On the Convergence Rate of Incremental Aggregated Gradient Algorithms

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1137/15M1049695">http://dx.doi.org/10.1137/15M1049695</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Society for Industrial &amp; Applied Mathematics (SIAM)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Mon Apr 01 00:37:37 EDT 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/114181">http://hdl.handle.net/1721.1/114181</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
ON THE CONVERGENCE RATE OF INCREMENTAL AGGREGATED GRADIENT ALGORITHMS

M. GÜRBÜZBALABAN†, A. OZDAGLAR†, AND P. A. PARRILO†

Abstract. Motivated by applications to distributed optimization over networks and large-scale data processing in machine learning, we analyze the deterministic incremental aggregated gradient method for minimizing a finite sum of smooth functions where the sum is strongly convex. This method processes the functions one at a time in a deterministic order and incorporates a memory of previous gradient values to accelerate convergence. Empirically it performs well in practice; however, no theoretical analysis with explicit rate results was previously given in the literature to our knowledge, in particular most of the recent efforts concentrated on the randomized versions. In this paper, we show that this deterministic algorithm has global linear convergence and we characterize the convergence rate. We also consider an aggregated method with momentum and demonstrate its linear convergence. Our proofs rely on a careful choice of a Lyapunov function that offers insight into the algorithm’s behavior and simplifies the proofs considerably.

Key words. convex optimization, first-order methods, convergence analysis, large-scale optimization

AMS subject classifications. 90C30, 90C06, 90C25

DOI. 10.1137/15M1049695

1. Introduction. We consider the following unconstrained optimization problem where the objective function is the sum of component functions:

\begin{equation}
\min_{x \in \mathbb{R}^n} f(x) = \sum_{i=1}^{m} f_i(x),
\end{equation}

where each \( f_i : \mathbb{R}^n \to \mathbb{R} \) is a convex, continuously differentiable function referred to as a component function. This problem arises in many applications including least squares problems [2, 18] or more general parameter estimation problems where \( f_i \) is the corresponding loss function of the \( i \)th data block [12], distributed optimization in wireless sensor networks [6], machine learning problems [7, 24], and minimization of expected value of a function (where the expectation is taken over a finite probability distribution or approximated by an \( m \)-sample average).

One widely studied approach is the (deterministic) incremental gradient (IG) method, which cycles through the component functions using a cyclic order and updates the iterates using the gradient of a single component function one at a time [3]. This method can be faster than nonincremental methods since each step is relatively cheaper (one gradient computation instead of \( m \) gradient computations in the nonincremental case) and each step makes reasonable progress on average [3]. However, IG requires the stepsize to go to zero to obtain convergence to an optimal solution of problem (1.1) even if it is applied to smooth and strongly convex component functions [4], unless a restrictive “gradient growth condition” holds [27]. As a consequence,
with a decaying stepsize, IG has typical sublinear convergence rate properties. The same observation applies both to stochastic gradient methods (which uses a random order for cycling through the component functions) [26] and to incremental Newton methods that are of second-order [12].

Another interesting class of methods includes the incremental aggregated gradient (IAG) method of Blatt and others (see [6, 28]) and closely related stochastic methods including the stochastic average gradient (SAG) method [24], the SAGA method [8], and the MISO method [16]. The applications include but are not limited to logistic regression and binary regression with $\ell_2$ regularisation and more recently training conditional random fields [25]. These methods process a single component function at a time as in incremental methods, but keep a memory of the most recent gradients of all component functions so that an approximate gradient descent (GD) direction of $f$ is taken at each iteration. They might require an excessive amount of memory when $m$ is large; however, they have fast convergence properties on strongly convex functions with constant stepsize without requiring the restrictive gradient growth condition. Furthermore, IAG forms approximations to the gradient of the objective function $\nabla f(x)$ at each step and this provides an accurate and efficient stopping criterion (stop if the norm of the approximate gradient is below a certain threshold), whereas it is often not clear “when to stop” with IG.

The global convergence of IAG was established by Blatt, Hero, and Gauchman in [6] under some assumptions. This paper also shows that in the special case when each component function $f_i$ is a quadratic, IAG exhibits global linear convergence if the stepsize is small enough; however, neither an explicit convergence rate nor an explicit upper bound on the stepsize that can lead to linear convergence was given. This result is based on a perturbation analysis of the eigenvalues of a periodic dynamic linear system which is of independent interest in terms of the techniques used but is also highly technical and computationally demanding as it requires estimating the derivatives of the eigenvalues of a one-parameter matrix family. Furthermore, it only applies to quadratic functions. More recently, Tseng and Yun [28] proved global convergence under less restrictive conditions and local linear convergence in a more general setting when each component function satisfies a local Lipschitzian error condition, a condition satisfied by locally strongly convex functions (around an optimal solution). Although the results are more general than those of [6] as they apply beyond quadratics, the proofs are still involved and do not contain any explicit rate estimates because (i) the constants involved in the analysis are implicit and hard to compute/approximate, and (ii) the results are asymptotic (they hold when the stepsize is small enough but bounds on the stepsize are not available). See Remark 3.4 for more detail.

In this paper, we present a novel convergence analysis for the deterministic IAG method for the general case when the component functions are convex and smooth (and the sum function is strongly convex) with several advantages and implications. First, our analysis is based on a careful choice of a Lyapunov function which leads to simple global and linear convergence proofs. Furthermore, viewing IAG as a perturbed GD method where gradient errors can be interpreted as shocks that are fading away as one gets closer to an optimal solution, our proofs provide explicit bounds on how fast these shocks decay, giving more insight into the behavior of IAG. Second, to our knowledge, our analysis is the first to provide explicit rate estimates for (deterministic) IAG methods. Third, we discuss an “IAG method with momentum” and show its global and linear convergence. To our knowledge, this is the first global convergence and linear convergence result for an aggregated gradient method with memory.
Considering deterministic orders within incremental algorithms (IG and IAG) is important since in many applications, the agents processing each of the component functions are arranged according to a fixed configuration, which imposes a deterministic order on the way these functions are accessed. Explicit examples include estimation and source localization problems over sensor networks in which a large number of sensors are connected to each other over a ring network [6, Section 4], [23]. In these problems, each sensor is equipped with one of the component functions (corresponding to a loss function for the data block acquired by this sensor) and can exchange information with the other sensors in his neighborhood (using wireless communication) according to the underlying network. This necessitates processing component functions in a cyclic order dictated by the ring network making randomized access to sensors or component functions impractical or impossible. Similar considerations arise for distributed empirical risk minimization problems where data is spatially distributed among multiple computational units connected to each other through a fixed network. Let us also note that in addition to being a natural order in some applications, deterministic orders are important to consider since their analysis for incremental methods could be a key step in understanding the convergence behavior of general component function sampling schemes (see the analysis of random reshuffling in [13] which built on the deterministic IG analysis provided in [11]).

There has been some recent work on the SAG algorithm, the stochastic version of the IAG method where the order is stochastic. For SAG, Le Roux, Schmidt, and Bach [24] and later Defazio, Bach, and Lacoste-Julien [8] established a global linear convergence rate in expectation which is a weaker average sense of convergence compared to the deterministic convergence we will consider in this work. In particular, our results apply to any order (as long as each function is visited at least once in a finite number of steps $K$) and hold deterministically (not in a probabilistic sense). As the performance of the deterministic incremental methods are sensitive to the specific order chosen (see, e.g., [3, Example 1.5.6], [5, Figure 2.1.9]), IAG can be slower than SAG if an unfavorable order is chosen and the analysis of IAG has to account for this worst-case scenarios. This is also reflected in Theorem 3.3, where the rate of IAG has a worse (quadratic) dependance in the condition number $Q$ (see (3.1) for a definition), whereas SAG has linear dependance [24, Proposition 1]. We also note that most of the ideas used in the stochastic setting (such as the fact that the expected gradient error is zero) do not apply to the deterministic setting, necessitating a new approach for analyzing IAG. We emphasize that our aim is not to provide an algorithm with better performance than existing stochastic algorithms (including SAG or other variants that use a stochastic order for processing the functions) for additive convex optimization but rather to fill a gap in the convergence theory of deterministic incremental algorithms which are motivated by networked applications (in particular, decentralized optimization in sensor networks) where stochastic algorithms are not feasible or are not practical due to physical constraints.

Agarwal and Bottou [1] established lower complexity bounds for a class of deterministic incremental algorithms including IAG for solving the problem (1.1). Their results imply the lower bound $\Omega(m + \sqrt{m(Q-1)\log(1/\varepsilon)}$ on the iteration complexity of the IAG algorithm where $Q$ is the condition number of the problem (1.1). If the component functions are processed in a cyclic fashion, Theorem 3.3 in this paper shows the upper bound of $O(m^2Q^2\log(1/\varepsilon))$ for the IAG method. There is still room for improvement for the gap between this upper bound and the lower bound not only for IAG but also for other deterministic and stochastic incremental algorithms [1].
2. IAG method. For a constant stepsize $\gamma > 0$, an integer $K \geq 0$, and arbitrary initial points $x^0, x^{-1}, \ldots, x^{-K} \in \mathbb{R}^n$, the IAG method consists of the iterations

\begin{align}
  g^k &= \sum_{i=1}^{m} \nabla f_i(x^{\tau^k_i}), \\
  x^{k+1} &= x^k - \gamma g^k, \quad k = 0, 1, 2, \ldots,
\end{align}

where the gradient sampling times $\{\tau^k_i\}_{i=1}^{m}$ can be arbitrary as long as they are sampled at least once in the last $K$ iterations, i.e.,

\begin{equation}
  k \geq \tau^k_i \geq k - K, \quad i = 1, 2, \ldots, m.
\end{equation}

In other words, $K$ is an upper bound on the delay encountered by the gradients of the component functions. The update (2.1) determines the direction of motion $-g^k$ by approximating the steepest descent direction $-\nabla f(x^k)$ at (iterate) time $k$ from the recently computed gradients of the component functions (at times $\{\tau^k_i\}_{i=1}^{m}$). For example, if the component functions are processed one by one using a deterministic cyclic order on the index set $\{1, 2, \ldots, m\}$ with initialization $\tau^0_i = 0$ for all $i$, then $\tau^k_i$ admits the recursion

\begin{equation}
  \tau^k_i = \begin{cases} 
  k & \text{if } i = (k - 1 \mod m) + 1, \\
  \tau^{k-1}_i & \text{else,}
  \end{cases} \quad 1 \leq i \leq m, \quad k = 1, 2, \ldots,
\end{equation}

which satisfies $\tau^k_i \geq k - (m - 1)$ for all $i,k$ where $K = m - 1$. This is the original IAG method introduced by Blatt, Hero, Gauchman [6]. Later on, Tseng and Yun [28] generalized this method by allowing more general gradient sampling times $\{\tau^k_i\}$ with bounded delays, i.e., satisfying (2.3).

As IAG takes an approximate steepest GD direction, it is natural to analyze it as a perturbed steepest GD method. In fact, in the special case, when $K = 0$, all the gradients are up to date and IAG reduces to the classical (nonincremental) GD method, which is well known. Therefore, the more interesting case that we will analyze is when $K$ is strictly positive. For simplicity of notation in our analysis, we will also take

\begin{equation}
  x^0 = x^{-1} = \cdots = x^{-K}.
\end{equation}

This results in the initialization $\tau^0_i = 0$ for all $i$. However, it will be clear that our analysis can be extended to other (arbitrary) choices of initial points $\{x^j\}_{j=-K}^{0}$ in a straightforward manner.
3. Convergence analysis.

3.1. Preliminaries. We will make the following assumptions that have appeared in a number of papers analyzing incremental methods including [27], [28], [18], [8], and [24].

Assumption 3.1 (strong convexity and Lipschitz gradients).

(i) Each $f_i$ is $L_i$-smooth, meaning it has Lipschitz continuous gradients on $\mathbb{R}^n$ satisfying

$$\|\nabla f_i(y) - \nabla f_i(z)\| \leq L_i \|y - z\| \quad \forall y, z \in \mathbb{R}^n,$$

where $L_i \geq 0$ is the Lipschitz constant, $i = 1, 2, \ldots, m$. Let $L = \sum_{i=1}^{m} L_i$.

(ii) The function $f$ is strongly convex on $\mathbb{R}^n$ with parameter $\mu > 0$, meaning that the function $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

It follows by the triangle inequality that $f$ has Lipschitz continuous gradients with a Lipschitz constant $L$.\(^1\) We define the condition number of $f$ as

$$Q = \frac{L}{\mu} \geq 1$$

(see, e.g., [20]). As an example, in the special case when each $f_i$ is a quadratic function, $\nabla^2 f_i(x)$ is a constant (matrix) for each $i$ and we can take $L_i$ to be its largest eigenvalue, whereas the strong convexity constant $\mu$ can be taken as the smallest eigenvalue of the Hessian of $f$.

A consequence of Assumption 3.1 on the strong convexity of $f$ is that there exists a unique optimal solution of the problem (1.1) which we denote by $x^*$. In addition to the strong convexity, by the gradient Lipschitzness assumption, we have

$$\|\nabla f(x)\| \leq L \|x - x^*\| \quad \forall x \in \mathbb{R}^n,$$

$$f(x) - f(x^*) \leq \frac{L}{2} \|x - x^*\|^2 \quad \forall x \in \mathbb{R}^n$$

(see [20, Theorem 2.1.5]). In addition, it follows from [20, Theorem 2.1.12] that

$$\langle \nabla f(x), x - x^* \rangle \geq \frac{\mu L}{\mu + L} \|x - x^*\|^2 + \frac{1}{\mu + L} \|\nabla f(x)\|^2 \quad \forall x \in \mathbb{R}^n.$$

We finally introduce the following lemma for proving the linear rate for IAG. We omit the proof due to space considerations; for a proof see Feyzmahdavian, Aytekin, and Johansson [9], where this lemma is used to analyze the effect of delays in a first-order method. Coarsely speaking, the intuition behind this lemma is the following: If a nonnegative sequence $\{V_k\}$ that decays to zero linearly obeying $V_{k+1} \leq pV_k$ for some $p < 1$ is perturbed with an additive (noise) shock term that depends on the recent history, i.e., the shock at step $k$ is on the order of $V_{k}$, where $\ell \in [k - d(k), k]$ and $d(k)$ is the time interval (duration) of the shock, the linear convergence property can be preserved if the shocks are small enough, but this comes at the expense of a degraded rate $r > p$ which is determined by the amplitude of the shocks (controlled by the parameter $q$) and the duration of the shocks.

Lemma 3.2. Let $\{V_k\}$ be a sequence of nonnegative real numbers satisfying

$$V_{k+1} \leq pV_k + q \max_{(k-d(k))_+ \leq \ell \leq k} V_{\ell}, \quad k \geq 0,$$

\(^1\) The Lipschitz constant $L = \sum_{i} L_i$ of $\nabla f_i$ may not be the best Lipschitz constant as Lipschitz constants $L_i$ of $\nabla f_i$ are subadditive.
for some nonnegative constants $p$ and $q$. If $p + q < 1$ and $0 \leq d(k) \leq d_{\max}$ for some positive constant $d_{\max}$, then

$$V_k \leq r^k V_0, \quad k \geq 1,$$

where $r = (p + q)^{1/d_{\max}}$.

### 3.2. Bounding gradient error.

We denote the distance to the optimal solution at iterate $k$ by

$$\text{(3.5)} \quad \text{dist}_k = \|x^k - x^*\|$$

and the gradient error by

$$\text{(3.6)} \quad e^k = g^k - \nabla f(x^k).$$

We will show that the gradient error can be bounded in terms of a finite sum involving distances of iterates to the optimal solution. Using the triangle inequality and the Lipschitzness of the gradients, for any $k \geq 0$,

$$\text{(3.7)} \quad \|e^k\| \leq \sum_{i=1}^{m} \|\nabla f_i(x^{\tau_i^k}) - \nabla f_i(x^k)\| \leq \sum_{i=1}^{m} L_i \|x^{\tau_i^k} - x^k\|.$$  

As the gradient delays are bounded by $K$ (see (2.3)), by a repetitive application of the triangle inequality, we obtain for any $k \geq 0$,

$$\text{(3.8)} \quad \|e^k\| \leq \sum_{i=1}^{m} L_i \sum_{j=\tau_i^k}^{k-1} \|x^{j+1} - x^j\| \leq L \sum_{j=(k-K)_+}^{k-1} \|x^{j+1} - x^j\|$$

$$\text{(3.9)} \quad \|e^k\| \leq \gamma L \sum_{j=(k-K)_+}^{k-1} \|g^j\| \leq \gamma L \sum_{j=(k-K)_+}^{k-1} (\|\nabla f(x^j)\| + \|e^j\|)$$

(with the convention that $\|e^0\| = 0$ which is implied by (2.4)). The inequality (3.9) provides a recursive upper bound for the gradient error that relates the gradient error $\|e^k\|$ to the previous gradient errors $\|e^j\|$ with $j < k$. Bounding the $\|e^j\|$ term in (3.9) with the previously obtained bound (3.7) on the gradient error and by the triangle inequality,

$$\text{(3.10)} \quad \|e^k\| \leq \gamma L \sum_{j=(k-K)_+}^{k-1} \left(\|\nabla f(x^j)\| + \sum_{i=1}^{m} L_i \|x^{\tau_i^j} - x^j\|\right)$$

$$\leq \gamma L \sum_{j=(k-K)_+}^{k-1} \left(\|\nabla f(x^j)\| + \sum_{i=1}^{m} L_i \left(\|x^{\tau_i^j} - x^*\| + \|x^* - x^j\|\right)\right)$$

$$\leq \gamma L \sum_{j=(k-K)_+}^{k-1} \left(\|\nabla f(x^j)\| + \sum_{i=1}^{m} L_i \left(\max_{t \in \{\tau_i^j\}_{i=1}^{m}} \text{dist}_t + \text{dist}_j\right)\right).$$
Invoking (2.3) on the boundedness of the gradient delays by $K$ once more and using (3.2) on gradient Lipschitzness to bound the norm of the gradient, we finally get

$$
\|e^k\| \leq \gamma L \sum_{j=(k-K)}^{k-1} \left( \|\nabla f(x^j)\| + 2L \max_{(j-K)+ \leq \ell \leq j} \text{dist}_\ell \right)
$$

$$
\leq \gamma L \sum_{j=(k-K)}^{k-1} \left( L \text{dist}_j + 2L \max_{(j-K)+ \leq \ell \leq j} \text{dist}_\ell \right)
$$

(3.11) \quad \leq \gamma L \sum_{j=(k-K)}^{k-1} \left( 3L \max_{(j-K)+ \leq \ell \leq j} \text{dist}_\ell \right) \leq 3\gamma L^2 K \max_{(k-2K)+ \leq \ell \leq k-1} \text{dist}_\ell.
$$

3.3. Linear convergence analysis. From the IAG update formulas (2.1)–(2.2) and the definition (3.6) of the gradient error, it follows directly by taking

(3.12) \quad \text{dist}^2_{k+1} = \text{dist}^2_k - 2\gamma \langle \nabla f(x^k), x^k - x^* \rangle + \gamma^2 \|\nabla f(x^k)\|^2 + E_k,

where the gradient errors are captured by the last term,

(3.13) \quad E_k = \gamma^2 \|e^k\|^2 - 2\gamma \langle x^k - x^*, -\nabla f(x^k), e^k \rangle.

Using inequality (3.4) for strongly convex functions,

(3.14) \quad \text{dist}^2_{k+1} \leq \left( 1 - 2\gamma \frac{\mu L}{\mu + L} \right) \text{dist}^2_k + \gamma \left( \gamma - \frac{2}{\mu + L} \right) \|\nabla f(x^k)\|^2 + E_k

(3.15) \quad \leq \left( 1 - 2\gamma \frac{\mu L}{\mu + L} \right) \text{dist}^2_k + E_k \quad \text{if} \quad \gamma \leq \frac{2}{\mu + L}.

Note that when $K = 0$, IAG reduces to the GD method where $e^k = 0$ and the $E_k$ term vanishes. In this special case, note that the equality (3.12) gives the identity

$$
\|x^k - x^* - \gamma \nabla f(x^k)\|^2 = \|x^k - x^*\|^2 - 2\gamma \langle \nabla f(x^k), x^k - x^* \rangle + \gamma^2 \|\nabla f(x^k)\|^2.
$$

Applying the inequality (3.4) to this expression, we obtain similarly

(3.16) \quad \|x^k - x^* - \gamma \nabla f(x^k)\| \leq \left( 1 - 2\gamma \frac{\mu L}{\mu + L} \right) \text{dist}^2_k \leq \text{dist}^2_k

for any $x^k \in \mathbb{R}^n$ as long as $\gamma \leq \frac{2}{\mu + L}$. Here, a choice of $\gamma = \frac{2}{\mu + L}$ leads to the global linear convergence satisfying

$$
\text{dist}_{k+1} = \|x^k - x^* - \frac{2}{\mu + L} \nabla f(x^k)\| \leq \left( \frac{Q - 1}{Q + 1} \right) \text{dist}_k \leq \left( \frac{Q - 1}{Q + 1} \right)^k \text{dist}_0.
$$

This is the standard analysis of the GD method (see [20, Theorem 2.1.15]). The next theorem shows that in the more interesting general case when there are gradient errors, i.e., when $K > 0$ and $e^k \neq 0$, a similar linear convergence argument can be done, although one has to use a smaller stepsize to compensate for the gradient errors. The main idea is to eliminate the gradient error $\|e^k\|$ terms in (3.15) by replacing them with terms involving only distances. This can be done by invoking (3.11) which essentially provides an upper bound for the gradient errors in terms of distances. Then, Lemma 3.2 with $V_k = \text{dist}^2_k$ applies and provides the convergence rate.
THEOREM 3.3. Suppose that Assumption 3.1 holds. Consider the IAG iterations (2.1)-(2.2) with an integer gradient delay parameter \( K > 0 \) and a constant stepsize \( 0 < \gamma < \bar{\gamma} \), where

\[
\bar{\gamma} = \left( \frac{a \mu}{KL} \right) \frac{1}{\mu + L}
\]

with \( a = 8/25 \). Then, IAG iterates \( \{x^k\} \) are globally linearly convergent. Furthermore, when \( \gamma = \gamma^* := \bar{\gamma}/2 \) we have for \( k = 1, 2, \ldots, \)

\[
\|x^k - x^*\| \leq \left( 1 - \frac{c_K}{(Q + 1)^2} \right)^k \|x^0 - x^*\|,
\]

\[
f(x^k) - f(x^*) \leq \frac{L}{2} \left( 1 - \frac{c_K}{(Q + 1)^2} \right)^{2k} \|x^0 - x^*\|^2,
\]

where \( c_K = \frac{2}{45} \left[ K(2K + 1) \right]^{-1} \).

Proof. As \( K > 0 \), there are gradient delays and the error term \( E_k \) defined by (3.13) that appears in the evolution (3.15) of iterates is nonzero in general. Therefore, the convergence rate is limited by how fast this error term decays. Assume \( \gamma \leq \frac{2}{\mu + L} \) so that (3.15) is applicable. Using the triangle inequality on \( E_k \), we see that

\[
|E_k| \leq \gamma^2 \|e^k\|^2 + 2\gamma \|e^k\| \|x^k - x^* - \gamma \nabla f(x^k)\|.
\]

Note that the inequality (3.16) is valid here because its derivation requires only Assumption 3.1 to be satisfied. By using this inequality to upper bound \( \|x^k - x^* - \gamma \nabla f(x^k)\| \), we obtain

\[
|E_k| \leq 9\gamma^4 L^4 K^2 \max_{(k-2) \leq \ell \leq k-1} \text{dist}_\ell^2 + 6\gamma^2 L^2 K \max_{(k-2) \leq \ell \leq k-1} \text{dist}_\ell \text{dist}_k
\]

\[
\leq \left( 9\gamma^4 L^4 K^2 + 6\gamma^2 L^2 K \right) \max_{(k-2) \leq \ell \leq k} \text{dist}_\ell^2.
\]

Plugging this bound into the recursive inequality (3.15) for distances leads to

\[
\text{dist}_{k+1}^2 \leq p(\gamma) \text{dist}_k^2 + q(\gamma) \max_{(k-2) \leq \ell \leq k} \text{dist}_\ell^2
\]

with

\[
p(\gamma) = 1 - 2\gamma \frac{\mu L}{\mu + L}, \quad q(\gamma) = 9\gamma^4 L^4 K^2 + 6\gamma^2 L^2 K.
\]

If the stepsize is small enough satisfying the condition \( s(\gamma) := p(\gamma) + q(\gamma) < 1 \), then by Lemma 3.2 applied with \( V_k = \text{dist}_k^2 \), IAG is globally linearly convergent.\(^2\) Ignoring the positive \( O(\gamma^4) \) term in \( q(\gamma) \), this condition would require at least

\[
1 - 2\gamma \frac{\mu L}{\mu + L} + 6\gamma^2 L^2 K < 1 \iff 0 < \gamma < \left( \frac{\mu L}{3L K} \right) \frac{1}{\mu + L} = \frac{25}{24} \bar{\gamma},
\]

which would imply

\[
\gamma^2 L^2 K \leq \frac{1}{9K(Q + 1)^2} \leq \frac{1}{36}
\]

\(^2\)Note that when \( K = 0 \), the inequality \( s(\gamma) < 1 \) is linear in \( \gamma \), whereas when \( K > 0 \) it is fourth-order in \( \gamma \) and is more restrictive.
as both $K \geq 1$ and $Q \geq 1$. We will show that under the slightly more restrictive condition $0 < \gamma < \bar{\gamma}$, one can also handle the $O(\gamma^4)$ term as well and guarantee $s(\gamma) < 1$. So assume $0 < \gamma < \bar{\gamma}$. The condition (3.21) holds and the inequality (3.22) is valid. This implies that

$$ q(\gamma) = 3\gamma^2 L^2 K (2 + 3\gamma^2 L^2 K) \leq 3\gamma^2 L^2 K \left( 2 + \frac{1}{12} \right) \leq \frac{25}{4} \gamma^2 L^2 K $$

and after straightforward calculations that

$$ s(\gamma) := p(\gamma) + q(\gamma) \leq 1 - 2\gamma \frac{\mu L}{\mu + L} + \frac{25}{4} \gamma^2 L^2 K < 1. $$

Then by Lemma 3.2, we have global linear convergence of the sequence $V_k = \text{dist}_k^2$ to zero with rate $\rho(\gamma) = s(\gamma)^{1/(2K+1)} < 1$. This shows the global linear convergence of IAG. It remains to show the claimed convergence rate for $\gamma = \gamma_*$. Let $\gamma = \gamma_*$. Note that this minimizes the quadratic with respect to $\gamma$ in (3.23), leading to

$$ s(\gamma_*) \leq 1 - \frac{4\mu^2}{25K(\mu^2 + L^2)} = 1 - \frac{4}{25K(Q+1)^2}. $$

Then, as the linear convergence rate of the sequence $\{\text{dist}_k^2\}$ is $\rho(\gamma_*) < 1$, by taking square roots, the sequence $\{\text{dist}_k\}$ is linearly convergent with rate $r_* = \rho(\gamma_*)^{1/2}$ satisfying

$$ \text{dist}_k \leq r_* \text{dist}_0, \quad r_* = \left( s(\gamma_*)^{1/(2K+1)} \right)^{1/2} \leq 1 - \frac{cK}{(Q+1)^2}, $$

where we used (3.24) and the inequality $(1 - x)^a \leq 1 - ax$ for $x, a \in [0, 1]$ to get an upper bound for $\bar{\rho}$. This proves the rate (3.18). Then, (3.19) follows directly from (3.3) and (3.18).

Remark 3.4 (comparison with previous results). In a related work, Tseng and Yun show the existence of a positive constant $C$ that if the stepsize $\gamma$ is small enough, then IAG has a $K$-step linear convergence rate of $\sqrt{1 - C\gamma}$ under a Lipschitzian error assumption which is a strong-convexity-like condition [28, Theorem 6.1]. However, in their analysis, there is no explicit rate estimate, as (i) results are asymptotic, holding for $\gamma$ small enough without giving a precise interval for $\gamma$, and (ii) the constant $C$ is implicit and hard to compute/approximate as it depends on several other implicit constants and a Lipschitzian error parameter $\tau$. Our analysis not only is simple using basic distance inequalities but also the constants are transparent and explicit.

Remark 3.5 (IAG versus IG). Theorem 3.3 shows that IAG with constant stepsize is globally linearly convergent; however, the same is not true for IG. In fact, IG with constant stepsize is linearly convergent to a neighborhood of the solution but does not in general converge to the optimal solution due to the existence of gradient errors that are typically bounded away from zero [3]. As a consequence, achieving global convergence with IG requires one to use a stepsize that goes to zero and this results in typically slow convergence [4]. In contrast, the gradient error in IAG is controlled by the distance of recent iterates to the optimal solution; therefore it is attenuated as the iterates get closer to the optimal solution and a diminishing stepsize is not needed to control the error.
Remark 3.6 (local strong convexity implies local linear rate). We note that when $f$ is not globally strongly convex but only locally strongly convex around a stationary point, for instance, when the Hessian is not degenerate around a locally optimal solution, by reasoning along the lines of the proof of Theorem 3.3, it is possible to show that IAG is locally linearly convergent.

4. IAG with momentum. An important variant of the GD method is the heavy-ball method [22], which extrapolates the direction implied by the previous two iterates by the following update rule:

$$x^{k+1} = x^k - \gamma \nabla f(x^k) + \beta (x^k - x^{k-1}),$$

where $\beta \geq 0$ is the momentum parameter. It can be shown that the heavy-ball method can achieve a faster local convergence than GD when $\beta$ is in a certain range [22, section 3.1]. There has also been much interest in understanding its global convergence properties [10, 14]. Accelerated gradient methods introduced by Nesterov [19, 21] can also be thought of as momentum methods where the momentum parameter is variable and appropriately chosen. There has been a lot of recent interest in these accelerated methods as they have optimal iteration complexity properties under some conditions [20].

In contrast to the recent advances in nonincremental methods with momentum, there has been less progress on incremental methods with momentum. In particular, no deterministic incremental methods with favorable convergence characteristics similar to those of accelerated gradient methods are currently known. However, there is the IG method with momentum, which consists of the inner iterations

$$x^{k+1}_i = x^k_i - \gamma^k \nabla f_i(x^k) + \beta (x^k - x^{k-1}), \quad i = 1, 2, \ldots, m, \quad k \geq 1,$$

starting from $x^1_i \in \mathbb{R}^n$ with the convention that $x^{k+1}_i = x^{k+1}_{m+1}$, where $\gamma^k$ is the stepsize [2, 17, 29]. This method can be faster than IG on some problems, especially when gradients have oscillatory behavior; however, it would still require the stepsize go to zero due to gradient errors, leading to typical sublinear convergence [29]. It is natural to ask whether IAG with such an additional momentum term, which we abbreviate by IAG-M,

$$x^{k+1} = x^k - \gamma^k g^k + \beta (x^k - x^{k-1}), \quad k = 0, 1, 2, \ldots,$$

would be globally convergent for $\beta$ in some range $(0, \bar{\beta})$. We expect that this algorithm can outperform IAG in problems where the individual gradients show oscillatory behavior because the momentum term provides an extra smoothing/averaging effect on the iterates. The global linear convergence of the IAG-M method for a certain range of $\beta$ values can be shown by similar reasoning along the lines presented in section 3. Most of the logic in the derivation of the inequalities (3.7)–(3.11) applies with the only difference that the $\|x^{j+1} - x^j\|$ terms will now contain an additional momentum term due to the modified update rule (4.1). We, however, provide a sketch of the proof in Appendix A for the sake of completeness.

5. Discussion. We analyzed the IAG method when component functions are strongly convex by viewing it as a GD method with errors. To the best of our knowledge, our analysis provides the first explicit linear rate result. Furthermore, it is different from the existing two approaches [6] and [28] in the sense that (i) it is based on simple basic inequalities that make global convergence analysis simpler, and
(ii) it gives more insight into the behavior of IAG. In particular, our analysis shows that the gradient errors can be treated as shocks with a finite duration which can be bounded in terms of distance of iterates to the optimal solution. Therefore, by choosing the stepsize small enough and using the strong convexity properties we can guarantee that the distance to the optimal solution shrinks down at each step by a factor less than one.

We also developed a new algorithm, IAG with momentum, and provided a linear convergence and rate analysis. It is expected that this algorithm can outperform IAG in problems where the individual gradients show oscillatory behavior, because the momentum term provides an extra (averaging) smoothing affect on the iterates.

We note that the extension of IAG to the generalized version of (1.1), \( \min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x) + h(x) \) with \( h : \mathbb{R}^n \to \mathbb{R} \) convex and possibly nonsmooth (such as the indicator of a function when there are constraints) is simple by an additional (proximal) step; see [28]. Extending our linear rate results to this case may be possible and is ongoing future work.

Appendix A. Proof sketch of the global linear convergence of the IAG-M method. Using the gradient error bound (3.8) and iterate update equation (4.1) of IAG-M,

\[
\|e^k\| = \|g^k - \nabla f(x^k)\| \leq L \sum_{j=(k-K)_+}^{k-1} \|x^{j+1} - x^j\| = L \sum_{j=(k-K)_+}^{k-1} \|g^j + \beta(x^j - x^{j-1})\|
\]

\[
\leq \gamma L \sum_{j=(k-K)_+}^{k-1} \|g^j\| + \beta L \sum_{j=(k-K)_+}^{k-1} \|x^j - x^{j-1}\|
\]

\[
\leq \gamma L \sum_{j=(k-K)_+}^{k-1} (\|\nabla f(x^j)\| + \|e^j\|) + \beta L \sum_{j=(k-K)_+}^{k-1} \|x^j - x^{j-1}\|.
\]

Then, using (3.2) to bound the norm of the gradient and (3.8) to bound \(\|e^j\|\), this becomes

\[
(A.1)
\]

\[
\|e^k\| \leq \gamma L \sum_{j=(k-K)_+}^{k-1} (L \text{dist}_j + \|e^j\|) + \beta L \sum_{j=(k-K)_+}^{k-1} \|x^j - x^{j-1}\|
\]

\[
\leq \gamma L \sum_{j=(k-K)_+}^{k-1} \left( L \text{dist}_j + L \sum_{\ell=(j-K)_+}^{j-1} \|\text{dist}_\ell + x^\ell\| \right) + \beta L \sum_{j=(k-K)_+}^{k-1} \|x^j - x^{j-1}\|
\]

\[
\leq \gamma L \sum_{j=(k-K)_+}^{k-1} \left( L \text{dist}_j + L \sum_{\ell=(j-K)_+}^{j-1} (\text{dist}_\ell + \text{dist}_{\ell+1}) \right) + \beta L \sum_{j=(k-K)_+}^{k-1} \|x^j - x^{j-1}\|
\]

\[
\leq (3\gamma L^2 K) \max_{(k-2K)_+ \leq \ell \leq k-1} \text{dist}_\ell + \beta L \sum_{j=(k-K)_+}^{k-1} \|x^j - x^{j-1}\|.
\]

Note that when \( \beta = 0 \), this inequality reduces to (3.11) obtained for IAG. From the inner update equation (4.1), we also have
\[ \|x^j - x^{j-1}\| \leq \gamma \|g^{j-1}\| + \beta \|x^{j-1} - x^j\| \\
\leq \gamma \|g^{j-1}\| + \beta (\|x^{j-1} - x^j\| + \|x^j - x^\ast\|) \\
\leq \gamma \|\nabla f(x^{j-1})\| + \gamma \|e^{j-1}\| + \beta (\text{dist}_{j-1} + \text{dist}_{j-2}) \\
\leq \gamma L \text{dist}_{j-1} + \gamma \|e^{j-1}\| + 2\beta \max(\text{dist}_{j-1}, \text{dist}_{j-2}), \tag{A.2} \]

where we used (3.2) to bound the norm of the gradient in the last inequality. We next bound the gradient error term \(\|e^{j-1}\|\) on the right-hand side. A consequence of (3.7), the triangle inequality, and the boundedness of the gradient delays is that gradient error is bounded by

\[ \|e^k\| \leq \sum_{i=1}^{m} L_i (\text{dist}_{r_i} + \text{dist}_k) \leq 2L \max_{(k-K)_+ \leq \ell \leq k} \text{dist}_\ell, \quad k \geq 0. \tag{A.3} \]

Combining all the inequalities (A.2), (A.2), and (A.3) together leads to

\[ \|e^k\| \leq (3\gamma L^2 K + \beta L K (3\gamma L + 2\beta)) \max_{(k-2K-1)_+ \leq \ell \leq k-1} \text{dist}_\ell, \tag{A.4} \]

which is an analogue of (3.11) for IAG-M. Then, following the same line of argument with the proof of Theorem 3.3, it is straightforward to show a linear rate for IAG-M as long as the momentum parameter \(\beta\) is not very large. This is the subject of the next theorem, whose proof is outlined for the sake of completeness.

**Theorem A.1.** Suppose that Assumption 3.1 holds. Consider the IAG-M iterates generated by the update rules (2.1) and (4.1) with constant stepsize \(\gamma < \tilde{\gamma}\), where \(\tilde{\gamma}\) is defined by (3.17) and momentum parameter \(\beta > 0\). Then, there exists a positive constant \(\tilde{\beta} > 0\) such that if \(\beta < \tilde{\beta}\), then the IAG-M iterates \(\{x^k\}\) are globally linearly convergent.

**Proof.** We can follow the same reasoning given in the proof of Theorem 3.3 with the only difference that the gradient error bound (A.4) for IAG-M is used instead of the gradient error bound (3.11) for IAG. Then, we obtain a perturbed version of (3.20):

\[ \text{dist}_{k+1}^2 \leq p(\gamma) \text{dist}_k^2 + q_{\beta}(\gamma) \max_{(k-2K)_+ \leq \ell \leq k} \text{dist}_\ell^2, \tag{A.5} \]

where \(q_{\beta}(\gamma)\) is of the form

\[ q_{\beta}(\gamma) = q(\gamma) + \tilde{q}(\gamma, \beta), \]

where \(\tilde{q}(\gamma, \beta) : [0, \infty) \times [0, \infty) \to \mathbb{R}\) is a function with a polynomial dependence in \(\alpha\) and \(\beta\) satisfying \(\tilde{q}(\gamma, 0) = 0\).\(^3\) As \(\tilde{q}(\gamma, \beta)\) is a polynomial in \(\beta\) (and therefore continuous in \(\beta\)), by reasoning along the lines of the proof of Theorem 3.3, we conclude that \(p(\gamma) + q_\beta(\gamma) < 1\) for \(\beta\) small enough and \(\gamma < \tilde{\gamma}\). This completes the proof. \(\square\)

**Appendix B. Numerical experiments for IAG-M.** Note that the analysis of IAG-M is based on the gradient error bound (A.4), which is weaker than the estimate (3.11). Therefore, Theorem A.1 does not lead to a rate faster than the rate of IAG given by Theorem 3.3. Nevertheless, in practice IAG-M is often faster than IAG for problems with a high condition number. This is illustrated in Figure 1, where we compare the performance of IAG and IAG-M over randomly generated quadratic functions.

\(^3\)Therefore, when \(\beta = 0\), (A.5) reduces to (3.20). Note that an explicit expression for \(\tilde{q}(\gamma, \beta)\) can also be derived, which we will not attempt here.
For the numerical experiments given by Figure 1, we generate the component functions to be random strongly convex quadratics of the form $f_i(x) = x^T H_i x$ where $H_i$ is diagonal with its largest and smallest eigenvalues satisfying $E(\lambda_{\text{max}}) = \sqrt{Q}$ and $E(\lambda_{\text{min}}) = \frac{1}{\sqrt{Q}}$ where $Q$ is the condition number. We experiment with the condition number by choosing it as $Q = 10^j m$ for $j = 0, 4, 8$ similar to the experiments carried out in [15]. Initial point $x^0$ is created randomly and the momentum parameter is set to $\beta = \frac{10 \sqrt{Q} - 1}{m \sqrt{Q} + 1}$. The first column in Figure 1 shows that if the condition number is relatively small, IAG and IAG-M behave similarly. However, the second and third columns show that when the condition number is larger, IAG-M outperforms IAG. We repeated the experiment with different initialization and $n$ and $m$ values and found similar results.

REFERENCES