table 2

$$\boldsymbol{e}^k_{\mathsf{CIAG}} = \sum_{i=1}^m \Big(\nabla f_i(\boldsymbol{\theta}^{\tau_i^k}) + \nabla^2 f_i(\boldsymbol{\theta}^{\tau_i^k})(\boldsymbol{\theta}^k - \boldsymbol{\theta}^{\tau_i^k}) - \nabla f_i(\boldsymbol{\theta}^k) \Big). \tag{59}$$

Applying Lemma 1:

$$\begin{aligned} \|\boldsymbol{e}^k_{\mathsf{CIAG}}\| &\leq \sum_{i=1}^m \frac{L_{H,i}}{2} \|\boldsymbol{\theta}^{\tau_i^k} - \boldsymbol{\theta}^k\|^2 \leq \sum_{i=1}^m \frac{L_{H,i}}{2} \underbrace{(k-\tau_i^k)}_{\leq K} \sum_{j=\tau_i^k}^{k-1} \|\boldsymbol{\theta}^{j+1} - \boldsymbol{\theta}^j\|^2 \\ &\leq \frac{KL_H}{2} \sum_{j=(k-K)_{++}}^{k-1} \|\boldsymbol{\theta}^{j+1} - \boldsymbol{\theta}^j\|^2 \leq \frac{KL_H}{2}\gamma^2 \sum_{j=(k-K)_{++}}^{k-1} \|\boldsymbol{e}^j_{\mathsf{CIAG}} + \nabla F(\boldsymbol{\theta}^j)\|^2 \\ &\leq \gamma^2 K L_H \sum_{j=(k-K)_{++}}^{k-1} \Big(\|\boldsymbol{e}^j_{\mathsf{CIAG}}\|^2 + \|\nabla F(\boldsymbol{\theta}^j)\|^2 \Big) . \end{aligned} \tag{60}$$

Furthermore, we have

$$\|\nabla F(\boldsymbol{\theta}^j)\|^2 = \|\nabla F(\boldsymbol{\theta}^j) - \nabla F(\boldsymbol{\theta}^\star)\|^2 \leq L^2 V^{(j)}, \tag{61}$$

$$\|\boldsymbol{e}^j_{\mathsf{CIAG}}\| \overset{(a)}{\leq} \sum_{i=1}^m L_{H,i}\Big(V^{(j)} + V^{(\tau_i^j)}\Big) \leq 2L_H \max_{\ell \in \{\tau_i^j\}_{i=1}^m \cup \{j\}} V^{(\ell)} , \tag{62}$$

where (a) is due to $\|\boldsymbol{a} - \boldsymbol{b}\|^2 \leq 2(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2)$. Plugging these back into (60) and using $\tau_i^{k-K} \geq k - 2K$ gives:

$$\begin{aligned} \|\boldsymbol{e}^k_{\mathsf{CIAG}}\| &\leq \gamma^2 K L_H \sum_{j=(k-K)_{++}}^{k-1} \left(L^2 V^{(j)} + \left(2L_H \max_{\ell \in \{\tau_i^j\}_{i=1}^m \cup \{j\}} V^{(\ell)} \right)^2 \right) \\ &\leq \gamma^2 K^2 L_H \left(L^2 \max_{(k-K)_{++} \leq \ell \leq k-1} V^{(\ell)} + 4L_H^2 \max_{(k-2K)_{++} \leq \ell \leq k-1} (V^{(\ell)})^2 \right) . \end{aligned} \tag{63}$$

B Step 3 in the Proof of Theorem 1

Combining Proposition 1 and 3 yields

$$\begin{aligned} V^{(k+1)} \leq &\left(1 - 2\gamma \frac{\mu L}{\mu + L}\right) V^{(k)} \\ &+ 2\gamma^3 K^2 L_H \left(L^2 \max_{(k-K)_{++} \leq \ell \leq k} (V^{(\ell)})^{\frac{3}{2}} + 4L_H^2 \max_{(k-2K)_{++} \leq \ell \leq k} (V^{(\ell)})^{\frac{5}{2}} \right) \\ &+ 2\gamma^6 K^4 L_H^2 \left(L^4 \max_{(k-K)_{++} \leq \ell \leq k-1} (V^{(\ell)})^2 + 16L_H^4 \max_{(k-2K)_{++} \leq \ell \leq k-1} (V^{(\ell)})^4 \right), \end{aligned} \tag{64}$$

which is the exact form for Eq. (44). The right hand side in (64) can be decomposed into two terms—the first term is of the same order as $V^{(k)}$, and the other terms are *delayed* and *higher-order* terms of $V^{(\ell)}$.

Observe that (64) is a special case of (48) in Proposition 5 with $R^{(k)} = V^{(k)}$, $M = 2K + 1$, $p = 1 - 2\gamma \mu L/(\mu + L)$ and

$$\begin{aligned} &q_1 = 2\gamma^3 K^2 L^2 L_H,\ \eta_1 = 3/2,\ q_2 = 8\gamma^3 K^2 L_H^3,\ \eta_3 = 5/2 , \\ &q_3 = 2\gamma^6 K^4 L_H^2 L^4,\ \eta_3 = 2,\ q_4 = 32\gamma^6 K^4 L_H^6,\ \eta_4 = 4 . \end{aligned} \tag{65}$$

The corresponding convergence condition in (49) can be satisfied if

$$\begin{aligned} &\gamma^5\ 2K^4 L_H^2 \big(L^4 V^{(1)} + 16L_H^4 (V^{(1)})^3\big) < \frac{\mu L}{\mu + L} \\ &\text{and}\ \ \gamma^2\ 2K^2 L_H \big(L^2 (V^{(1)})^{1/2} + 4L_H^2 (V^{(1)})^{3/2}\big) < \frac{\mu L}{\mu + L} , \end{aligned} \tag{66}$$

which can be implied by (28). The proof is thus concluded.

C Proof of Proposition 5

The proof of the proposition is divided into two parts. We first show that under (49), the sequence $\{R^{(k)}\}_{k\geq 1}$ converges linearly as in part (a) of the proposition; then we show that the rate of convergence is asymptotically given by p as in part (b) of the proposition [cf. (50)].

The first part of the proof is achieved using induction on all $\ell \geq 1$ with:

$$R^{(k)} \leq \delta^{\ell}\ R^{(1)},\ \forall\ k = (\ell - 1)M + 2, ..., \ell M + 1\ . \tag{67}$$

The base case when $\ell = 1$ can be straightforwardly established:

$$\begin{aligned} &R^{(2)} \leq pR^{(1)} + \textstyle\sum_{j=1}^{J} q_j (R^{(1)})^{\eta_j} \leq \delta R^{(1)}\ , \\ &\vdots \\ &R^{(M+1)} \leq pR^{(M)} + \textstyle\sum_{j=1}^{J} q_j (R^{(0)})^{\eta_j} \leq \delta R^{(1)}\ . \end{aligned} \tag{68}$$

Suppose that the statement (67) is true up to $\ell = c$, for $\ell = c + 1$, we have:

$$\begin{aligned} R^{(cM+2)} &\leq pR^{(cM+1)} + \sum_{j=1}^{J} q_j \max_{k' \in [(c-1)M+2, cM+1]} (R^{(k')})^{\eta_j} \\ &\leq p\left(\delta^c R^{(1)}\right) + \sum_{j=1}^{J} q_j \left(\delta^c R^{(1)}\right)^{\eta_j} \leq \delta^c \left(pR^{(1)} + \sum_{j=1}^{J} q_j (R^{(1)})^{\eta_j} \right) \leq \delta^{c+1} R^{(1)}\ . \end{aligned}$$

Similar statement also holds for $R^{(k)}$ with $k = cM + 3, ..., (c + 1)M + 1$. We thus conclude with:

$$R^{(k)} \leq \delta^{\lceil (k-1)/M \rceil}\ R^{(1)},\ \forall\ k \geq 1\ , \tag{69}$$

which proves the first part of the proposition.

The second part of the proof establishes the asymptotic linear rate of convergence in (50). We consider the upper bound sequence $\{\bar{R}^{(k)}\}_{k\geq 1}$ such that $\bar{R}^{(1)} = R^{(1)}$ and the inequality (48) is tight for $\{\bar{R}^{(k)}\}_{k\geq 1}$. Obviously, it also holds that $\bar{R}^{(k)} \leq \delta^{\lceil (k-1)/M \rceil} \bar{R}^{(1)}$ for all $k \geq 1$. Now, observe that

$$\frac{\bar{R}^{(k+1)}}{\bar{R}^{(k)}} = p + \frac{\sum_{j=1}^{J} q_j \max_{k' \in [(k-M+1)_{++}, k]} (R^{(k')})^{\eta_j}}{\bar{R}^{(k)}}\ . \tag{70}$$

For any $k' \in [k - M + 1, k]$ and any $\eta > 1$, we have:

$$\frac{(\bar{R}^{(k')})^{\eta}}{\bar{R}^{(k)}} = \frac{\bar{R}^{(k')}}{\bar{R}^{(k)}} (\bar{R}^{(k')})^{\eta-1} \leq \frac{\bar{R}^{(k')}}{\bar{R}^{(k)}} (R^{(1)})^{\eta-1} \delta^{(\lceil \frac{k'-1}{M} \rceil)(\eta-1)}\ . \tag{71}$$

Note that as $\bar{R}^{(k+1)}/\bar{R}^{(k)} \geq p$, we have:

$$\frac{(\bar{R}^{(k')})^{\eta}}{\bar{R}^{(k)}} \leq p^{-M}(R^{(1)})^{\eta-1}\delta^{(\lceil\frac{k'-1}{M}\rceil)(\eta-1)} \,. \tag{72}$$

Taking $k \to \infty$ shows that the right hand side vanishes. As a result, we have $\lim_{k\to\infty} \bar{R}^{(k+1)}/\bar{R}^{(k)} = p$. This proves part (b) of the proposition.

D Proof of Proposition 2

The following proof is partially inspired by [7, 21, 25]. For simplicity, we drop the subscript ACIAG in $\boldsymbol{g}^k_{\mathsf{ACIAG}}$ and $\boldsymbol{e}^k_{\mathsf{ACIAG}}$. Define $\rho := 1 - \sqrt{\mu\gamma}$ and the estimation sequence as:

$$\begin{aligned}
\boldsymbol{\Phi}_1(\boldsymbol{\theta}) &:= F(\boldsymbol{\theta}^1_{ex}) + \frac{\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^1_{ex}\|^2 \\
\boldsymbol{\Phi}_{k+1}(\boldsymbol{\theta}) &:= \rho\, \boldsymbol{\Phi}_k(\boldsymbol{\theta}) + \sqrt{\mu\gamma}\Big(F(\boldsymbol{\theta}^k_{ex}) + \langle \boldsymbol{g}^k, \boldsymbol{\theta} - \boldsymbol{\theta}^k_{ex}\rangle + \frac{\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^k_{ex}\|^2\Big) \,,
\end{aligned} \tag{73}$$

where $\boldsymbol{g}^k := \boldsymbol{b}^k + \boldsymbol{H}^k\boldsymbol{\theta}^k_{ex}$ is the gradient surrogate used in (17). Recall that $\boldsymbol{e}^k := \boldsymbol{g}^k - \nabla F(\boldsymbol{\theta}^k_{ex})$ is the gradient error. The following inequality, which holds for all $\boldsymbol{\theta} \in \mathbb{R}^d$, can be immediately obtained using (73) and the μ-strong convexity of $F(\boldsymbol{\theta})$:

$$\begin{aligned}
\boldsymbol{\Phi}_{k+1}(\boldsymbol{\theta}) - F(\boldsymbol{\theta}) &= \rho\boldsymbol{\Phi}_k(\boldsymbol{\theta}) - F(\boldsymbol{\theta}) \\
&\quad + \sqrt{\mu\gamma}\Big(F(\boldsymbol{\theta}^k_{ex}) + \langle \nabla F(\boldsymbol{\theta}^k_{ex}) + \boldsymbol{e}^k, \boldsymbol{\theta} - \boldsymbol{\theta}^k_{ex}\rangle + \frac{\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^k_{ex}\|^2\Big) \\
&\leq \rho\big(\boldsymbol{\Phi}_k(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\big) + \sqrt{\mu\gamma}\langle \boldsymbol{e}^k, \boldsymbol{\theta} - \boldsymbol{\theta}^s_{ex}\rangle \\
&\leq \rho^k\big(\boldsymbol{\Phi}_1(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\big) + \sum_{\ell=1}^{k} \rho^{k-\ell}\sqrt{\mu\gamma}\langle \boldsymbol{e}^\ell, \boldsymbol{\theta} - \boldsymbol{\theta}^\ell_{ex}\rangle \,.
\end{aligned} \tag{74}$$

To facilitate our development, let us denote:

$$\boldsymbol{\Phi}^\star_k := \min_{\boldsymbol{\theta}} \boldsymbol{\Phi}_k(\boldsymbol{\theta}), \quad \boldsymbol{v}^k := \arg\min_{\boldsymbol{\theta}} \boldsymbol{\Phi}_k(\boldsymbol{\theta}) \,. \tag{75}$$

By setting $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$ in (74), we have:

$$\begin{aligned}
\boldsymbol{\Phi}^\star_{k+1} - F(\boldsymbol{\theta}^\star) &\leq \boldsymbol{\Phi}_{k+1}(\boldsymbol{\theta}^\star) - F(\boldsymbol{\theta}^\star) \\
&\leq \rho^k\Big(\frac{\mu}{2}\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^1_{ex}\|^2 + F(\boldsymbol{\theta}^1_{ex}) - F(\boldsymbol{\theta}^\star)\Big) + \sum_{\ell=1}^{k} \rho^{k-\ell}\sqrt{\mu\gamma}\langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\star - \boldsymbol{\theta}^\ell_{ex}\rangle \\
&\leq 2\rho^k\big(F(\boldsymbol{\theta}^1) - F(\boldsymbol{\theta}^\star)\big) + \sum_{\ell=1}^{k} \rho^{k-\ell}\sqrt{\mu\gamma}\langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\star - \boldsymbol{\theta}^\ell_{ex}\rangle \,.
\end{aligned} \tag{76}$$

Now, if $F(\boldsymbol{\theta}^{k+1}) \leq \boldsymbol{\Phi}^\star_{k+1}$, then the inequality above shows the evolution of the optimality gap $h^{(k)}$. This motivates our next step, relating $F(\boldsymbol{\theta}^{k+1})$ to $\boldsymbol{\Phi}^\star_{k+1}$.

Lower bounding $\boldsymbol{\Phi}_{k+1}^{\star}$ in the presence of errors. Since $\nabla^2 \boldsymbol{\Phi}_k(\boldsymbol{\theta}) = \mu \boldsymbol{I}$, the function $\boldsymbol{\Phi}_k(\boldsymbol{\theta})$ is quadratic and we can represent $\boldsymbol{\Phi}_k(\boldsymbol{\theta})$ alternatively as

$$\boldsymbol{\Phi}_k(\boldsymbol{\theta}) = \boldsymbol{\Phi}_k^{\star} + \frac{\mu}{2} \| \boldsymbol{\theta} - \boldsymbol{v}^k \|^2 \ . \tag{77}$$

By substituting (77) into the definition of $\boldsymbol{\Phi}_{k+1}(\boldsymbol{\theta})$ in (73) and evaluating the first order optimality condition of the latter, we have:

$$\begin{aligned} &\sqrt{\mu\gamma}(\boldsymbol{g}^k + \mu(\boldsymbol{v}^{k+1} - \boldsymbol{\theta}_{ex}^k)) + \rho\ \mu(\boldsymbol{v}^{k+1} - \boldsymbol{v}^k) = \boldsymbol{0} \ , \\ &\Longrightarrow \boldsymbol{v}^{k+1} = \rho \boldsymbol{v}^k + \sqrt{\mu\gamma}\boldsymbol{\theta}_{ex}^k - \sqrt{\frac{\gamma}{\mu}} \boldsymbol{g}^k \ . \end{aligned} \tag{78}$$

By setting $\boldsymbol{\theta} = \boldsymbol{\theta}_{ex}^k$ in (73) and using the recursive definition of $\boldsymbol{\Phi}_{k+1}(\boldsymbol{\theta})$, we obtain

$$\boldsymbol{\Phi}_{k+1}(\boldsymbol{\theta}_{ex}^k) = \rho \boldsymbol{\Phi}_k(\boldsymbol{\theta}_{ex}^k) + \sqrt{\mu\gamma} F(\boldsymbol{\theta}_{ex}^k) = \rho\Big(\boldsymbol{\Phi}_k^{\star} + \frac{\mu}{2} \| \boldsymbol{\theta}_{ex}^k - \boldsymbol{v}^k \|^2 \Big) + \sqrt{\mu\gamma} F(\boldsymbol{\theta}_{ex}^k) \ , \tag{79}$$

while setting $\boldsymbol{\theta} = \boldsymbol{\theta}_{ex}^k$ in (77) and using (78) gives us:

$$\boldsymbol{\Phi}_{k+1}(\boldsymbol{\theta}_{ex}^k) = \boldsymbol{\Phi}_{k+1}^{\star} + \frac{\mu}{2} \left(\rho^2 \| \boldsymbol{\theta}_{ex}^k - \boldsymbol{v}^k \|^2 + \frac{\gamma}{\mu} \| \boldsymbol{g}^k \|^2 + 2\rho \sqrt{\frac{\gamma}{\mu}} \langle \boldsymbol{g}^k, \boldsymbol{\theta}_{ex}^k - \boldsymbol{v}^k \rangle \right) \ . \tag{80}$$

Comparing the right hand side of (79) and (80) shows:

$$\begin{aligned} \boldsymbol{\Phi}_{k+1}^{\star} &= \rho\Big(\boldsymbol{\Phi}_k^{\star} + \frac{\mu}{2} \| \boldsymbol{\theta}_{ex}^k - \boldsymbol{v}^k \|^2 \Big) + \sqrt{\mu\gamma} F(\boldsymbol{\theta}_{ex}^k) \\ &\quad - \frac{\mu}{2} \left(\rho^2 \| \boldsymbol{\theta}_{ex}^k - \boldsymbol{v}^k \|^2 + \frac{\gamma}{\mu} \| \boldsymbol{g}^k \|^2 + 2\rho \sqrt{\frac{\gamma}{\mu}} \langle \boldsymbol{g}^k, \boldsymbol{\theta}_{ex}^k - \boldsymbol{v}^k \rangle \right) \\ &= \rho \boldsymbol{\Phi}_k^{\star} + \sqrt{\mu\gamma} F(\boldsymbol{\theta}_{ex}^k) + \frac{\mu}{2} \rho \sqrt{\mu\gamma} \| \boldsymbol{\theta}_{ex}^k - \boldsymbol{v}^k \|^2 - \frac{\gamma}{2} \| \boldsymbol{g}^k \|^2 - \rho \sqrt{\mu\gamma} \langle \boldsymbol{g}^k, \boldsymbol{\theta}_{ex}^k - \boldsymbol{v}^k \rangle \ . \end{aligned}$$

Using the fact $\boldsymbol{v}^k - \boldsymbol{\theta}_{ex}^k = (\sqrt{\mu\gamma})^{-1} \big(\boldsymbol{\theta}_{ex}^k - \boldsymbol{\theta}^k \big)$ (proven in Sect. D.1), we have

$$\boldsymbol{\Phi}_{k+1}^{\star} = \rho \boldsymbol{\Phi}_k^{\star} + \sqrt{\mu\gamma} F(\boldsymbol{\theta}_{ex}^k) + \frac{\mu}{2} \frac{\rho}{\sqrt{\mu\gamma}} \| \boldsymbol{\theta}_{ex}^k - \boldsymbol{\theta}^k \|^2 - \frac{\gamma}{2} \| \boldsymbol{g}^k \|^2 - \rho \langle \boldsymbol{g}^k, \boldsymbol{\theta}^k - \boldsymbol{\theta}_{ex}^k \rangle \ . \tag{81}$$

We obtain the following chain:

$$
\begin{aligned}
F(\boldsymbol{\theta}^{k+1}) - \boldsymbol{\Phi}_{k+1}^\star &\overset{(a)}{\leq} F(\boldsymbol{\theta}_{ex}^k) - \gamma \langle \nabla F(\boldsymbol{\theta}_{ex}^k), \boldsymbol{g}^k \rangle + \frac{L\gamma^2}{2} \|\boldsymbol{g}^k\|^2 - \boldsymbol{\Phi}_{k+1}^\star \\
&\overset{(b)}{=} \rho \left(F(\boldsymbol{\theta}_{ex}^k) + \langle \boldsymbol{g}^k, \boldsymbol{\theta}^k - \boldsymbol{\theta}_{ex}^k \rangle - \boldsymbol{\Phi}_k^\star \right) \\
&\quad - \gamma \langle \nabla F(\boldsymbol{\theta}_{ex}^k), \boldsymbol{g}^k \rangle + \frac{\gamma}{2}(1 + L\gamma)\|\boldsymbol{g}^k\|^2 - \frac{\mu}{2} \frac{\rho}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^k - \boldsymbol{\theta}^k\|^2 \\
&\overset{(c)}{=} \rho \left(F(\boldsymbol{\theta}_{ex}^k) + \langle \nabla F(\boldsymbol{\theta}_{ex}^k), \boldsymbol{\theta}^k - \boldsymbol{\theta}_{ex}^k \rangle - \boldsymbol{\Phi}_k^\star \right) - \gamma \langle \nabla F(\boldsymbol{\theta}_{ex}^k), \boldsymbol{g}^k \rangle \\
&\quad + \rho \langle \boldsymbol{e}^k, \boldsymbol{\theta}^k - \boldsymbol{\theta}_{ex}^k \rangle + \frac{\gamma}{2}(1 + L\gamma)\|\boldsymbol{g}^k\|^2 - \frac{\mu}{2} \frac{\rho}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^k - \boldsymbol{\theta}^k\|^2 \\
&\overset{(d)}{\leq} \rho \left(F(\boldsymbol{\theta}^k) - \boldsymbol{\Phi}_k^\star + \langle \boldsymbol{e}^k, \boldsymbol{\theta}^k - \boldsymbol{\theta}_{ex}^k \rangle \right) - \frac{\mu}{2} \frac{1 - \mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^k - \boldsymbol{\theta}^k\|^2 \\
&\quad + \frac{\gamma}{2}(1 + L\gamma)\|\boldsymbol{g}^k\|^2 - \gamma \langle \nabla F(\boldsymbol{\theta}_{ex}^k), \boldsymbol{g}^k \rangle \\
&\overset{(e)}{\leq} \rho \left(F(\boldsymbol{\theta}^k) - \boldsymbol{\Phi}_k^\star + \langle \boldsymbol{e}^k, \boldsymbol{\theta}^k - \boldsymbol{\theta}_{ex}^k \rangle \right) - \frac{\mu}{2} \frac{1 - \mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^k - \boldsymbol{\theta}^k\|^2 + \gamma \|\boldsymbol{e}^k\|^2 ,
\end{aligned}
\tag{82}
$$

where (a) is due to the L-smoothness of F; (b) is due to (81); (c) is obtained by expanding $\boldsymbol{g}^k$ as $\nabla F(\boldsymbol{\theta}_{ex}^k) + \boldsymbol{e}^k$; (d) is obtained by adding and subtracting $(\mu/2)\|\boldsymbol{\theta}^k - \boldsymbol{\theta}_{ex}^k\|^2$ inside the first bracket, applying the identity $\rho + \rho/\sqrt{\mu\gamma} = (1 - \mu\gamma)/\sqrt{\mu\gamma}$, and using the μ-strong convexity of F; and (e) is due to the following chain of inequalities:

$$
\begin{aligned}
&\frac{\gamma}{2}(1 + L\gamma)\|\boldsymbol{g}^k\|^2 - \gamma \langle \nabla F(\boldsymbol{\theta}_{ex}^k), \boldsymbol{g}^s \rangle \\
&\quad \leq \frac{\gamma}{2}(1 + L\gamma)\big(\|\boldsymbol{e}^k\|^2 + \|\nabla F(\boldsymbol{\theta}_{ex}^k)\|^2\big) + \frac{L\gamma^2}{2}\big(\|\nabla F(\boldsymbol{\theta}_{ex}^k)\|^2 + \|\boldsymbol{e}^k\|^2\big) - \gamma \|\nabla F(\boldsymbol{\theta}_{ex}^k)\|^2 \\
&\quad = \left(\frac{\gamma}{2} + L\gamma^2\right)\|\boldsymbol{e}^k\|^2 + \left(-\frac{\gamma}{2} + L\gamma^2\right)\|\nabla F(\boldsymbol{\theta}_{ex}^k)\|^2 \leq \gamma \|\boldsymbol{e}^k\|^2 .
\end{aligned}
$$

As $\boldsymbol{\Phi}_1(\boldsymbol{\theta}^1) = F(\boldsymbol{\theta}^1) = \boldsymbol{\Phi}_1^\star$, applying the inequality (82) recursively shows:

$$
\begin{aligned}
&F(\boldsymbol{\theta}^{k+1}) - \boldsymbol{\Phi}_{k+1}^\star \leq \\
&\sum_{\ell=1}^{k} \rho^{k-\ell} \left((1 - \sqrt{\mu\gamma}) \langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\ell - \boldsymbol{\theta}_{ex}^\ell \rangle + \gamma \|\boldsymbol{e}^\ell\|^2 - \frac{\mu}{2} \frac{1 - \mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^\ell - \boldsymbol{\theta}^\ell\|^2 \right) .
\end{aligned}
\tag{83}
$$

Importantly, (83) establishes a lower bound on $\boldsymbol{\Phi}_{k+1}^\star$ in terms of $F(\boldsymbol{\theta}^{k+1})$ and $\boldsymbol{e}^k$.

Proving Proposition 2. Finally, summing up (83) and (76) gives:

$$
\begin{aligned}
h^{(k+1)} &\leq 2\rho^k h^{(1)} + \sum_{\ell=1}^{k} \rho^{k-\ell} \Big(\sqrt{\mu\gamma} \langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\star - \boldsymbol{\theta}_{ex}^\ell \rangle \\
&\qquad + \rho \langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\ell - \boldsymbol{\theta}_{ex}^\ell \rangle + \gamma \|\boldsymbol{e}^\ell\|^2 - \frac{\mu}{2} \frac{1-\mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^\ell - \boldsymbol{\theta}^\ell\|^2 \Big) \\
&= 2\rho^k h^{(1)} + \sum_{\ell=1}^{k} \rho^{k-\ell} \Big(\sqrt{\mu\gamma} \langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\star - \boldsymbol{\theta}^\ell \rangle \\
&\qquad + \langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\ell - \boldsymbol{\theta}_{ex}^\ell \rangle + \gamma \|\boldsymbol{e}^\ell\|^2 - \frac{\mu}{2} \frac{1-\mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^\ell - \boldsymbol{\theta}^\ell\|^2 \Big).
\end{aligned} \tag{84}
$$

Let us take a look at the last summands in the above inequality: for any $\ell \geq 1$,

$$
\begin{aligned}
&\sqrt{\mu\gamma} \langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\star - \boldsymbol{\theta}^\ell \rangle + \langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\ell - \boldsymbol{\theta}_{ex}^\ell \rangle + \gamma \|\boldsymbol{e}^\ell\|^2 - \frac{\mu}{2} \frac{1-\mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^\ell - \boldsymbol{\theta}^\ell\|^2 \\
&\overset{(a)}{\leq} \sqrt{\mu\gamma} \|\boldsymbol{e}^\ell\| \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}^\ell\| + \left(\gamma + \frac{\sqrt{\gamma/\mu}}{1-\mu\gamma} \right) \|\boldsymbol{e}^\ell\|^2 - \frac{\mu}{4} \frac{1-\mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^\ell - \boldsymbol{\theta}^\ell\|^2 \\
&\overset{(b)}{\leq} \sqrt{2\gamma h^{(\ell)}} \|\boldsymbol{e}^\ell\| + \left(\gamma + \frac{\sqrt{\gamma/\mu}}{1-\mu\gamma} \right) \|\boldsymbol{e}^\ell\|^2 - \frac{\mu}{4} \frac{1-\mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^\ell - \boldsymbol{\theta}^\ell\|^2 \\
&\overset{(c)}{\leq} \sqrt{2\gamma h^{(\ell)}} \|\boldsymbol{e}^\ell\| + \sqrt{\frac{9\gamma}{\mu}} \|\boldsymbol{e}^\ell\|^2 - \frac{\mu}{4} \frac{1-\mu\gamma}{\sqrt{\mu\gamma}} \|\boldsymbol{\theta}_{ex}^\ell - \boldsymbol{\theta}^\ell\|^2 ,
\end{aligned} \tag{85}
$$

where (a) is resulted from the fact $\langle \boldsymbol{e}^\ell, \boldsymbol{\theta}^\ell - \boldsymbol{\theta}_{ex}^\ell \rangle \leq (1/2)(\|\boldsymbol{e}^\ell\|^2/c + c\|\boldsymbol{\theta}^\ell - \boldsymbol{\theta}_{ex}^\ell\|^2)$ for any $c > 0$ and we have set $c = \frac{\mu}{2} \frac{1-\mu\gamma}{\sqrt{\mu\gamma}}$ therein; (b) is due to the relation $\|\boldsymbol{\theta}^\ell - \boldsymbol{\theta}^\star\| \leq \sqrt{2h^{(\ell)}/\mu}$; (c) is due to $\gamma + \frac{\sqrt{\gamma/\mu}}{1-\mu\gamma} \leq 3\sqrt{\gamma/\mu}$, which can be verified through replacing γ by its upper bound $1/(2L)$ in the denominator of the fraction on the left-hand-side. Combining the two equations above yields the desired result of Proposition.

D.1 Proof of the equality

We prove $\boldsymbol{v}^k - \boldsymbol{\theta}_{ex}^k = (\sqrt{\mu\gamma})^{-1}\big(\boldsymbol{\theta}_{ex}^k - \boldsymbol{\theta}^k\big)$ using induction on k. Clearly, the said equality holds for $k = 1$ since $\boldsymbol{v}^1 = \boldsymbol{\theta}^1 = \boldsymbol{\theta}_{ex}^1$, and we assume that it holds up to k. Consider:

$$
\begin{aligned}
\boldsymbol{v}^{k+1} - \boldsymbol{\theta}_{ex}^{k+1} &= \rho \boldsymbol{v}^k + \sqrt{\mu\gamma} \boldsymbol{\theta}_{ex}^k - \sqrt{\frac{\gamma}{\mu}} \boldsymbol{g}^k - \boldsymbol{\theta}_{ex}^{k+1} \\
&= \rho(\boldsymbol{v}^k - \boldsymbol{\theta}_{ex}^k) + \boldsymbol{\theta}_{ex}^k - \sqrt{\frac{\gamma}{\mu}} \boldsymbol{g}^k - \boldsymbol{\theta}_{ex}^{k+1} = \frac{\rho}{\sqrt{\mu\gamma}} (\boldsymbol{\theta}_{ex}^k - \boldsymbol{\theta}^k) + \boldsymbol{\theta}_{ex}^k - \sqrt{\frac{\gamma}{\mu}} \boldsymbol{g}^k - \boldsymbol{\theta}_{ex}^{k+1} ,
\end{aligned}
$$

where we have used the induction hypothesis. Furthermore, using $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}_{ex}^k - \gamma \boldsymbol{g}^k$,

$$\begin{aligned} \boldsymbol{v}^{k+1} - \boldsymbol{\theta}_{ex}^{k+1} &= \sqrt{\mu\gamma}^{-1}\Big(\rho(\boldsymbol{\theta}_{ex}^{k} - \boldsymbol{\theta}^{k}) + \sqrt{\mu\gamma}(\boldsymbol{\theta}_{ex}^{k} - \boldsymbol{\theta}_{ex}^{k+1}) - \gamma \boldsymbol{g}^{k}\Big) \\ &\overset{(a)}{=} \sqrt{\mu\gamma}^{-1}\Big(\sqrt{\mu\gamma}(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}_{ex}^{k+1}) + \rho(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k})\Big) = \sqrt{\mu\gamma}^{-1}\big(\boldsymbol{\theta}_{ex}^{k+1} - \boldsymbol{\theta}^{k+1}\big)\,, \end{aligned} \tag{86}$$

where (a) is due to $\rho(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^{k}) = (1 + \sqrt{\mu\gamma})(\boldsymbol{\theta}_{ex}^{k+1} - \boldsymbol{\theta}^{k+1})$.

E Proof of Proposition 4

We begin by observing that due to the $L_{H,i}$-Lipschitz continuity of the Hessian of f_i and using Lemma 1, we have:

$$\|\boldsymbol{e}_{\mathsf{ACIAG}}^{\ell}\| = \|\boldsymbol{g}_{\mathsf{ACIAG}}^{\ell} - \nabla F(\boldsymbol{\theta}_{ex}^{\ell})\| \leq \sum_{i=1}^{m} \frac{L_{H,i}}{2} \|\boldsymbol{\theta}_{ex}^{\ell} - \boldsymbol{\theta}_{ex}^{\tau_i^{\ell}}\|^2\,. \tag{87}$$

Now, expanding the right hand side of (87) gives:

$$\begin{aligned} \|\boldsymbol{e}_{\mathsf{ACIAG}}^{\ell}\| &\leq \sum_{i=1}^{m} \frac{L_{H,i}}{2} \Big\|\boldsymbol{\theta}_{ex}^{\ell} - \boldsymbol{\theta}_{ex}^{\tau_i^{\ell}}\Big\|^2 \leq \sum_{i=1}^{m} \frac{L_{H,i}}{2} \underbrace{(\ell - \tau_i^{\ell})}_{\leq K} \sum_{j=\ell-\tau_i^{\ell}}^{\ell-1} \|\boldsymbol{\theta}_{ex}^{j+1} - \boldsymbol{\theta}_{ex}^{j}\|^2 \\ &\leq \frac{KL_H}{2} \sum_{j=(\ell-K)_{++}}^{\ell-1} \|\boldsymbol{\theta}_{ex}^{j+1} - \boldsymbol{\theta}_{ex}^{j}\|^2 = \frac{KL_H}{2} \sum_{j=(\ell-K)_{++}}^{\ell-1} \|\gamma \boldsymbol{g}_{\mathsf{ACIAG}}^{j} + \underbrace{\alpha(\boldsymbol{\theta}^{j+1} - \boldsymbol{\theta}^{j})}_{=\boldsymbol{\theta}_{ex}^{j+1} - \boldsymbol{\theta}^{j+1}}\|^2 \\ &\leq \frac{3KL_H}{2} \sum_{j=(\ell-K)_{++}}^{\ell-1} \big(\gamma^2\big(\|\boldsymbol{e}^{j}\|^2 + \|\nabla F(\boldsymbol{\theta}_{ex}^{j})\|^2\big) + \|\boldsymbol{\theta}_{ex}^{j+1} - \boldsymbol{\theta}^{j+1}\|^2\big)\,. \end{aligned} \tag{88}$$

Remarkably, the above bound resembles that of Proposition 3 with the exception of the last term that depends on $\boldsymbol{\theta}_{ex}^{j+1} - \boldsymbol{\theta}^{j+1}$. This is included to account for the extrapolated iterates used in the A-CIAG method.

To find an upper bound of $\|\boldsymbol{e}_{\mathsf{ACIAG}}^{\ell}\|$ to corroborate Proposition 4, in what follows, we will upper bound $\|\boldsymbol{e}_{\mathsf{ACIAG}}^{j}\|^2$ and $\|\nabla F(\boldsymbol{\theta}_{ex}^{j})\|^2$, respectively. Firstly,

$$\begin{aligned} \|\boldsymbol{e}_{\mathsf{ACIAG}}^{j}\| &\leq \sum_{i=1}^{m} \frac{L_{H,i}}{2} \Big\|\boldsymbol{\theta}_{ex}^{j} - \boldsymbol{\theta}_{ex}^{\tau_i^{j}}\Big\|^2 \\ &\leq \sum_{i=1}^{m} L_{H,i}\Big((1+\alpha)^2 \|\boldsymbol{\theta}^{j} - \boldsymbol{\theta}^{\tau_i^{j}}\|^2 + \alpha^2 \|\boldsymbol{\theta}^{j-1} - \boldsymbol{\theta}^{\tau_i^{j}-1}\|^2\Big)\,. \end{aligned} \tag{89}$$

Noticing that as $\|\boldsymbol{\theta}^{j} - \boldsymbol{\theta}^{\tau_i^{j}}\|^2 \leq 2(\|\boldsymbol{\theta}^{j} - \boldsymbol{\theta}^{\star}\|^2 + \|\boldsymbol{\theta}^{\tau_i^{j}} - \boldsymbol{\theta}^{\star}\|^2) \leq (4/\mu)(h^{(j)} + h^{(\tau_i^{j})})$, it follows from (89) that

$$\|\boldsymbol{e}^j_{\mathsf{ACIAG}}\| \le \frac{4}{\mu}\sum_{i=1}^m L_{H,i}\Big((1+\alpha)^2(h^{(j)}+h^{(\tau_i^j)})+\alpha^2(h^{(j-1)}+h^{(\tau_i^j-1)})\Big)$$
$$\le \frac{8L_H}{\mu}\big((1+\alpha)^2+\alpha^2\big)\max_{(j-K-1)_{++}\le q\le j} h^{(q)} \le \frac{40L_H}{\mu}\max_{(j-K-1)_{++}\le q\le j} h^{(q)}\,, \tag{90}$$

which implies

$$\sum_{j=(\ell-K)_{++}}^{\ell-1}\|\boldsymbol{e}^j_{\mathsf{ACIAG}}\|^2 \le K\left(\frac{40L_H}{\mu}\right)^2 \max_{(\ell-2K-1)_{++}\le q\le \ell}(h^{(q)})^2\,. \tag{91}$$

Secondly,

$$\|\nabla F(\boldsymbol{\theta}^j_{ex})\|^2 \le 2L^2\big(\|\boldsymbol{\theta}^j-\boldsymbol{\theta}^\star\|^2+\|\boldsymbol{\theta}^j-\boldsymbol{\theta}^{j-1}\|^2\big) \le \frac{4L^2}{\mu}\big(3h^{(j)}+2h^{(j-1)})\big)\,, \tag{92}$$

thus

$$\sum_{j=(\ell-K)_{++}}^{\ell-1}\|\nabla F(\boldsymbol{\theta}^j_{ex})\|^2 \le \frac{20L^2K}{\mu}\max_{(\ell-K-1)_{++}\le q\le \ell-1} h^{(q)}\,. \tag{93}$$

Substituting (91) and (93) into the right hand side of (88) verifies Proposition 4.

F Step 3 in the Proof of Theorem 2

To proceed with the proof, let us define the following quantity:

$$\tilde{E}^{(\ell)} := \gamma^{\frac{5}{2}}\sqrt{\frac{9}{2}}K^2L_H\left(\left(\frac{40L_H}{\mu}\right)^2\max_{(\ell-2K-1)_{++}\le q\le\ell}(h^{(q)})^2+\frac{20L^2}{\mu}\max_{(\ell-K-1)_{++}\le q\le\ell}h^{(q)}\right)$$
$$+\gamma^{\frac{9}{2}}\frac{81K^4L_H^2}{4\sqrt{\mu}}\left(\left(\frac{40L_H}{\mu}\right)^4\max_{(\ell-2K-1)_{++}\le q\le\ell}(h^{(q)})^4+\left(\frac{20L^2}{\mu}\right)^2\max_{(\ell-K-1)_{++}\le q\le\ell}(h^{(q)})^2\right).$$

Using Proposition 4, we obtain:

$$\sqrt{2\gamma h^{(\ell)}}\|\boldsymbol{e}^\ell_{\mathsf{ACIAG}}\| + \sqrt{\frac{9\gamma}{\mu}}\|\boldsymbol{e}^\ell_{\mathsf{ACIAG}}\|^2$$
$$\le \tilde{E}^{(\ell)} + \sum_{j=(\ell-K+1)_{++}}^{\ell}\left(\sqrt{\frac{9\gamma h^{(\ell)}K^2L_H^2}{2}}\|\boldsymbol{\theta}^j-\boldsymbol{\theta}^j_{ex}\|^2+\frac{27K^3L_H^2}{4}\sqrt{\frac{9\gamma}{\mu}}\|\boldsymbol{\theta}^j-\boldsymbol{\theta}^j_{ex}\|^4\right). \tag{94}$$

We need to further bound $h^{(k)}$ [recall for (41) in Proposition 2] in terms of itself to create a ‘recursion’ for $h^{(k)}$. To upper bound the right hand side of (41), let us start from (94). It follows that

$$\sum_{\ell=1}^{k} \rho^{k-\ell}\left(\sqrt{2\gamma h^{(\ell)}}\|\boldsymbol{e}^{\ell}\| + \sqrt{\frac{9\gamma}{\mu}}\|\boldsymbol{e}^{\ell}\|^2 - \frac{\mu}{4}\frac{1-\mu\gamma}{\sqrt{\mu\gamma}}\|\boldsymbol{\theta}_{ex}^{\ell} - \boldsymbol{\theta}^{\ell}\|^2\right) \leq \sum_{\ell=1}^{k} \rho^{k-\ell}\left(\tilde{E}^{(\ell)}\right.$$

$$\left. + \left(\sum_{j=\ell}^{\min\{k,\ell+K-1\}}\left(\sqrt{\frac{9\gamma K^2 L_H^2 h^{(j)}}{2}} + \frac{81K^3L_H^2}{4}\sqrt{\frac{\gamma}{\mu}}\|\boldsymbol{\theta}^{\ell} - \boldsymbol{\theta}_{ex}^{\ell}\|^2\right) - \frac{\mu}{4}\frac{1-\mu\gamma}{\sqrt{\mu\gamma}}\right)\|\boldsymbol{\theta}^{\ell} - \boldsymbol{\theta}_{ex}^{\ell}\|^2\right).$$

Moreover, we observe for $\ell \geq 2$:

$$\|\boldsymbol{\theta}^{\ell} - \boldsymbol{\theta}_{ex}^{\ell}\|^2 \leq 2(\|\boldsymbol{\theta}^{\ell} - \boldsymbol{\theta}^{\star}\|^2 + \|\boldsymbol{\theta}^{\ell-1} - \boldsymbol{\theta}^{\star}\|^2) \leq \frac{4}{\mu}\left(h^{(\ell)} + h^{(\ell-1)}\right), \tag{95}$$

The coefficient in front of the last $\|\boldsymbol{\theta}^{\ell} - \boldsymbol{\theta}_{ex}^{\ell}\|^2$ term can be upper bounded as:

$$\tilde{C}^{(\ell,k)} := \gamma K^2 L_H \sqrt{\frac{9}{2}} \max_{\ell \leq q \leq \min\{\ell+K-1,k\}} (h^{(q)})^{\frac{1}{2}} + \gamma \frac{81K^4L_H^2}{\mu^{\frac{3}{2}}}\left(h^{(\ell)} + h^{(\ell-1)}\right) - \frac{\mu}{4}\frac{1-\mu\gamma}{\sqrt{\mu}}.$$

If we define

$$E^{(\ell,k)} := \tilde{E}^{(\ell)} + \tilde{C}^{(\ell,k)}\frac{\|\boldsymbol{\theta}^{\ell} - \boldsymbol{\theta}_{ex}^{\ell}\|^2}{\sqrt{\gamma}}, \tag{96}$$

where $E^{(\ell,k)} = E^{(\ell,k-1)}$ for all $k \geq \ell + m$. Applying Proposition 2 readily shows

$$h^{(k+1)} \leq 2(1-\sqrt{\mu\gamma})^k h^{(1)} + \sum_{\ell=1}^{k}(1-\sqrt{\mu\gamma})^{k-\ell}E^{(\ell,k)}. \tag{97}$$

Concluding the Proof of Theorem 2. Our goal is to analyze (97) using Proposition 6. Let us recognize that:

$$\begin{aligned}
R^{(k)} =& \bar{h}^{(k)},\ p = (1-\sqrt{\mu\gamma}),\ b = 2,\ M = 2K+1,\ \eta_1 = \frac{3}{2},\ \eta_2 = \frac{5}{2}, \eta_3 = 2,\ \eta_4 = 4\\
s_1 =& \gamma^{\frac{5}{2}}\sqrt{\frac{9}{2}}K^2L_H\frac{20L^2}{\mu},\ s_2 = \gamma^{\frac{5}{2}}\sqrt{\frac{9}{2}}K^2L_H\left(\frac{40L_H}{\mu}\right)^2,\\
s_3 =& \gamma^{\frac{9}{2}}\frac{81K^4L_H^2}{4\sqrt{\mu}}\left(\frac{20L^2}{\mu}\right)^2,\ s_4 = \gamma^{\frac{9}{2}}\frac{81K^4L_H^2}{4\sqrt{\mu}}\left(\frac{40L_H}{\mu}\right)^4,\\
c =& \frac{\mu}{4}\frac{1-\mu\gamma}{\sqrt{\mu}},\ D^{(\ell)} = \frac{\|\boldsymbol{\theta}^{\ell} - \boldsymbol{\theta}_{ex}^{\ell}\|^2}{\sqrt{\gamma}},\ f(\bar{h}^{(q)}) = \gamma\left(K^2L_H\sqrt{\frac{9}{2}}(\bar{h}^{(q)})^{\frac{1}{2}} + \frac{162K^4L_H^2}{\mu^{\frac{3}{2}}}\bar{h}^{(q)}\right).
\end{aligned}$$

The conditions in (55) are satisfied when

$$
\begin{aligned}
&\frac{\sqrt{\mu}}{4} - \gamma\left(K^2 L_H \sqrt{9} (\bar{h}^{(1)})^{\frac{1}{2}} + \frac{324 K^4 L_H^2}{\mu^{\frac{3}{2}}} \bar{h}^{(1)} + \frac{\mu^{\frac{3}{2}}}{4} \right) \geq 0 \\
&\iff \gamma \leq \frac{\sqrt{\mu}}{4} \left(K^2 L_H \sqrt{9} (\bar{h}^{(1)})^{\frac{1}{2}} + \frac{324 K^4 L_H^2}{\mu^{\frac{3}{2}}} \bar{h}^{(1)} + \frac{\mu^{\frac{3}{2}}}{4} \right)^{-1} := \frac{\bar{c}_3}{L} ,
\end{aligned}
\tag{98}
$$

and

$$
\begin{aligned}
1 > (1 - \sqrt{\mu\gamma}) + \gamma^{\frac{5}{2}} \sqrt{\frac{9}{2}} K^2 L_H \left(\frac{20 L^2}{\mu} (2\bar{h}^{(1)})^{\frac{1}{2}} + \left(\frac{40 L_H}{\mu} \right)^2 (2\bar{h}^{(1)})^{\frac{3}{2}} \right) \\
+ \gamma^{\frac{9}{2}} \frac{81 K^4 L_H^2}{4\sqrt{\mu}} \left(\left(\frac{20 L^2}{\mu} \right)^2 (2\bar{h}^{(1)}) + \left(\frac{40 L_H}{\mu} \right)^4 (2\bar{h}^{(1)})^3 \right),
\end{aligned}
\tag{99}
$$

that can be implied by

$$
\begin{aligned}
&\gamma < \left(\frac{\sqrt{\mu}}{\sqrt{18} K^2 L_H} \left(\frac{20 L^2}{\mu} (2\bar{h}^{(1)})^{\frac{1}{2}} + \left(\frac{40 L_H}{\mu} \right)^2 (2\bar{h}^{(1)})^{\frac{3}{2}} \right)^{-1} \right)^{\frac{1}{2}} := \frac{\bar{c}_1}{L} \quad \text{and} \\
&\gamma < \left(\frac{2\mu}{81 K^4 L_H^2} \left(\left(\frac{20 L^2}{\mu} \right)^2 (2\bar{h}^{(1)}) + \left(\frac{40 L_H}{\mu} \right)^4 (2\bar{h}^{(1)})^3 \right)^{-1} \right)^{\frac{1}{4}} := \frac{\bar{c}_2}{L} .
\end{aligned}
\tag{100}
$$

Substituting these constants into Proposition 6 proves the claims in Theorem 2.

G Proof of Proposition 6

Define $\{\bar{R}^{(k)}\}_{k\geq 1}$ that satisfies:

$$
\bar{R}^{(k+1)} = p^k b \bar{R}^{(1)} + \sum_{\ell=1}^{k} p^{k-\ell} \left(\sum_{j=1}^{J} s_j \max_{(\ell-M)_{++} \leq q \leq \ell} (\bar{R}^{(q)})^{\eta_j} \right), \quad \bar{R}^{(1)} = R^{(1)} , \tag{101}
$$

By subtracting $p\bar{R}^{(k)}$ from $\bar{R}^{(k+1)}$, (101) can be alternatively expressed as:

$$
\bar{R}^{(k+1)} - p\bar{R}^{(k)} = \sum_{j=1}^{J} s_j \max_{(k-M)_{++} \leq q \leq k} (\bar{R}^{(q)})^{\eta_j} . \tag{102}
$$

Now, consider the statements (1) and (2) in (56) as the following event:

$$
\mathcal{E}_z = \left\{ \bar{R}^{((z-1)M+k+1)} \geq R^{((z-1)M+k+1)},\ \bar{R}^{((z-1)M+k+1)} \leq \delta^z (b\bar{R}^{(1)}),\ k = 1, ..., M \right\} ,
$$

for all $z \geq 1$. We shall prove that $\mathcal{E}_z$ is true for $z = 1, 2, ...$ using induction.

Base case with $z = 1$. To prove $\mathcal{E}_1$, let us apply another induction on k inside the event. For the base case of $k = 1$,

$$\bar{R}^{(2)} \geq p(bR^{(1)}) + \sum_{j=1}^{J} s_j (R^{(1)})^{\eta_j} - (\bar{f} - f(R^{(1)}))D^{(1)} = R^{(2)} \,, \tag{103}$$

where we used the fact $\bar{f} \geq f(bR^{(1)}) \geq f(R^{(1)})$. Furthermore, the base case holds as:

$$\bar{R}^{(2)} = (b\bar{R}^{(1)})\left(p + (1/b)\sum_{j=1}^{J} s_j (\bar{R}^{(1)})^{\eta_j - 1}\right) \leq \delta(b\bar{R}^{(1)}) \,. \tag{104}$$

For the induction step, suppose that the statements in (103) are also true up to $k = k' - 1$ with $z = 1$ such that $\bar{R}^{(k')} \geq R^{(k')}$ and $\bar{R}^{(k')} \leq \delta(b\bar{R}^{(1)})$. Consider the case of $k = k'$, we observe that $\bar{f} \geq f(bR^{(1)}) \geq f(\delta bR^{(1)}) \geq f(\bar{R}^{(q)}) \geq f(R^{(q)})$ for all $q = 1, ..., k'$. Therefore, we can lower bound $\bar{R}^{(k'+1)}$ as:

$$\begin{aligned}
\bar{R}^{(k'+1)} &= p^{k'}(b\bar{R}^{(1)}) + \sum_{\ell=1}^{k'} p^{k'-\ell}\left(\sum_{j=1}^{J} s_j \max_{(\ell-M)_{++} \leq q \leq \ell} (\bar{R}^{(q)})^{\eta_j}\right) \\
&\geq p^{k'}(bR^{(1)}) + \sum_{\ell=1}^{k'} p^{k'-\ell}\left(\sum_{j=1}^{J} s_j \max_{(\ell-M)_{++} \leq q \leq \ell} (R^{(q)})^{\eta_j} - \left(\bar{f} - \max_{\ell \leq q \leq k'} f(R^{(q)})\right) V^{(\ell)}\right),
\end{aligned}$$

where the right hand side is exactly $R^{(k'+1)}$; also, using (102), we can show:

$$\bar{R}^{(k'+1)} \leq (b\bar{R}^{(1)})\left(\delta p + \sum_{j=1}^{J} s_j (b\bar{R}^{(1)})^{\eta_j - 1}\right) \leq \delta(b\bar{R}^{(1)}) \,. \tag{105}$$

Induction Case. For the induction case, suppose that $\mathcal{E}_z$ is true for all z up to z'. We consider the case when $z = z' + 1$. Once again, we apply another induction on k. In the base case of $k = 1$ and $z = z' + 1$, we have

$$\begin{aligned}
\bar{R}^{(z'M+2)} &= p^{z'M+1}(b\bar{R}^{(1)}) + \sum_{\ell=1}^{z'M+1} p^{z'M+1-\ell}\left(\sum_{j=1}^{J} s_j \max_{(\ell-M)_{++} \leq q \leq \ell} (\bar{R}^{(q)})^{\eta_j}\right) \\
&\geq p^{z'M+1}(bR^{(1)}) + \sum_{\ell=1}^{z'M+1} p^{z'M+1-\ell}\left(\sum_{j=1}^{J} s_j \max_{(\ell-M)_{++} \leq q \leq \ell} (R^{(q)})^{\eta_j}\right. \\
&\quad \left. - \left(\bar{f} - \max_{\ell \leq q \leq z'M+1} f(R^{(q)})\right) V^{(\ell)}\right) = R^{(z'M+2)} \,,
\end{aligned}$$

where we used $\bar{f} \geq f(bR^{(1)}) \geq f(\bar{R}^{(q)}) \geq f(R^{(q)})$ for all q up to $q = z'M + 1$ (by the induction hypothesis). Furthermore, the base case holds since:

$$\begin{aligned}\bar{R}^{(z'M+2)} &= p\bar{R}^{(z'M+1)} + \sum_{j=1}^{J} s_j \max_{(z'M+1-M)_{++} \le q \le z'M+1} (\bar{R}^{(q)})^{\eta_j} \\ &\le \delta^{z'}(b\bar{R}^{(1)}) \left(p + \sum_{j=1}^{J} s_j (\delta^{z'})^{\eta_j-1} (b\bar{R}^{(1)})^{\eta_j-1} \right) \le \delta^{z'+1}(b\bar{R}^{(1)}) \ .\end{aligned} \tag{106}$$

Let the statements in $\mathcal{E}_z$ be true up to $k = k' - 1$, $z = z' + 1$. With $k = k'$,

$$\begin{aligned}\bar{R}^{(z'M+k'+1)} &\ge p^{z'M+k'}(bR^{(1)}) + \sum_{\ell=1}^{z'M+k'} p^{z'M+k'-\ell} \left(\sum_{j=1}^{J} s_j \max_{(\ell-M)_{++} \le q \le \ell} (R^{(q)})^{\eta_j} \right. \\ &\qquad \left. - \left(\bar{f} - \max_{\ell \le q \le z'M+k'} f(R^{(q)}) \right) V^{(\ell)} \right) = R^{(z'M+k'+1)} \ , \\ \bar{R}^{(z'M+k'+1)} &\le \delta^{z'}(b\bar{R}^{(1)}) \left(\delta p + \sum_{j=1}^{J} s_j (\delta^{z'})^{\eta_j-1} (b\bar{R}^{(1)})^{\eta_j-1} \right) \le \delta^{z'+1}(b\bar{R}^{(1)}) \ .\end{aligned} \tag{107}$$

The induction case is thus proven. This shows that the event $\mathcal{E}_z$ is true for all $z \ge 1$.

Proving statement (iii). We apply statement (ii) to prove (iii). From (102),

$$\frac{\bar{R}^{(k+1)}}{\bar{R}^{(k)}} = p + \frac{1}{\bar{R}^{(k)}} \sum_{j=1}^{J} s_j \max_{(k-M)_{++} \le q \le k} (\bar{R}^{(q)})^{\eta_j} \ . \tag{108}$$

For any $q \in [(k-M)_{++}, k]$, we have

$$\frac{(\bar{R}^{(q)})^{\eta_j}}{\bar{R}^{(k)}} = \frac{\bar{R}^{(q)}}{\bar{R}^{(k)}} (\bar{R}^{(q)})^{\eta_j-1} \le \frac{\bar{R}^{(q)}}{\bar{R}^{(k)}} \left(\delta^{\lceil (q-1)/M \rceil} (bR^{(1)}) \right)^{\eta_j-1} \ . \tag{109}$$

Since $\eta_j > 1$ and $|q-k| \le M$, we have $\delta^{\lceil (q-1)/M \rceil (\eta_j - 1)} \to 0$ as $k \to \infty$, moreover as $\bar{R}^{(k+1)}/\bar{R}^{(k)} \ge p$ for all $k \ge 1$, $\bar{R}^{(q)}/\bar{R}^{(k)} \le p^{-M}$ for all q. Therefore, we get

$$\lim_{k\to\infty} \frac{\max_{(k-M)_{++} \le q \le k} (\bar{R}^{(q)})^{\eta_j}}{\bar{R}^{(k)}} = 0, \ \forall j \Longrightarrow \lim_{k\to\infty} \frac{\bar{R}^{(k+1)}}{\bar{R}^{(k)}} = p \ . \tag{110}$$

References

1. Agarwal, A., Bottou, L.: A lower bound for the optimization of finite sums. In: International Conference on Machine Learning, pp. 78–86 (2015)
2. Arjevani, Y., Shamir, O.: Dimension-free iteration complexity of finite sum optimization problems. In: Advances in Neural InformationProcessing Systems 29, pp. 3540–3548 (2016)
3. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont (1999)
4. Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. Optim. Mach. Learn. **2010**(1–38), 3 (2011)
5. Blatt, D., Hero, A.O., Gauchman, H.: A convergent incremental gradient method with a constant step size. SIAM J. Optim. **18**(1), 29–51 (2007)
6. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)

7. Bubeck, S., et al.: Convex optimization: algorithms and complexity. Found. Trends Mach. Learn. **8**(3–4), 231–357 (2015)
8. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011)
9. Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: a fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in Neural Information Processing Systems 27, pp. 1646–1654 (2014)
10. Feyzmahdavian, H.R., Aytekin, A., Johansson, M.: A delayed proximal gradient method with linear convergence rate. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE (2014)
11. Gower, R.M., Roux, N.L., Bach, F.: Tracking the gradients using the hessian: a new look at variance reducing stochastic methods. In: AISTATS (2018)
12. Gürbüzbalaban, M., Ozdaglar, A., Parrilo, P.: A globally convergent incremental newton method. Math. Program. **151**(1), 283–313 (2015)
13. Gürbüzbalaban, M., Ozdaglar, A., Parrilo, P.: Why random reshuffling beats stochastic gradient descent. Math. Program. https://doi.org/10.1007/s10107-019-01440-w (2019)
14. Gürbüzbalaban, M., Ozdaglar, A., Parrilo, P.: On the convergence rate of incremental aggregated gradient algorithms. SIAM J. Optim. **27**(2), 1035–1048 (2017)
15. Lan, G., Zhou, Y.: An optimal randomized incremental gradient method. Math. Program. **171**, 167–215 (2018)
16. Mairal, J.: Incremental majorization-minimization optimization with application to large-scale machine learning. SIAM J. Optim. **25**(2), 829–855 (2015)
17. Mokhtari, A., Eisen, M., Ribeiro, A.: Iqn: an incremental quasi-newton method with local superlinear convergence rate. SIAM J. Optim. **28**(2), 1670–1698 (2018)
18. Nedić A., Bertsekas D.: Convergence Rate of Incremental Subgradient Algorithms. In: Uryasev S., Pardalos P.M. (eds) Stochastic Optimization: Algorithms and Applications. Applied Optimization, vol. 54. Springer, Boston, MA (2001)
19. Nedic, A., Bertsekas, D.P.: Incremental subgradient methods for nondifferentiable optimization. SIAM J. Optim. **12**(1), 109–138 (2001)
20. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, Berlin (2013)
21. Nitanda, A.: Stochastic proximal gradient descent with acceleration techniques. In: Advances in Neural Information Processing Systems, pp. 1574–1582 (2014)
22. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**(3), 400–407 (1951)
23. Rodomanov, A., Kropotov, D.: A superlinearly-convergent proximal Newton-type method for the optimization of finite sums. In: International Conference on Machine Learning, pp. 2597–2605 (2016)
24. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**(1–2), 83–112 (2017)
25. Schmidt, M., Roux, N.L., Bach, F.R.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems 24, pp. 1458–1466 (2011)
26. So, A.M.C., Zhou, Z.: Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity. Optim. Methods Softw. **32**(4), 963–992 (2017)
27. Vanli, N.D., Gürbüzbalaban, M., Ozdaglar, A.: A stronger convergence result on the proximal incremental aggregated gradient method. arXiv preprint arXiv:1611.08022 (2016)
28. Vapnik, V.N.: An overview of statistical learning theory. IEEE Trans. Neural Netw. **10**(5), 988–999 (1999)
29. Wai, H.T., Shi, W., Nedić, A., Scaglione, A.: Curvature-aided incremental aggregated gradient method. In: Proceedings of Allerton (2017)
30. Xiao, L., Zhang, T.: A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. **24**(4), 2057–2075 (2014)
31. Zheng, S., Meng, Q., Wang, T., Chen, W., Yu, N., Ma, Z.M., Liu, T.Y.: Asynchronous stochastic gradient descent with delay compensation. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 4120–4129. JMLR.org (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Hoi-To Wai[1] · Wei Shi[2] · César A. Uribe[4] · Angelia Nedić[3] · Anna Scaglione[3]

César A. Uribe
cauribe@mit.edu

Angelia Nedić
Angelia.Nedich@asu.edu

Anna Scaglione
Anna.Scaglione@asu.edu

1 Department of SEEM, Chinese University of Hong Kong, Sha Tin, Hong Kong

2 Department of EE, Princeton University, NJ, USA

3 School of ECEE, Arizona State University, Tempe, AZ, USA

4 LIDS, Massachusetts Institute of Technology, Cambridge, MA, USA