SIAM J. OPTIM.
Vol. 29, No. 4, pp. 2542–2565

CONVERGENCE RATE OF INCREMENTAL GRADIENT AND INCREMENTAL NEWTON METHODS*

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

Abstract. The incremental gradient (IG) method is a prominent algorithm for minimizing a finite sum of smooth convex functions and is used in many contexts including large-scale data processing applications and distributed optimization over networks. It is a first-order method that processes the functions one at a time based on their gradient information. The incremental Newton method, on the other hand, is a second-order variant which additionally exploits the curvature information of the underlying functions and can therefore be faster. In this paper, we focus on the case when the objective function is strongly convex and present new convergence rate estimates for the incremental gradient and incremental Newton methods under constant and diminishing step sizes. For a decaying step-size rule $\alpha_k = R/k^s$ with $s \in (0,1]$ and $R > 0$, we show that the distance of the IG iterates to the optimal solution converges at a rate $\mathcal{O}(1/k^s)$ (which translates into a $\mathcal{O}(1/k^{2s})$ rate in the suboptimality of the objective value). For $s > 1/2$, this improves the previous $\mathcal{O}(1/\sqrt{k})$ results in distances obtained for the case when functions are nonsmooth under the additional assumption that the functions are smooth. We show that to achieve the fastest $\mathcal{O}(1/k)$ rate with a step size $\alpha_k = R/k$, IG needs a step-size parameter R to be a function of the strong convexity constant whereas the incremental Newton method does not. The results are based on viewing the IG method as a gradient descent method with gradient errors, developing upper bounds for the gradient error to derive inequalities that relate distances of the consecutive iterates to the optimal solution and finally applying Chung's lemmas from the stochastic approximation literature to these inequalities to determine their asymptotic behavior. In addition, we construct examples to show tightness of our rate results in terms of their dependency in k.

Key words. convex optimization, incremental algorithms, first-order methods, convergence rate

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

DOI. 10.1137/17M1147846

1. Introduction.

We consider the following additive cost optimization problem

$$\min \sum_{i=1}^{m} f_i(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n, \tag{1.1}$$

where the objective function is the sum of a large number of convex *component functions* $f_i : \mathbb{R}^n \to \mathbb{R}$. Such problems arise in a number of settings including distributed optimization across m agents, where the component function f_i corresponds to the local objective function of agent i [10, 30, 31, 39], and statistical estimation problems, where each f_i represents the loss function associated with one of the data blocks [5, 9, 43, 45]. Our goal is to exploit the additive structure of problem (1.1) and solve it using incremental methods which involve sequential processing of component functions.

We first consider the *incremental gradient (IG) method* for solving problem (1.1). The IG method is similar to the standard gradient method with the key difference that at each iteration, the decision vector is updated incrementally by taking sequential steps along the gradient of the component functions f_i in a cyclic order. Hence, we

[*] Received by the editors September 18, 2017; accepted for publication (in revised form) July 5, 2019; published electronically October 15, 2019.
https://doi.org/10.1137/17M1147846

[†] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (mertg@mit.edu, asuman@mit.edu, parrilo@mit.edu).

can view each outer iteration k as a cycle of m *inner iterations*: starting from an initial point $x_1^1 \in \mathbb{R}^n$, for each $k \geq 1$ we update the iterate x_i^k as

$$x_{i+1}^k := x_i^k - \alpha_k \nabla f_i(x_i^k), \quad i = 1, 2, \ldots, m, \tag{1.2}$$

where $\alpha_k > 0$ is a step size. We set $x_1^{k+1} = x_{m+1}^k$ and refer to $\{x_1^k\}$ as the *outer iterates*. When the component functions are not smooth, we can replace gradients with subgradients and the corresponding method is called the *incremental subgradient method*. Using the update relation (1.2), for each $k \geq 1$, we can write down the relation between the outer iterates as

$$x_1^{k+1} = x_1^k - \alpha_k \sum_{i=1}^{m} \nabla f_i(x_i^k), \tag{1.3}$$

where $\sum_{i=1}^m \nabla f_i(x_i^k)$ is the *aggregated component gradients* and serve as an approximation to the full gradient $\nabla f(x_1^k)$ with the difference that it is evaluated at different inner iterates.

IG is a prominent algorithm with a long history that has appeared in many contexts. In the artificial intelligence literature, it has been used in training neural networks since the 1980s and is known as the online backpropagation algorithm [5, 23, 48]. When the component functions are quadratics, IG reduces to the well-known Kaczmarz method for solving linear systems [6].

Due to the simplicity and long history of the IG method, its global convergence has been studied under various conditions (see [5] for a survey). However, characterizing its convergence rate has been the subject of more recent work. Among the papers relevant to our work, Kohonen [21] focused on quadratic component functions with constant step size, $\alpha_k = \alpha > 0$ for all k, and showed that the iterates may converge to a limit cycle (subsequence of inner iterates converge to different limits close to optimal). The papers [2, 3, 7, 14, 15, 23, 24, 25, 46] focused on diminishing step size and showed convergence of the algorithm and its variants under different assumptions. The papers [44] and [28] studied IG with a constant step size and under different assumptions on the component functions, and showed that the iterates converge to a neighborhood of the optimal solution (where the size of the neighborhood is a function of the step size). Nedić and Bertsekas [29] focused on the convergence analysis of IG under different assumptions on the step size. Most closely related to our paper is a convergence rate result provided by Nedić and Bertsekas [28], which under a strong-convexity-type condition on the sum function $f(x) = \sum_{i=1}^m f_i(x)$, but without assuming differentiability of the component functions, shows that the distance of the iterates generated by the incremental subgradient method converges at rate $\mathcal{O}(\frac{1}{\sqrt{k}})$ to the optimal solution with a properly selected diminishing step size.[1]

Luo [23] considered a special case of problem (1.1) in dimension one when there are two convex quadratic component functions with an identical nonzero curvature and showed that IG iterates converge in this particular case at rate $\mathcal{O}(\frac{1}{k})$ to the optimal solution. Motivated by this example, in this paper we show that Nédic and Bertsekas's $\mathcal{O}(\frac{1}{\sqrt{k}})$ result can be improved when the component functions are smooth. In particular, when the component functions are quadratics and the sum function $f(x)$ is strongly convex, we first prove that the distances of the iterates generated by the IG method converge at rate $\mathcal{O}(\frac{1}{k})$ (which translates into $\mathcal{O}(\frac{1}{k^2})$ in function values by

[1] Given sequences $\{a_k\}$ and $\{b_k\}$, we write $a_k = \mathcal{O}(b_k)$ if $|a_k| \leq b_k$ for any k large enough, where $\mathcal{O}(\cdot)$ is also known as Landau's symbol.

the smoothness and strong convexity of f). Then, we generalize this result to twice continuously differentiable component functions under some assumptions. Achieving this rate with IG requires using a diminishing step size that adapts to the strong convexity constant c of the sum function, i.e., a step size that takes the form R/k, where $R > 1/c$.[2] We then consider alternative "robust" step sizes $\alpha_k = \Theta(\frac{1}{k^s})$ for $s \in (0,1)$, which does not require knowledge of the strong convexity constant, and show that the IG method with these step sizes achieves a rate $\mathcal{O}(\frac{1}{k^s})$ in distances (which translates into $\mathcal{O}(\frac{1}{k^{2s}})$ in function values). We also provide lower bounds showing that these rates cannot be improved using IG in terms of their dependency on k. We note however that our lower bounds are based on the construction of specific examples and are limited in the sense that they do not show the tightness of our IG analysis with respect to the dimension n and the number of component functions m.

Our results play a key role in the recently obtained convergence results for the *random reshuffling* (RR) method [17]. The random reshuffling method is a stochastic variant of IG where the order of visit to the functions is selected as a random permutation of $\{1, 2, \dots, m\}$ at the beginning of each cycle instead of the deterministic fixed order $\{1, 2, \dots, m\}$ of IG (hence the name RR refers to the random reshuffling of the order). Providing convergence rate results for the random reshuffling method has been a long-standing open question; see [40] and [41, section 5]. Fundamental to the analysis in [17], which provides the first asymptotic convergence rate results for RR methods with decaying step sizes under some technical assumptions on the objective function f, is the fast convergence rate results introduced in this paper, which applies to any order of visit to the component functions. In addition to providing the iteration complexity $\mathcal{O}(\frac{1}{k})$, our rate estimates also highlight the dependency on strong convexity and Lipschitz constant of the sum function.

The goal of our current paper is to focus on the smoothness assumptions typically considered in the recent literature and show that under these assumptions, IG admits better $O(1/k^{2s})$ bounds with respect to the existing literature. We note that the performance bounds in function values we provide for IG in this paper hide constants that depend linearly on m, which is expected since these are worst-case bounds, i.e., they apply to all possible orders. We also show in the paper that these bounds are tight for incremental methods. This is in contrast with stochastic incremental gradient methods such as stochastic gradient descent (SGD), which sample the component functions randomly and admit performance guarantees in expectation with leading constants that are independent of m (see [36] and [17] for an asymptotic theory of SGD and RR). If gradients of the component functions can be accessed in a random fashion, there are also variance-reduced stochastic gradient methods that require more ($O(m)$) memory but can converge linearly in expectation [13, 22]. Nevertheless, in many applications in distributed optimization, random access to component functions cannot be implemented because of communication constraints between the nodes, yet IG methods with an order consistent with the underlying network topology are applicable. As an example, consider a set of sensors arranged as a ring network, each collecting decentralized data, which therefore has access to a locally known cost function. The natural order to process the component functions in this example is a cyclic order whereby each sensor after local processing with his component function

[2] We note that a consequence of a paper by Hazan and Kale [19] is that when each of the component functions is strongly convex, IG with iterate averaging and step size $\alpha_k = R/k$, where R is the multiplicative inverse of the strong convexity constant, converges at rate $\mathcal{O}(\log k/k)$ in the suboptimality of the function value. However, the rate we obtain in this paper with a similar step size corresponds to $\mathcal{O}(1/k^2)$ in the suboptimality of the objective value, which is much faster.

passes the iterate to the neighboring sensor [8]. There are also many other examples beyond sensor networks that necessitate decentralized computation where stochastic incremental methods are impractical but the deterministic incremental gradient is applicable. Examples include but are not limited to multiagent control and coordination [32], learning [37], decentralized regression [42], estimation [20], and sparse optimization [49].

We next consider an *incremental Newton (IN) method*, introduced in [16] for solving problem (1.1), which scales the gradients of the component functions with an estimate of the Hessian of the sum function $f(x)$: starting from initial point $x_1^1 \in \mathbb{R}^n$, initial step size $\alpha_1 > 0$, and initial Hessian estimate $H_0^1 = I_n$, for each $k \geq 1$, the IN method updates the iterate x_i^k as

$$x_{i+1}^k := x_i^k - \alpha_k (\bar{H}_i^k)^{-1} \nabla f_i(x_i^k), \tag{1.4}$$

where

$$H_i^k := H_{i-1}^k + \nabla^2 f_i(x_i^k), \quad \bar{H}_i^k = H_i^k / k, \tag{1.5}$$

with the convention that $x_1^{k+1} = x_{m+1}^k$ and $H_0^{k+1} = H_m^k$. For IN, we provide rate estimates which do not depend on the Lipschitz constant. We show that the IN method, unlike IG, converges with rate $\mathcal{O}(\frac{1}{k})$ without using a step size that adapts to the strong convexity constant.

Notation. For nonnegative sequences $\{a_k\}$ and $\{b_k\}$, we write $a_k \geq \Omega(b_k)$ if there exists a real constant $h > 0$ and a finite integer k_0 such that $a_k \geq h b_k$ for every $k \geq k_0$. The norm $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices. We also write $a_k \geq \tilde{\Omega}(b_k)$ if there exists a real constant $h > 0$ and infinitely many k such that $a_k \geq h b_k$ is true.[3] The matrix I_n denotes the $n \times n$ identity. The sets $\mathbb{R}_+$ and $\mathbb{N}_+$ denote the positive real numbers and positive integers, respectively. We refer to twice continuously differentiable functions on $\mathbb{R}^n$ as *smooth* functions.

2. Preliminaries.

We introduce the following lemma, known as Chung's lemma, which we will make use of in our rate analysis. The proof of part (i) of this lemma can be found in [35, section 2.2]. For the proof of part (ii), we refer the reader to [12, Lemma 4].

LEMMA 2.1. *Let $\{u_k\}$ be a sequence of nonnegative numbers. Assume there exists k_0 such that*

$$u_{k+1} \leq \left(1 - \frac{a}{k^s}\right) u_k + \frac{d}{k^{s+t}} \quad \forall k \geq k_0,$$

where $0 < s \leq 1$, $d > 0$, $a > 0$, and $t > 0$ are given real numbers. Then we have the following.

(i) *If $s = 1$, then*

$$\begin{aligned} &\limsup_{k\to\infty} k^t u_k &&\leq \frac{d}{a-t} && \textit{for} \quad a > t, \\ &\limsup_{k\to\infty} \frac{k^a}{\log k} u_k &&< \infty && \textit{for} \quad a = t, \\ &\limsup_{k\to\infty} k^a u_k &&< \infty && \textit{for} \quad a < t. \end{aligned}$$

[3] The $\tilde{\Omega}$ function defined here was introduced by Littlewood and Hardy in 1914. It is a weaker alternative to the Ω function and satisfies $a_k \geq \Omega(b_k) \implies a_k \geq \tilde{\Omega}(b_k)$ but not vice versa.

(ii) *If* $0 < s < 1$, *then*[4]

$$\limsup_{k\to\infty} k^t u_k \le \frac{d}{a}.$$

3. Convergence rate analysis for IG.

3.1. Rate for quadratic functions.

We first analyze the convergence behavior when the component functions are quadratic functions before proceeding to the more general case when functions are twice continuously differentiable. Let $f_i(x) : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function of the form

$$f_i(x) = \frac{1}{2} x^T P_i x - q_i^T x + r_i, \quad i = 1, 2, \ldots, m, \tag{3.1}$$

where P_i is a symmetric $n \times n$ square matrix, $q_i \in \mathbb{R}^n$ is a column vector, and r_i is a real scalar. The gradient and the Hessian of f are given by

$$\nabla f_i(x) = P_i x - q_i, \quad \nabla^2 f_i(x) = P_i. \tag{3.2}$$

The sum f is also a quadratic which we next assume to be strongly convex.

Assumption 3.1. The sum function $f(x)$ is strongly convex on $\mathbb{R}^n$, i.e., there exists a constant $c > 0$ such that the function $f(x) - \frac{c}{2}\|x\|^2$ is convex on $\mathbb{R}^n$.

Under this assumption, the optimal solution to problem (1.1) is unique, and we denote it by x^*. In the particular case when each f_i is a quadratic function given by (3.2), the Hessian matrix of the sum satisfies

$$P := \nabla^2 f(x) = \sum_{i=1}^m P_i \succeq c I_n \succ 0, \tag{3.3}$$

and the optimal solution is

$$x^* = P^{-1} \sum_{i=1}^m q_i. \tag{3.4}$$

For this case, the inner iterations of IG become

$$x_{i+1}^k = (I_n - \alpha_k P_i) x_i^k + \alpha_k q_i, \quad i = 1, 2, \ldots, m.$$

Therefore, the outer iterations are given by

$$\begin{aligned} x_1^{k+1} &= \prod_{i=1}^m (I_n - \alpha_k P_i) x_1^k + \alpha_k \sum_{i=1}^m \prod_{j=i+1}^m (I_n - \alpha_k P_j) q_i \qquad (3.5)\\ &= \big(I_n - \alpha_k P + \mathcal{O}(\alpha_k^3)\big) x_1^k + \alpha_k \sum_{i=1}^m q_i + \alpha_k^2 T(\alpha_k) + \mathcal{O}(\alpha_k^3) \qquad (3.6)\\ &= \big(I_n - \alpha_k P\big) x_1^k + \alpha_k \sum_{i=1}^m q_i + \alpha_k^2 E(\alpha_k), \qquad (3.7) \end{aligned}$$

[4] Part (ii) of Lemma 2.1 is still correct when u_k is allowed to take negative values. However, this will not be needed in our analysis.

where

$$T(\alpha_k) = \sum_{1\leq i<j\leq m} P_j(P_i x_1^k - q_i) = \sum_{1\leq i<j\leq m} P_j \nabla f_i(x_1^k), \tag{3.8}$$

$$E(\alpha_k) = T(\alpha_k) + \mathcal{O}(\alpha_k) + \mathcal{O}(\alpha_k x_1^k). \tag{3.9}$$

Subtracting x^* from both sides of (3.6) and using the identity (3.4),

$$x_1^{k+1} - x^* = \big(I_n - \alpha_k P + \mathcal{O}(\alpha_k^3)\big)(x_1^k - x^*) + \alpha_k^2 T(\alpha_k) + \mathcal{O}(\alpha_k^3). \tag{3.10}$$

Similarly, (3.7) and (3.4) lead to

$$x_1^{k+1} - x^* = \big(I_n - \alpha_k P\big)(x_1^k - x^*) + \alpha_k^2 E(\alpha_k).$$

Taking norms of both sides of the last expression, defining

$$\mathrm{dist}_k = \|x_1^k - x^*\|$$

as the distance to the optimal solution, and using the lower bound (3.3) on the eigenvalues of P, we obtain

$$\begin{aligned} \mathrm{dist}_{k+1} &\leq \big\|I_n - \alpha_k P\big\| \mathrm{dist}_k + \alpha_k^2 \|E(\alpha_k)\| \\ &\leq (1-\alpha_k c)\mathrm{dist}_k + \alpha_k^2 \|E(\alpha_k)\| \quad (\text{if} \quad \alpha_k \|P\| \leq 1) &(3.11)\\ &\leq (1-\alpha_k c)\mathrm{dist}_k + \alpha_k^2 M_\infty \quad (\text{if} \quad \alpha_k \|P\| \leq 1), &(3.12)\end{aligned}$$

where

$$M_\infty := \sup_{k\geq 1} \|E(\alpha_k)\|. \tag{3.13}$$

The next theorem analyzes this recursion and the behavior of $\|E(\alpha_k)\|$ as k goes to infinity in order to establish convergence rate estimates for IG with quadratic component functions for different step-size rules. For this purpose, we introduce

$$M := \limsup_{k\to\infty} \|E(\alpha_k)\|, \tag{3.14}$$

which will be studied in the next theorem.

THEOREM 3.2. *Let each $f_i(x) = \frac{1}{2}x_i^T P_i x - q_i^T x + r_i$ be a quadratic function as in* (3.1) *for $i = 1, 2, \ldots, m$. Suppose Assumption* 3.1 *holds. Consider the iterates $\{x_1^k\}$ generated by the IG method with step size $\alpha_k = R/k^s$, where $R > 0$ and $s \in [0, 1]$. Then, we have the following.*

(i) *If $0 < s \leq 1$, then*

$$M = \Big\|\sum_{1\leq i<j\leq m} P_j \nabla f_i(x^*)\Big\|, \tag{3.15}$$

where M is defined by (3.14).

(ii) *If $s = 1$, then*

$$\begin{aligned} &\limsup_{k\to\infty} k\,\mathrm{dist}_k &&\leq \frac{R^2 M}{Rc-1} && \textit{for} \quad R > 1/c,\\ &\limsup_{k\to\infty} \frac{k}{\log k}\mathrm{dist}_k &&< \infty && \textit{for} \quad R = 1/c,\\ &\limsup_{k\to\infty} k^{Rc}\mathrm{dist}_k &&< \infty && \textit{for} \quad R < 1/c,\end{aligned}$$

where M is given by (3.15).

(iii) *If* $0 < s < 1$, *then*

$$\limsup_{k\to\infty} k^s \mathrm{dist}_k \le \frac{RM}{c},$$

where M *is given by* (3.15).

(iv) *If* $s = 0$ *and* $R \le \frac{1}{\|P\|}$, *then*

$$\mathrm{dist}_{k+1} \le (1 - c\alpha)^k \mathrm{dist}_1 + \frac{\alpha M_\infty}{c} \quad \forall k \ge 1, \tag{3.16}$$

where the step size $\alpha = \alpha_k = R$ *is a constant and* M_∞ *is defined by* (3.13).

Proof. We first prove parts (i), (ii), and (iii). Assume $0 < s \le 1$. Plugging the expression for the step size into (3.10), taking norms of both sides and using the inequality $\|P\| \ge c$, we obtain

$$\mathrm{dist}_{k+1} \le \left(1 - \frac{Rc}{k^s} + \mathcal{O}\left(\frac{1}{k^{3s}}\right)\right)\mathrm{dist}_k + \frac{R^2}{k^{2s}}\|T(\alpha^k)\| + \mathcal{O}\left(\frac{1}{k^{3s}}\right). \tag{3.17}$$

We define

$$M_* := \left\| \sum_{1\le i<j\le m} P_j \nabla f_i(x^*) \right\|.$$

We see from (3.8) and (3.9) that if IG is globally convergent, i.e., if $\mathrm{dist}_k = \|x_1^k - x^*\| \to 0$, then $E(\alpha_k)$ converges as $k \to \infty$ and in particular, by the definition of (3.14), we have $M = \lim_{k\to\infty} E(\alpha_k) = \lim_{k\to\infty} T(\alpha_k) = M_*$, which would imply part (i). Therefore, for the proof of part (i), it suffices to show that $\mathrm{dist}_k \to 0$. It is easy to see from (3.8) that

$$\begin{aligned}\|T(\alpha_k)\| &\le M_* + \left\| \sum_{1\le i<j\le m} P_j\big(\nabla f_i(x_1^k) - \nabla f_i(x^*)\big) \right\| \\ &= M_* + \left\| \sum_{1\le i<j\le m} P_j P_i (x_1^k - x^*) \right\| \\ &\le M_* + h_1 \mathrm{dist}_k \qquad (3.18)\end{aligned}$$

for a positive constant h_1 that depends only on $\{L_i\}_{i=1}^m$, where we used the triangle inequality and the fact that $\|P_i\| \le L_i$ in the last step. Then, from (3.17),

$$\mathrm{dist}_{k+1} \le \left(1 - \frac{Rc}{k^s} + \frac{R^2 h_1}{k^{2s}} + \mathcal{O}\left(\frac{1}{k^{3s}}\right)\right)\mathrm{dist}_k + \frac{R^2 M_*}{k^{2s}} + \mathcal{O}\left(\frac{1}{k^{3s}}\right).$$

Finally, applying Lemma 2.1 with a choice of $0 < a < Rc$, $t = s$, and $d > R^2 M_*$ and letting $a \to Rc$ and $d \to R^2 M_*$ shows the rate estimates for dist_k given in parts (ii) and (iii). In particular, these rate estimates also imply that $\mathrm{dist}_k \to 0$ as desired and proves part (i). To prove part (iv), assume $s = 0$ and $R \le \frac{1}{\|P\|}$. Then, the step size $\alpha_k = \alpha = R$ is a constant and by (3.11), for all $k \ge 1$,

$$\mathrm{dist}_{k+1} \le (1 - \alpha c)\mathrm{dist}_k + \alpha^2 M_\infty.$$

From this relation, by induction we obtain, for all $k \ge 1$,

$$\mathrm{dist}_{k+1} \le (1 - c\alpha)^k \mathrm{dist}_1 + \alpha^2 M_\infty \sum_{j=0}^{k-1} (1 - c\alpha)^j.$$

As the geometric sum satisfies $\sum_{j=0}^{k-1}(1-c\alpha)^j \le \frac{1}{c\alpha}$ for all $k \ge 1$, this proves part (iv). □

Remark 3.3. Part (ii) of Theorem 3.2 shows that for the step-size rule $\alpha_k = R/k$ with $R > 1/c$, we have $\mathrm{dist}_k = \mathcal{O}(1/k)$. Choosing R to satisfy this inequality requires the estimation or the knowledge of a lower bound for c. The following example illustrates that the convergence can be slower when R is not properly adjusted to c and the necessary condition $R > 1/c$ in our analysis cannot be omitted to achieve the rate results we report in Theorem 3.2. Similar issues with $1/k$-decay step sizes are also widely noted in the analysis of the stochastic gradient descent method in the stochastic approximation literature; see, e.g., [1, 12, 27, 33].

Example 3.4. Let $f_i(x) = x^2/20$ for $i = 1, 2$, $x \in \mathbb{R}$. Then, we have $m = 2$, $c = 1/5$, and $x^* = 0$. Take $R = 1$, which corresponds to the step size $1/k$. The IG iterations are

$$x_1^{k+1} = \left(1 - \frac{1}{10k}\right)^2 x_1^k.$$

If $x_1^1 = 1$, a simple analysis similar to [33] shows $x_1^k = \mathrm{dist}_k > \Omega(\frac{1}{k^{1/5}})$.

Remark 3.5. Under the setting of Theorem 3.2, in the special case when x^* is a global minimizer for each of the component functions, we have $\nabla f_i(x^*) = 0$ for each $i = 1, 2, \ldots, m$. This implies that $M = 0$ and for step size R/k with $R > 1/c$, by Theorem 3.2, we have $\limsup_{k\to\infty} k\mathrm{dist}_k = 0$, i.e., $\mathrm{dist}_k = o(1/k)$. In this special case, this rate result obtained from Theorem 3.2 can be refined further with an alternative analysis as follows: we can assume without loss of generality that $x^* = 0$ (the more general case can be treated similarly by shifting the coordinates and considering the functions $f_i(x - x^*)$ and $f(x - x^*)$). Then, this implies that $q_i = 0$ for all i, and therefore from (3.5) we have

$$\begin{aligned}\mathrm{dist}_{j+1} = \left\| \prod_{i=1}^{m} \left(I_n - \frac{R}{k} P_i \right) \right\| \mathrm{dist}_j &= \left\| I_n - \frac{R}{j} P + \mathcal{O}(1/j^2) \right\| \mathrm{dist}_j \\ &\leq \left(1 - \frac{Rc}{j} + \mathcal{O}(1/j^2)\right) \mathrm{dist}_j \leq \left(1 - \frac{\delta}{j}\right) \mathrm{dist}_j,\end{aligned}$$

where the last inequality holds for any $1 < \delta < Rc$ and j large enough. As

$$\prod_{j=2}^{k} \left(1 - \frac{\delta}{j}\right) \approx \prod_{j=2}^{k} \left(1 - \frac{1}{j}\right)^{\delta} = 1/k^{\delta},$$

it follows that $\mathrm{dist}_k = \mathcal{O}(1/k^{\delta})$ for any $1 < \delta < Rc$. This estimate on dist_k is more precise than $\mathrm{dist}_k = o(1/k)$, which we obtained from Theorem 3.2.

Remark 3.6. The constant M appearing in the upper bounds in Theorem 3.2 depend on the Lipschitz constant L and the number of component functions m. To illustrate this dependency, consider the simple example (also studied in [4, Example 1.5.6]) with

$$f_i(x) = \begin{cases} F_1(x) := \frac{(x-1)^2}{2}, & i = 1, 2, \ldots, \frac{m}{2}, \\ F_2(x) := \frac{(x+1)^2}{2}, & i = \frac{m}{2} + 1, \frac{m}{2} + 2, \ldots, m, \end{cases} \tag{3.19}$$

where $m \geq 2$ is even. This is a least square problem consisting of $m/2$ copies of two functions $(x-1)^2/2$ and $(x+1)^2/2$ in dimension one with the Hessian matrices $P_j = 1$

for all j, $\|P\| = L = c = m$, $x^* = 0$, and $\nabla f_i(x^*) = -1$ if $i \leq m/2$ and $\nabla f_i(x^*) = +1$ if $i > m/2$. After a straightforward computation using the formula (3.15), we obtain

$$M = \frac{1}{4} Lm. \tag{3.20}$$

Taking the limit superior of both sides of (3.11) would lead to

$$\limsup_{k\to\infty} \operatorname{dist}_k \leq \alpha \frac{M}{c} = \alpha \frac{m}{4} \quad \text{if} \quad \alpha \leq \frac{1}{\|P\|} = \frac{1}{m}, \tag{3.21}$$

where we plugged in (3.20) for M. For this particular example, Bertsekas [4, Example 1.5.6] showed that IG iterates converge to a limit cycle satisfying $\lim_{k\to\infty} \operatorname{dist}_k = \alpha\frac{m}{4}$, which matches our upper bound in (3.21). This shows that our analysis is tight for some quadratic functions and one would not be able to remove the m dependency in the constant M in general because of such worst-case examples. Studying the cycle gradient errors e^k defined in (3.31) for this example would also explain why the factor m in the upper bound for e^k is needed.[5] We also note that if the functions are selected according to a permutation Γ of $\{1, 2, \dots, m\}$ instead of the deterministic cyclic order, then Theorem 3.2 yields performance bounds with a constant M that will depend on Γ, highlighting the performance with respect to the specific order Γ chosen. For instance, for this particular example, if the odd numbered component functions are processed first, and then the even numbered functions are processed second, this would correspond to the choice of $\Gamma = \{1, 3, \dots, m-1, 2, 4, \dots, m\}$ in which case the constant M defined in Theorem 3.2 will become

$$M = \left\| \sum_{1\leq i<j\leq m} P_{\Gamma(j)} \nabla f_{\Gamma(i)}(x^*) \right\| = \frac{L}{2}. \tag{3.22}$$

Note that this constant is $O(m)$ times smaller than the constant in (3.20) obtained for the standard cyclic order $\Gamma_* := \{1, 2, \dots, m\}$. In fact, it can be shown that the standard cyclic order Γ_* corresponds to the worst-case scenario that maximizes M in (3.22) over all choices of permutations Γ. This is inline with the well-known fact that the performance of IG is quite sensitive to the choice of σ in practice (see, e.g., [40]). To our knowledge, our analysis is the first that can give performance bounds for IG that highlights the dependency on the order chosen. This is in contrast with the stochastic gradient descent (SGD) method, which is a stochastic variant of IG that samples the component functions randomly with replacement. Because SGD randomizes the choice of the index i over $\{1, 2, \dots, m\}$, the running time of SGD will be the same over this example if we remove the duplicate functions in (3.19) and minimize the objective $F_1(x) + F_2(x)$ instead with two component functions. Therefore, for this example, we see that the running time of SGD is independent of m in expectation whereas our upper bounds for IG will grow linearly with m. Beyond this example, under the assumptions of this paper, SGD with a decaying

[5] We note that similar scaling in m of the upper bounds for the distance to the optimizer and the cycle gradient error also arises in the study of some other deterministic methods such as the incremental aggregated gradient (IAG) method with strongly convex objectives and its proximal variants (see, e.g., [18, equation (3.11)] and [47]). On a related note, if the component functions are selected randomly without replacement instead, we show in work subsequent to this paper that one can improve the upper bounds by a factor of $O(m)$ if we consider the expected distance of the iterates to the optimizer (which is a weaker notion of convergence than the deterministic convergence considered in this paper); see [17, Remark 1].

step size has a slower convergence rate of $O(1/k)$ in function values compared to the $O(1/k^{2s})$ rate of IG; however, for SGD the constants appearing in the convergence rate are independent of m (see, e.g., [11, 27]) whereas for IG the constants depend linearly on m. Therefore, when m is very large and k is small to moderate, we expect SGD to perform better than IG (although SGD guarantees are in a weaker notion of convergence (in expectation) as opposed to deterministic convergence guarantees for IG), which is also inline with practice [4, 6]. A similar observation can be made for the RR method (a stochastic variant of IG where the component functions are sampled randomly without replacement) showing that RR converges at a rate $O(1/k^{2s})$ in expectation for $0 < s < 1$ in such a way that the constant in front of $1/k^{2s}$ does not depend on m (see [17, Example 3.2 and Theorem 2]).

3.2. Rate for smooth component functions. In addition to Assumption 3.1 on the strong convexity of f, we adopt the following assumptions that have appeared in a number of papers in the literature for analyzing incremental methods including [2, 16, 26].

Assumption 3.7. The functions f_i are twice continuously differentiable on $\mathbb{R}^n$ for each $i = 1, 2, \dots, m$.

Assumption 3.8. The iterates $\{x_1^k, x_2^k, \dots, x_m^k\}_{k\geq 1}$ are uniformly bounded, i.e., there exists a nonempty compact Euclidean ball $\mathcal{X} \subset \mathbb{R}^n$ that contains all the iterates.[6]

A consequence of these two assumptions is that the first and second derivatives of f on the compact set $\mathcal{X}$ are continuous, and hence are bounded. In other words, there exists a constant G such that

$$\max_{1\leq i\leq m} \sup_{x\in\mathcal{X}} \|\nabla f_i(x)\| \leq G \tag{3.23}$$

and there exists constants $L_i := \max_{z\in\mathcal{X}} \|\nabla^2 f_i(z)\| \geq 0$ such that

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\| \quad \text{for all} \quad x \in \mathcal{X}, \quad i = 1, 2, \dots, m. \tag{3.24}$$

From the triangle inequality, f has also Lipschitz gradients on $\mathcal{X}$ with constant

$$L = \sum_{i=1}^m L_i. \tag{3.25}$$

Another consequence is that an optimal solution to the problem (1.1), which we denote by x^*, exists and is unique by the strong convexity of f. Furthermore, these two assumptions are sufficient for global convergence of both the incremental Newton and the incremental gradient methods to x^* (see [4, 16] for a more general convergence theory). In this paper, we are interested in the rate of convergence.

3.2.1. Analyzing IG as a gradient descent with errors. We can rewrite the inner iterations (1.2) more compactly as

$$x_1^{k+1} = x_1^k - \alpha_k\big(\nabla f(x_1^k) - e^k\big), \quad k \geq 1, \quad i = 1, 2, \dots, m, \tag{3.26}$$

where the term

$$e^k = \sum_{i=1}^m \big(\nabla f_i(x_1^k) - \nabla f_i(x_i^k)\big) \tag{3.27}$$

[6] We note that Assumption 3.8 is not restrictive, because for strongly convex f, it can be shown (see [44, 46]) that the iterates are contained in a bounded set of the form $\{x : f(x) \leq \rho_1\} + \{x : \|x\| \leq \rho_2\}$ for some $\rho_2 > 0$ and $\rho_1 > f(x_1^1)$.

can be viewed as the gradient error. If Assumption 3.7 holds, we can substitute

$$\nabla f(x_1^k) = A_k(x_1^k - x^*)$$

into (3.26), where $A_k = \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))d\tau$ is the average of the Hessian matrices on the line segment $[x_1^k, x^*]$, to obtain

$$x_1^{k+1} - x^* = (I_n - \alpha_k A_k)(x_1^k - x^*) + \alpha_k e^k, \quad k \geq 1, \quad i = 1, 2, \ldots, m. \tag{3.28}$$

Taking norms of both sides, this implies that

$$\mathrm{dist}_{k+1} \leq \|I_n - \alpha_k A_k\| \mathrm{dist}_k + \alpha_k \|e^k\|. \tag{3.29}$$

These relations show that the evolution of the distance to the optimal solution is controlled by the decay of the step size α_k and the gradient error $\|e^k\|$. This motivates deriving tight upper bounds for the gradient error. Note also that under Assumptions 3.1, 3.7, and 3.8, the Hessian of f and the averaged Hessian matrix A_k admit the bounds

$$cI_n \preceq \nabla^2 f(x), \quad A_k \preceq LI_n, \quad x \in \mathcal{X} \tag{3.30}$$

(see also (3.25)). The gradient error consists of the difference of gradients evaluated at different inner steps (see (3.27)). This error can be controlled by the Lipschitz-ness of the gradients as follows: for any $k \geq 1$,

$$\begin{aligned}\|e^k\| &\leq \sum_{i=2}^m L_i \|x_1^k - x_i^k\| \leq \sum_{i=2}^m L_i \sum_{j=1}^{i-1} \|x_j^k - x_{j+1}^k\| \\ &\leq \sum_{i=2}^m L_i \alpha_k \sum_{j=1}^{i-1} \|\nabla f_j(x_j^k)\| \\ &\leq \alpha_k \widetilde{M}. \end{aligned} \tag{3.31}$$

Here,

$$\widetilde{M} := LGm, \tag{3.32}$$

where L is a Lipschitz constant for the gradient of f as in (3.25) and G is an upper bound on the gradients as in (3.23). Finally, plugging this into (3.29) and using the bounds (3.30) on the eigenvalues of A_k,

$$\begin{aligned}\mathrm{dist}_{k+1} &\leq \max(\|1 - \alpha_k c\|, \|1 - \alpha_k L\|)\mathrm{dist}_k + \alpha_k^2 \, \widetilde{M} \\ &\leq (1 - \alpha_k c)\mathrm{dist}_k + \alpha_k^2 \, \widetilde{M} \quad \text{if} \quad \alpha_k L \leq 1. \end{aligned} \tag{3.33}$$

This is the analogue of the recursion (3.12) obtained for quadratics with the only difference that the constants M_∞ and $\|P\|$ are replaced by their analogues $\widetilde{M}$ and L, respectively. Then, a reasoning along the lines of the proof of Theorem 3.2 yields the following convergence result, which generalizes Theorem 3.2 from quadratic functions to smooth functions just by modifying the constants properly (by replacing M and M_∞ with $\widetilde{M}$ and replacing $\|P\|$ with L). We skip the proof for the sake of brevity.

THEOREM 3.9. *Let $f_i(x) : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2, \ldots, m$, be component functions satisfying Assumptions* 3.1 *and* 3.7*. Consider the iterates $\{x_1^k, x_2^k, \ldots, x_m^k\}_{k \geq 1}$ obtained by the IG iterations* (1.2) *with a decaying step size $\alpha_k = R/k^s$, where $R > 0$ and $s \in [0, 1]$. Suppose that Assumption* 3.8 *is also satisfied. Then, we have the following.*

(i) *If* $s = 1$, *then*

$$\begin{aligned}
&\limsup_{k\to\infty} k\,\mathrm{dist}_k &&\le \frac{R^2\,\widetilde{M}}{Rc-1} && \textit{for} \quad R > 1/c,\\
&\limsup_{k\to\infty} \frac{k}{\log k}\mathrm{dist}_k &&< \infty && \textit{for} \quad R = 1/c,\\
&\limsup_{k\to\infty} k^{Rc}\mathrm{dist}_k &&< \infty && \textit{for} \quad R < 1/c,
\end{aligned}$$

where $\widetilde{M}$ *is given by* (3.32).

(ii) *If* $0 < s < 1$, *then*

$$\limsup_{k\to\infty} k^s \mathrm{dist}_k \le \frac{R\,\widetilde{M}}{c},$$

where $\widetilde{M}$ *is given by* (3.32).

(iii) *If* $s = 0$ *and* $R \le \frac{1}{L}$, *then*

$$\mathrm{dist}_{k+1} \le (1-c\alpha)^k \mathrm{dist}_1 + \frac{\alpha\,\widetilde{M}}{c} \quad \forall k \ge 1, \tag{3.34}$$

where the step size $\alpha = \alpha_k = R$ *is a constant and* $\widetilde{M}$ *is given by* (3.32).

Remark 3.10. Under the conditions of Theorem 3.2, the quadratic functions f_i have Lipschitz continuous gradients with constants $L_i = \|P_i\|$. Thus, for $0 < s \le 1$,

$$M \le \sum_{1\le i\le m} \sum_{j=i+1}^{m} L_j \,\|\nabla f_i(x^*)\| \le \sum_{1\le i\le m} L\,\|\nabla f_i(x^*)\| \le \widetilde{M} = LGm$$

by the definitions of L and $\widetilde{M}$ from (3.25) and (3.32), where M satisfies (3.15). This shows how the constants M and $\widetilde{M}$ that arise in Theorems 3.2 and 3.9 are related. In particular, the upper bounds obtained are smaller in the quadratic case.

Under a strong convexity-type condition and subgradient boundedness, Nedić and Bertsekas consider the IG method with constant step size and show that when f_i are convex but not necessarily smooth or differentiable, for any given $\varepsilon > 0$, it suffices to have $\mathcal{O}(\log(\frac{1}{\varepsilon})/\varepsilon^2)$ cycles of IG for convergence to the ε-neighborhood $\{x \in \mathbb{R}^n : \|x - x^*\| \le \varepsilon\}$ of an optimal solution x^* [28, Proposition 2.4]. The following corollary of Theorem 3.9 shows that this result can be improved to $\mathcal{O}(\log(\frac{1}{\varepsilon})/\varepsilon)$ when the component functions are smooth and strongly convex.

Corollary 3.11. *Let* $f_i(x) : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2, \ldots, m$, *be component functions satisfying Assumptions* 3.1 *and* 3.7. *Consider the iterates* $\{x_1^k, x_2^k, \ldots, x_m^k\}_{k\ge1}$ *obtained by the IG iterations* (1.2) *with constant step size* $\alpha = \varepsilon c/(2\,\widetilde{M})$, *where* $\widetilde{M}$ *is defined by* (3.32). *Suppose that Assumption* 3.8 *is also satisfied. Then, IG requires at most*

$$\mathcal{O}\left(\frac{\widetilde{M}}{c^2}\frac{\log(1/\varepsilon)}{\varepsilon}\right) \tag{3.35}$$

cycles to guarantee convergence to an ε*-neighborhood of the optimal solution* x^*.

Proof. Given such $\varepsilon > 0$ and step size α, we note that $c\alpha < 1$ and $\alpha\, \widetilde{M}/c = \varepsilon/2$. Furthermore, by Theorem 3.9, the inequality (3.16) holds with M_∞ replaced by $\widetilde{M}$. Therefore, there exists a constant K such that

$$(1-c\alpha)^k \mathrm{dist}_1 \leq \exp(-c\alpha k)\mathrm{dist}_1 < \frac{\varepsilon}{2} \quad \forall k \geq K, \tag{3.36}$$

so that $\mathrm{dist}_{k+1} < \varepsilon$ for all $k \geq K$, i.e., the iterates lie inside an ε-neighborhood of the optimizer after K cycles. By taking the log of both sides in (3.36) and using $\log(1-z) \approx z$ for z around zero, straightforward calculations show that this condition is satisfied for K satisfying (3.35). ∎

Remark 3.12. When $s = 0$, the step size $\alpha_k = \alpha$ is a constant and there exists simple examples (with two quadratics in dimension one) which necessitate $\Omega\big(\log(1/\varepsilon)/\varepsilon\big)$ cycles to reach to an ε-neighborhood of an optimal solution (see [23, Proposition 2.2]). Therefore, it can be argued that the dependency on ε of the iteration complexity in Corollary 3.11 cannot be improved further when we restrict ourselves to smooth and strongly convex functions.

For decaying step sizes of the form $\alpha_k = R/k^s$ with $s \in (0,1]$, Theorem 3.9 shows that $\mathrm{dist}_k = \mathcal{O}(1/k^s)$ (as long as $R > 1/c$ when $s = 1$). The next lemma shows that there exist quadratic examples that satisfy $\mathrm{dist}_k = \Omega(1/k^s)$. This shows that our convergence analysis for the step-size rule $\alpha_k = R/k^s$ in Theorem 3.9 is tight in the sense that the exponent s in the sublinear convergence rate $1/k^s$ cannot be improved.[7] In fact, the simple example given in the proof of part (ii) of Theorem 3.14 provides a lower bound for $s \in (0,1]$ via an analysis similar to the proof of part (ii) of Theorem 3.14, which leads to the following lemma.

LEMMA 3.13. *Consider the iterates $\{x_1^k\}$ generated by the IG method with decaying step size $\alpha = R/k^s$, where $s \in (0,1]$. There exist quadratic component functions $\{f_i\}_{i=1}^m$ such that the sum function f is strongly convex and the resulting IG iterates satisfy* $\mathrm{dist}_k = \Omega(1/k^s)$.

This lower bound in distances to the optimal solution in Lemma 3.13 is based on an example in dimension one. However, one can also construct similar examples in higher dimensions. In Appendix B, we provide an alternative example in dimension two to illustrate this fact.

3.3. Comparison of IG with GD, SGD, and RR.

We recall that the convergence rates we provide for IG methods hide constants that depend linearly on m. This is expected as these results are worst-case deterministic bounds that are applicable to any order to process the component functions. Furthermore, Remark 3.6 shows that this dependency on m is not improvable in general for IG methods. This is in contrast to SGD and RR methods which sample the functions randomly admitting $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ asymptotic convergence guarantees in expectation with leading constants that are independent of m [17, 36]. The GD method on the other hand enjoys linear convergence for strongly convex objectives. However, its running time also scales with m linearly as computing the gradient of the objective f requires computing the gradients of all the m component functions. Summarizing these observations, the performance of IG, GD, SGD, and RR methods depends on the value of m and the target accuracy ε required in function values to solve problem (1.1). In particular, if

[7] We note that for the special case in which $s \in (1/2, 1]$, Luo gives a first heuristic analysis which suggests that one would expect to have $\mathrm{dist}_k = \Omega(1/k^s)$ for least square problems (see [23, Remark 2, after the proof of Theorem 3.1]).

the target accuracy ε is extremely small and m is small to moderate, GD or IG can clearly outperform SGD. On the other hand, for applications where m is very large and low-to-medium target accuracy ε is enough, SGD and RR can be better than GD or IG.

If we compare the $\mathcal{O}(m/k^2)$ performance of IG to $\mathcal{O}(1/k)$ performance of SGD, for small m IG can outperform SGD, but when m is large (which is the more interesting scenario for the problem (1.1)), SGD will typically perform better than IG for moderate values of k.

3.4. Lower bounds. Consider the following set of quadratic functions which are strongly convex with parameter c and have Lipschitz gradients with constant L:

$$\mathcal{C}_{c,L} = \bigcup_{n=1}^{\infty} \left\{ \bar{f}(x) = \frac{1}{2} x^T P x - q^T x + r \;\middle|\; P \text{ symmetric, } cI_n \preceq P \preceq LI_n; x, q \in \mathbb{R}^n; r \in \mathbb{R} \right\}.$$

Theorem 3.2 and Remark 3.10 shows that when IG is applied to quadratic functions $\bar{f}_i : \mathbb{R}^n \to \mathbb{R}$ with a sum $\bar{f} \in \mathcal{C}_{c,L}$ using a step size $\alpha_k = R/k$, where $R > 1/c$, it results in

$$\limsup_{k\to\infty} k\text{dist}_k \le \frac{M}{Rc-1} \le \frac{\widetilde{M}}{Rc-1} = \frac{LGm}{Rc-1}.$$

In other words, $\text{dist}_k = \mathcal{O}(1/k)$. A natural question is whether one could improve this rate by choosing an alternative step size. For instance, would it be possible to obtain $\text{dist}_k = o(1/k)$ uniformly (for every m and $\{\bar{f}_i\}_{i=1}^m$ with a sum $\bar{f} \in \mathcal{C}_{c,L}$)? The next result gives a negative answer to this question, showing that no matter which fixed deterministic step size we choose (that depends only on the problem parameters c, L, and iteration number k), there exist simple quadratic functions that result in iterates $\{x_1^k\}$ satisfying $\text{dist}_k \ge \tilde{\Omega}(1/k)$. The proof is based on explicit construction of some quadratic examples.

THEOREM 3.14. *Consider the following IG iterations applied to quadratic component functions $\bar{f}_i : \mathbb{R}^n \to \mathbb{R}$, where $\bar{f} = \left(\sum_{i=1}^m \bar{f}_i\right) \in \mathcal{C}_{c,L}$:*

$$x_{i+1}^k = x_i^k - \sigma(c, L, k)\nabla \bar{f}_i(x_i^k), \quad k \ge 1, \quad i = 1, 2, \dots, m,$$

where the step-size sequence $\alpha_k = \sigma(c, L, k) : \mathbb{R}_+^3 \to \mathbb{R}_+$ is a fixed deterministic sequence where the function σ determines the step size and depends only on c, L, and k. Suppose that for every choice of m, n and such $\{\bar{f}_i : \mathbb{R}^n \to \mathbb{R}\}_{i=1}^m$, we have

$$\limsup_{k\to\infty} k\text{dist}_k \le \bar{b} \tag{3.37}$$

for some $\bar{b} > 0$ which depends only on L, G, m, c and the function $\sigma = \sigma(c, L, k)$ defined above. Then, the following statements are true.

(i) *The step-size sequence satisfies $\limsup_{k\to\infty} k\alpha_k \ge \underline{b}$, where $\underline{b} = \frac{1}{2L}$.*

(ii) *There exist positive integers $\tilde{m}$, $\tilde{n}$ and functions $\{\tilde{f}_i : \mathbb{R}^{\tilde{n}} \to \mathbb{R}\}_{i=1}^{\tilde{m}}$ such that $\tilde{f} = \left(\sum_{i=1}^{\tilde{m}} \tilde{f}_i\right) \in \mathcal{C}_{c,L}$ and the iterates $\{x_1^k\}$ generated by the IG iterations* (1.3) *with component functions $f_i = \tilde{f}_i$ satisfy*

$$\text{dist}_k \ge \tilde{\Omega}(1/k).$$

Proof.

(i) We follow a similar approach to the analysis in [38, Appendix A]. Consider the simple example $\bar{f}(x) = \bar{f}_1(x) = \frac{L}{2}x^2 \in \mathcal{C}_{c,L}$ with only one component function in dimension one ($m = n = 1$) or a similar alternative example $\bar{f}(x) = \frac{L}{2}\big(x(1)^2 + x(2)^2\big) \in \mathcal{C}_{c,L}$ with two component functions $\bar{f}_1(x) = \frac{L}{2}x(1)^2$ and $\bar{f}_2(x) = \frac{L}{2}x(2)^2$ in dimension two ($n = m = 2$), where $x(\ell)$ denotes the ℓth coordinate of the vector x. In any of these two examples, IG becomes the classical gradient descent (GD) method leading to the iterations $x_1^{k+1} = \prod_{j=1}^k (1-\alpha_j L)x_1^1$. Since $x^* = 0$, this implies

$$\mathrm{dist}_{k+1} = \left|\prod_{j=1}^{k}(1-\alpha_j L)\right| \mathrm{dist}_1.$$

By assumption (3.37), we need at least

$$\left|\prod_{j=1}^{k}(1-\alpha_j L)\right| \le \frac{\bar{b}}{k} + o\left(\frac{1}{k}\right) \le \frac{2\bar{b}}{k} \quad \text{for } k \text{ large} \tag{3.38}$$

and $\alpha_k \to 0$ (otherwise simple examples show the global convergence may not be obtained from an arbitrary initial point). By taking the natural logarithm of both sides, this is equivalent to requiring

$$\sum_{j=1}^{k} -\ln|1-\alpha_j L| \ge \log k - \log(2\bar{b}) \quad \text{for } k \text{ large.}$$

Using $2z \ge -\ln(1-z)$ for $0 \le z \le \frac{1}{2}$, it follows that

$$\sum_{j=1}^{k} \alpha_j \ge \underline{b} \log k - \frac{\log(2\bar{b})}{2L} \tag{3.39}$$

when k is large enough with $\underline{b} = \frac{1}{2L}$. Assume there exists δ such that $\limsup_{k\to\infty} k\alpha_k < \delta < \underline{b} = \frac{1}{2L}$. Then, by definition of the limit superior, we have $\alpha_k \le \frac{\delta}{k}$ for any k large enough. By summing this inequality over the iterations k, we obtain $\sum_{j=1}^k \alpha_j \le \delta \log(k) + b_2$ for a constant b_2 and for any k large enough. By (3.39), we also have $\delta \ge \underline{b}$. This contradicts our earlier assumption that $\delta < \underline{b}$. Therefore, no such δ exists, i.e., $\limsup_{k\to\infty} k\alpha_k \ge \underline{b}$. This completes the proof.

(ii) Consider the following simple example with two quadratics $\bar{f} = \bar{f}_1 + \bar{f}_2$ with $\bar{f}_1(x) = \frac{L}{2}(x-1)^2$ and $\bar{f}_2(x) = \frac{L}{2}(x+1)^2$ in dimension one ($m = 2$, $n = 1$). Then, applying IG with an initial point $x_1^1 \in \mathbb{R}$ results in the iterates $\{x_1^k, x_2^k\}$ with

$$x_2^k = x_1^k - \bar{\alpha}_k(x_1^k - 1), \tag{3.40}$$

$$x_1^{k+1} = (1-\bar{\alpha}_k)^2 x_1^k - (\bar{\alpha}_k)^2, \tag{3.41}$$

$$x_2^{k+1} = (1-\bar{\alpha}_k)^2 x_2^k + (\bar{\alpha}_k)^2, \tag{3.42}$$

where $\bar{\alpha}_k = \alpha_k L$ is the normalized step size. Define $y^k = x_1^k + x_2^k$. By summing up (3.41) and (3.42), we see that

$$y^{k+1} = (1-\bar{\alpha}_k)^2 y^k = \prod_{j=1}^{k}(1-\bar{\alpha}_j)^2 y^1. \tag{3.43}$$

By the necessary condition (3.38), we also have

$$0 \leq |y^k| \leq \mathcal{O}(1/k^2). \tag{3.44}$$

Finally, plugging $y^k = x_1^k + x_2^k$ into (3.40), we obtain

$$x_1^k = \frac{y^k}{2} + \bar{\alpha}_k \frac{(x_1^k - 1)}{2}.$$

As $\alpha_k = \tilde{\Omega}(1/k)$ by part (i) and x_1^k is converging to $x^* = 0$, it follows from (3.44) and the triangle inequality that

$$|x_1^k| = \text{dist}_k \geq \bar{\alpha}_k \frac{|x_1^k - 1|}{2} - \frac{|y^k|}{2} = \tilde{\Omega}(1/k).$$

This completes the proof. □

4. Convergence rate analysis for IN.

To analyze the gradient errors introduced in the IN iterations (1.4) and (1.5), we rewrite the outer IN iterations using [16, equation (2.12)] as

$$x_1^{k+1} = x_1^k - \alpha_k (\bar{H}_m^k)^{-1} (\nabla f(x_1^k) + e_g^k), \tag{4.1}$$

where

$$e_g^k = \sum_{j=1}^m \left(\nabla f_j(x_j^k) - \nabla f_j(x_1^k) + \frac{1}{\alpha_k k} \nabla^2 f_j(x_j^k)(x_1^k - x_j^k) \right) \tag{4.2}$$

is the gradient error and

$$\bar{H}_m^k = \frac{H_0^1 + \sum_{i=1}^k \sum_{j=1}^m \nabla^2 f_j(x_j^i)}{k} = \frac{\sum_{i=1}^k \nabla^2 f(x_1^i)}{k} + e_h^k \tag{4.3}$$

is an averaged Hessian up to an error term

$$e_h^k := \frac{H_0^1 + \sum_{i=1}^k \sum_{j=1}^m \left(\nabla^2 f_j(x_j^i) - \nabla^2 f_j(x_1^i) \right)}{k}. \tag{4.4}$$

We let $\alpha_k = R/k$ and introduce the norm

$$\|z\|_* := \left(z^T H_* z \right)^{1/2}, \quad z \in \mathbb{R}^n, \quad \text{where} \quad H_* := \nabla^2 f(x^*), \tag{4.5}$$

which is a norm that arises frequently in the analysis of self-concordant functions and Newton's method [34]. The next theorem shows that unlike IG, IN can achieve the $\mathcal{O}(1/k)$ rate without requiring knowing or estimating the strong convexity constant of f. Furthermore, the constants arising in IN when considered in the $*$-norm do not have the Lipschitz constant L unlike the previous rate estimates we obtained for IG in the Euclidean norm.

THEOREM 4.1. *Let f_i be convex component functions on $\mathbb{R}^n$ satisfying Assumptions* 3.1 *and* 3.7 *for $i = 1, 2, \ldots, m$. Consider the iterates $\{x_1^k, \ldots, x_m^k\}$ generated by the IN method with step size $\alpha_k = R/k$, where $R > 1$. Suppose also that Assumption* 3.8 *holds. Then, we have*

$$\limsup_{k \to \infty} k \|x_1^k - x_*\|_* \leq \frac{BR(R+1)}{R-1}, \tag{4.6}$$

where $\|\cdot\|_*$ and H_* are defined by (4.5) *and $B = \sum_{i=1}^m \|H_*^{-1/2} \nabla f_i(x^*)\| \leq G/\sqrt{c}$, where G is defined by* (3.23).

The proof of this theorem is given in Appendix A. The main idea is to change variables $y = H_*^{1/2}x$ and analyze the corresponding iterates $y_1^k = H_*^{1/2}x_1^k$. By this change of variables, it can be shown that $\{y_1^k\}$ follows a similar recursion to the IN iterates $\{x_1^k\}$ converging to $y_* = H_*^{1/2}x_*$. Then, one can analyze how fast the sequence $\|y_1^k - y_*\| = \|x_1^k - x_*\|_*$ decays to zero by exploiting the fact that y-coordinates have the advantage that the local strong convexity constant and the local Lipschitz constant of f around y_* are both equal to one due to the normalization obtained by this change of variable.

5. Conclusion.

We analyzed the convergence rate of the IG and IN algorithms when the component functions are smooth and the sum of the component functions is strongly convex. This covers the interesting case of many regression problems including the ℓ_2 regularized linear and nonlinear regression problems. For IG, we show that the distance of the iterates converges at rate $\mathcal{O}(1/k^s)$ to the optimal solution with a diminishing step size of the form $\alpha_k = \mathcal{O}(1/k^s)$ for $s \in (0,1]$. This improves the previously known $\mathcal{O}(1/\sqrt{k})$ rate (when $s \in (1/2,1]$) and translates into convergence at rate $\mathcal{O}(1/k^{2s})$ of the suboptimality of the function value. For constant step size, we also improve the existing iteration complexity results for IG from $\mathcal{O}(\frac{\log(1/\varepsilon)}{\varepsilon^2})$ to $\mathcal{O}(\frac{\log(1/\varepsilon)}{\varepsilon})$ to reach an ε-neighborhood of an optimal solution. In addition, we show that our analysis with this choice of step size is tight in the sense that the exponent s of the sublinear convergence rates cannot be improved.

Achieving the fastest $\mathcal{O}(1/k)$ rate in distances with IG for the step size $\alpha_k = R/k$ requires a good knowledge or approximation of the strong convexity constant of the sum function f in order to be able to tune the parameter R. However, we showed that IN as a second-order method can achieve this fast rate without the knowledge of the strong convexity constant. Furthermore, the results we obtain in this paper yield performance guarantees for IG when the component functions are selected with respect to any fixed permutation Γ of $\{1,2,\ldots,m\}$ (see Remark 3.10).

The RR method we analyze in a subsequent work [17] is based on selecting a random permutation Γ of $\{1,2,\ldots,m\}$ for each cycle where we develop convergence guarantees for the distance to the optimizer both in expectation and with probability one (with respect to random permutations encountered during the iterations). In expectation results in [17], we build on the main results of this paper and show after a detailed analysis that some terms appearing in the upper bounds for a fixed permutation Γ cancel out when we take expectation over all choices of Γ, leading to different constants in the upper bounds for the RR method compared to the IG method. The analysis in [17] in expectation also studies an alternative recursion for the dist_k sequence (as opposed to the recursions studied in Theorems 3.2 and 3.9) to get the best constants. Furthermore, since RR is a stochastic variant of IG, analyzing RR iterates with probability one requires several tools from stochastic approximation theory and probability theory such as martingale convergence theorems and concentration inequalities (see [17] for the details) in contrast to the deterministic proof techniques used in this work.

Appendix A. Proof of Theorem 4.1.

Proof. By a change of variable let $y = H_*^{1/2}x$ and define $\hat{f}(y) = f(x)$ for $y \in \mathbb{R}^n$. Consider the IN iterates in the y-coordinates. By the chain rule, we have

$$\nabla f(x) = H_*^{1/2}\nabla \hat{f}(y), \quad \nabla^2 f(x) = H_*^{1/2}\nabla^2 \hat{f}(y) H_*^{1/2}. \tag{A.1}$$

Using these identities, the IN iterations (1.4) and (1.5) become

$$y_{i+1}^k := y_i^k - \alpha_k (\bar{D}_i^k)^{-1} \nabla \hat{f}_i(y_i^k), \quad i = 1, 2, \ldots, m, \tag{A.2}$$

where $\bar{D}_i^k = D_i^k / k$ with

$$y_i^k = H_*^{1/2} x_i^k, \quad D_i^k := D_{i-1}^k + \nabla^2 \hat{f}_i(y_i^k) = H_*^{-1/2} H_i^k H_*^{-1/2}. \tag{A.3}$$

Furthermore, it is known that the IN method is globally convergent under these assumptions (see [16]), i.e., $x_1^k \to x^*$, although studying the rate of convergence is the subject of this theorem. More generally, due to the cyclic structure, we have also $x_i^k \to x^*$ for each $i = 1, 2, \ldots, m$. Then, from the Hessian update formula (1.5), it follows that $\bar{H}_i^k \to H_*$ for each $i = 1, 2, \ldots, m$ fixed and

$$\bar{D}_i^k \to \nabla^2 \hat{f}(y^*) = H_*^{-1/2} H_* H_*^{-1/2} = I_n, \quad i = 1, 2, \ldots, m, \tag{A.4}$$

where we used the second change of variable identity from (A.1) to calculate $\nabla^2 \hat{f}(y^*)$. Comparing the IN iterations (1.4) in the x-coordinates and the IN iterations (A.2) in the y-coordinates, we see that they have exactly the same form: the only differences are that in the latter the gradients and the Hessian matrices are taken with respect to y (instead of x) and f is replaced with $\hat{f}$. Therefore, inequalities (4.1) and (4.2) hold if we replace f with $\hat{f}$ and x_j^i with y_j^i, leading to

$$y_1^{k+1} = y_1^k - \alpha_k (\bar{D}_k)^{-1} (\nabla \hat{f}(y_1^k) + e_y^k), \quad \text{where} \quad \bar{D}_k := \bar{D}_m^k, \tag{A.5}$$

and the gradient error becomes

$$e_y^k = \sum_{j=1}^m \left(\nabla \hat{f}_j(y_j^k) - \nabla \hat{f}_j(y_1^k) + \frac{1}{R} \nabla^2 \hat{f}_j(y_j^k)(y_1^k - y_j^k) \right), \tag{A.6}$$

where we set $\alpha_k = R/k$. Setting $\nabla \hat{f}(y_1^k) = Y_k (y_1^k - y^*)$ in (A.5) with an averaged Hessian

$$Y_k = \int_0^1 \nabla^2 \hat{f}(y^* + \tau(y_1^k - y^*)) d\tau, \tag{A.7}$$

where $y^* = H_*^{1/2} x_*$, and using the triangle inequality we obtain

$$\|y_1^{k+1} - y^*\| \leq \underbrace{\left\| \left(I_n - \frac{R}{k} \bar{D}_k^{-1} Y_k \right) (y_1^k - y^*) \right\|}_{:= m_k} + \frac{R}{k} \underbrace{\|(\bar{D}_k)^{-1} e_y^k\|}_{:= n_k}. \tag{A.8}$$

The remainder of the proof consists of estimating the terms m_k and n_k on the right-hand side separately in the following three steps, and this will imply the desired convergence rate of the left-hand side $\|y_1^{k+1} - y^*\| = \|x_1^{k+1} - x^*\|_*$.

Step 1 (bounding m_k). We first observe that

$$m_k^2 = \left\| \left(I_n - \frac{R}{k} \bar{D}_k^{-1} Y_k \right) (y_1^k - y^*) \right\|^2 = (y_1^k - y^*)^T S_k (y_1^k - y^*), \tag{A.9}$$

where

$$S_k = I_n - \frac{R}{k} Z_k, \quad Z_k = Y_k \bar{D}_k^{-1} + \bar{D}_k^{-1} Y_k - \frac{R}{k} Y_k \bar{D}_k^{-2} Y_k. \tag{A.10}$$

From (A.4), we have $\bar{D}_k = \bar{D}_m^k \to I_n$. Furthermore, as y_1^k converges to y^* by the global convergence property of IN, Y_k defined in (A.7) converges to $\nabla_y^2 f(y^*) = I_n$ as well. Therefore, we have $Z_k = 2I_n + o(1)$ in (A.10), which leads to

$$S_k = \left(1 - \frac{2R}{k}\right) I_n + o\left(\frac{1}{k}\right).$$

Then, for every $\varepsilon \in (0,1)$, there exists a finite $k_1 = k_1(\varepsilon)$ such that for $k \geq k_1(\varepsilon)$,

$$S_k \preceq \left(1 - \frac{2R(1-\varepsilon)}{k}\right) I_n,$$

and therefore

$$m_k^2 = (y_1^k - y^*)^T S_k (y_1^k - y^*) \leq \left(1 - \frac{2R(1-\varepsilon)}{k}\right) \|y_1^k - y^*\|^2$$

for $k \geq k_1(\varepsilon)$. By taking the square roots of both sides, for $k \geq \max\{k_1, 2R\}$, we obtain

$$m_k \leq \left(1 - \frac{R(1-\varepsilon)}{k}\right) \|y_1^k - y^*\|, \tag{A.11}$$

where we used $(1-z)^{1/2} \leq 1 - z/2$ for $z \in [0,1]$ with $z = \frac{2R(1-\varepsilon)}{k}$.

Step 2 (bounding n_k). Similarly we can write

$$\nabla \hat{f}_j(y_j^k) - \nabla \hat{f}_j(y_1^k) = Y_{k,j}(y_j^k - y_1^k)$$

with an averaged Hessian satisfying

(A.12)

$$Y_{k,j} = \int_0^1 \nabla^2 \hat{f}_j(y_1^k + \tau(y_j^k - y_1^k)) d\tau \underset{k\to\infty}{\to} \nabla^2 \hat{f}_j(y^*) \preceq \sum_{i=1}^m \nabla^2 \hat{f}_i(y^*) = \nabla^2 \hat{f}(y^*) = I_n$$

as $k \to \infty$ for all $j = 1, 2, \ldots, m$, where we used (A.4) in the last equality and the fact that $\nabla^2 \hat{f}_i(y^*) \succeq 0$, which is implied by the convexity of f_i. Next, we decompose the gradient error term (A.6) into two parts as

$$e_y^k = e_{y,1}^k + e_{y,2}^k$$

with

$$e_{y,1}^k = \sum_{j=1}^m Y_{k,j}(y_j^k - y_1^k), \quad e_{y,2}^k = \frac{1}{R}\sum_{j=1}^m \nabla^2 \hat{f}_j(y_j^k)(y_1^k - y_j^k).$$

From the triangle inequality for n_k defined in (A.8), we have

$$n_k \leq \sum_{\ell=1}^2 n_{k,\ell} \quad \text{with} \quad n_{k,\ell} := \|\bar{D}_k^{-1} e_{y,\ell}^k\|. \tag{A.13}$$

We then estimate $n_{k,\ell}$ for $\ell = 1$ and $\ell = 2$:

$$\begin{aligned} n_{k,1} &= \left\| \sum_{j=1}^m \bar{D}_k^{-1} Y_{k,j}(y_j^k - y_1^k) \right\| \\ &= \frac{R}{k} \left\| \sum_{j=1}^m \sum_{\ell=1}^{j-1} \bar{D}_k^{-1} Y_{k,j} \left(\bar{D}_\ell^k\right)^{-1} \nabla \hat{f}_\ell(y_\ell^k) \right\|. \qquad \text{(A.14)} \end{aligned}$$

From (A.4) and (A.12), for every $\ell, j \in \{1, 2, \dots, m\}$, each summand above in the last equality satisfies

$$\lim_{k\to\infty} \bar{D}_k^{-1} Y_{k,j} \big(\bar{D}_\ell^k\big)^{-1} \nabla \hat{f}_\ell(y_\ell^k) = \nabla^2 \hat{f}_j(y^*) \nabla \hat{f}_\ell(y^*) \tag{A.15}$$

so that

$$\begin{aligned}\lim_{k\to\infty} k n_{k,1} &= R \Bigg\| \sum_{\ell=1}^m \sum_{j=\ell+1}^m \nabla^2 \hat{f}_j(y^*) \nabla \hat{f}_\ell(y^*) \Bigg\| \\ &\le R \sum_{\ell=1}^m \Bigg\| \sum_{j=\ell+1}^m \nabla^2 \hat{f}_j(y^*) \Bigg\| \|\nabla \hat{f}_\ell(y^*)\| \le R \sum_{\ell=1}^m \|\nabla^2 \hat{f}(y^*)\| \|\nabla \hat{f}_\ell(y^*)\| \\ &\le RB,\end{aligned}$$

where in the last step we used the fact that $\nabla^2 \hat{f}(y^*) = I_n$ and the change of variable formula (A.1) on gradients. Similarly,

$$\begin{aligned} n_{k,2} = \|\bar{D}_k^{-1} e_{y,2}^k\| = \frac{1}{R} \Bigg\| \sum_{j=1}^m \bar{D}_k^{-1} \nabla^2 \hat{f}_j(y_j^k)(y_j^k - y_1^k) \Bigg\| \\ = \frac{1}{k} \Bigg\| \sum_{j=1}^m \bar{D}_k^{-1} \nabla^2 \hat{f}_j(y_j^k) \sum_{\ell=1}^{j-1} (\bar{D}_\ell^k)^{-1} \nabla \hat{f}_\ell(y_\ell^k) \Bigg\|.\end{aligned}$$

Then, as $\nabla^2 \hat{f}_j(y_j^k) \to \nabla^2 \hat{f}_j(y^*)$, it follows similarly from (A.4) that

$$\lim_{k\to\infty} k n_{k,2} = \Bigg\| \sum_{j=1}^m \sum_{\ell=1}^{j-1} \nabla^2 \hat{f}(y^*) \nabla \hat{f}(y^*) \Bigg\| \le B. \tag{A.16}$$

Going back to the triangle inequality bound (A.13) on n_k, we arrive at

$$\begin{aligned}\lim_{k\to\infty} k n_k &\le \limsup_{k\to\infty} k n_{k,1} + \limsup_{k\to\infty} k n_{k,2} \\ &\le RB + B = (R+1)B.\end{aligned}$$

In other words, for any $\varepsilon > 0$, there exists $k_2 = k_2(\varepsilon)$ such that

$$n_k \le (1+\varepsilon)(R+1)B \frac{1}{k} \quad \forall k \ge k_2(\varepsilon). \tag{A.17}$$

Step 3 (deriving the rate). Let $\varepsilon \in (0, \frac{R-1}{2R})$ so that $R_\varepsilon := R(1-\varepsilon) > 1$. Then, it follows from (A.8), (A.11), and (A.17) that for $k \ge \max\{k_1(\varepsilon), 2R, k_2(\varepsilon)\}$,

$$\|y_1^{k+1} - y^*\| \le \left(1 - \frac{R_\varepsilon}{k}\right) \|y_1^k - y^*\| + \frac{(1+\varepsilon)BR(R+1)}{k^2}.$$

Applying Lemma 2.1 with $u_k = \|y_1^k - y^*\| = \|x_1^k - x^*\|_*$, $a = R_\varepsilon > 1$, and $s = 1$ leads to

$$\limsup_{k\to\infty} k \|x_1^k - x_*\|_* \le \frac{(1+\varepsilon)BR(R+1)}{R(1-\varepsilon) - 1}. \tag{A.18}$$

Letting $\varepsilon \to 0$ completes the proof. □

Appendix B. An example in dimension two with $\mathbf{dist}_{\boldsymbol{k}} = \boldsymbol{\Omega(1/k^s)}$. Our aim is to construct a set of component functions in dimension two such that if IG is applied with step size $\alpha_k = \Theta(1/k^s)$ with $0 < s \leq 1$, the resulting iterates satisfy $\text{dist}_k \geq \Omega(1/k^s)$.

Consider the following least squares example in dimension two ($n = 2$) with $m = 8$ quadratics:

$$\tilde{f}_i(x) = \frac{1}{2}(c_i^T x + 1)^2, \quad i = 1, 2, \dots, 8,$$

where the vectors $c_i \in \mathbb{R}^2$ are

$$c_1 = c_6 = -c_2 = -c_5 = [-1, 0]^T, \tag{B.1}$$

$$c_3 = c_8 = -c_4 = -c_7 = [0, -1]^T. \tag{B.2}$$

It is easy to check that the sum $\tilde{f} := \sum_{i=1}^{8} \tilde{f}_i$ is strongly convex as

$$\nabla^2 \tilde{f}(x) = \sum_{i=1}^{8} c_i c_i^T = 4 I_n \succ 0. \tag{B.3}$$

Starting from an initial point $\tilde{x}_1^1$, the IG method with step size α_k leads to the iterations

$$\tilde{x}_{i+1}^k = (I_n - \alpha_k c_i c_i^T)\tilde{x}_i^k - \alpha_k c_i, \quad i = 1, 2, \dots, 8, \tag{B.4}$$

which implies

$$\tilde{x}_1^{k+1} = \prod_{i=1}^{8}(I_n - \alpha_k c_i c_i^T)\tilde{x}_1^k - \alpha_k \sum_{i=1}^{8} c_i + \alpha_k^2 \sum_{1 \leq i < j \leq 8} (c_j^T c_i) c_j + \mathcal{O}(\alpha_k^3), \tag{B.5}$$

$$\tilde{x}_1^{k+1} = \prod_{i=1}^{8}(I_n - \alpha_k c_i c_i^T)\tilde{x}_1^k + \mathcal{O}(\alpha_k^3), \tag{B.6}$$

where in the second step we used the fact that the terms with α_k and α_k^2 above vanish due to symmetry properties imposed by relations (B.1) and (B.2). The cyclic order $\{1, 2, \dots, 8\}$ is special in the sense that it takes advantage of the symmetry in the problem leading to cancellations of the $\mathcal{O}(\alpha_k)$ and $\mathcal{O}\big(\alpha_k^2\big)$ terms in (B.5) leading to smaller $\mathcal{O}\big(\alpha_k^3\big)$ additive error terms, whereas it can be checked that this is not the case for the order $\{2, 3, \dots, 8, 1\}$. With this intuition in mind, we next show that the sequence $\{\tilde{x}_2^k\}$ converges to the optimal solution $x^* = 0$ more slowly than the sequence $\{\tilde{x}_1^k\}$ does.

Using (B.3), the fact that $x^* = 0$ for this specific example, and the triangle inequality on (B.6),

$$\begin{aligned} \text{dist}_{k+1} &\leq \left\| \prod_{i=1}^{8} \left(I_n - \alpha_k c_i c_i^T\right) \right\| \text{dist}_k + \mathcal{O}(\alpha_k^3) \\ &\leq \left|1 - 4\alpha_k + \mathcal{O}(\alpha_k^2)\right| \text{dist}_k + h_3(\alpha_k)^3 \end{aligned}$$

for some constant $h_3 > 0$. As $\alpha_k = \Theta(1/k^s)$, applying part (ii) of Lemma (2.1) with $t = 2s$ gives

$$\|\tilde{x}_1^k\| = \mathcal{O}(1/k^{2s}). \tag{B.7}$$

Then, for $i = 1$ the inner iteration (B.4) gives

$$\tilde{x}_2^k = \tilde{x}_1^k - \alpha_k(c_1^T \tilde{x}_1^k + 1)c_1. \tag{B.8}$$

As $\tilde{x}_1^k \to 0$, $(c_1^T \tilde{x}_1^k + 1)c_1 \to c_1$. Then, it follows from (B.7) and (B.8) that $\text{dist}(\tilde{x}_2^k) = \|\tilde{x}_2^k\| = \Theta(\alpha_k) = \Theta(1/k^s)$. As the order is cyclic, if we apply IG to the functions with an alternative order $f_1 = \tilde{f}_2$, $f_2 = \tilde{f}_3$, $\ldots$, $f_{m-1} = \tilde{f}_m$, and $f_m = \tilde{f}_1$ instead, the resulting iterates are $x_1^k = \tilde{x}_2^k$, which satisfy $\text{dist}(x_1^k) = \text{dist}(\tilde{x}_2^k) = \Theta(1/k^s)$. This shows that there exists component functions where the convergence behavior is $\text{dist}_k \geq \Omega(1/k^s)$. Therefore, we do not expect to improve the rate results of Theorem 3.9 in terms of dependency in k.

REFERENCES

[1] F. Bach, *Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression*, J. Mach. Learn. Res., 15 (2014), pp. 595–627.

[2] D. Bertsekas, *Incremental least squares methods and the extended Kalman filter*, SIAM J. Optim., 6 (1996), pp. 807–822.

[3] D. Bertsekas, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.

[4] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[5] D. Bertsekas, *Incremental gradient, subgradient, and proximal methods for convex optimization: A survey*, Optim. Mach. Learn., 2010 (2011), pp. 1–38.

[6] D. Bertsekas, *Convex Optimization Algorithms*, Athena Scientific, Belmont, MA, 2015.

[7] D. Bertsekas and J. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.

[8] D. Blatt, A. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM J. Optim., 18 (2007), pp. 29–51.

[9] L. Bottou and Y. Le Cun, *On-line learning for very large data sets.*, Appl. Stoch. Models Bus. Ind., 21 (2005), pp. 137–151.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.

[11] S. Bubeck, *Theory of Convex Optimization for Machine Learning*, preprint, https://arxiv.org/abs/1405.4980, 2014.

[12] K. L. Chung, *On a stochastic approximation method*, Ann. Math. Statist., 25 (1954), pp. 463–483.

[13] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Adv. Neural Inf. Process. Syst. 27, Curran Associates, Red Hook, NY, 2014, pp. 1646–1654.

[14] A. A. Gaivoronski, *Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods. Part* 1, Optim. Methods Softw., 4 (1994), pp. 117–134, https://doi.org/10.1080/10556789408805582.

[15] L. Grippo, *A class of unconstrained minimization methods for neural network training*, Optim. Methods Softw., 4 (1994), pp. 135–150, https://doi.org/10.1080/10556789408805583.

[16] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo, *A globally convergent incremental Newton method*, Math. Program., 151 (2015), pp. 283–313, https://doi.org/10.1007/s10107-015-0897-y.

[17] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo, *Why random reshuffling beats stochastic gradient descent*, Math. Program., to appear.

[18] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo, *On the convergence rate of incremental aggregated gradient algorithms*, SIAM J. Optim., 27 (2017), pp. 1035–1048.

[19] E. Hazan, A. Agarwal, and S. Kale, *Logarithmic regret algorithms for online convex optimization*, Mach. Learn., 69 (2007), pp. 169–192.

[20] V. Kekatos and G. B. Giannakis, *Distributed robust power system state estimation*, IEEE Trans. Power Syst., 28 (2013), pp. 1617–1626, https://doi.org/10.1109/TPWRS.2012.2219629.

[21] T. Kohonen, *An adaptive associative memory principle*, IEEE Trans. Comput., 23 (1974), pp. 444–445, http://doi.ieeecomputersociety.org/10.1109/T-C.1974.223960.

[22] H. Lin, J. Mairal, and Z. Harchaoui, *A universal catalyst for first-order optimization*, in Adv. Neural Inf. Process. Syst. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, Red Hook, NY, 2015, pp. 3384–3392.

[23] Z. Luo, *On the convergence of the LMS algorithm with adaptive learning rate for linear feed-forward networks*, Neural Computat., 3 (1991), pp. 226–245.

[24] Z. Luo and P. Tseng, *Analysis of an approximate gradient projection method with applications to the backpropagation algorithm*, Optim. Methods Softw., (2008).

[25] O. Mangasarian and M. Solodov, *Serial and parallel backpropagation convergence via non-monotone perturbed minimization*, Optim. Methods Softw., 4 (1994), pp. 103–116.

[26] H. Moriyama, N. Yamashita, and M. Fukushima, *The incremental Gauss–Newton algorithm with adaptive stepsize rule*, Comput. Optim. Appl., 26 (2003), pp. 107–141, https://doi.org/10.1023/A:1025703629626.

[27] E. Moulines and F. R. Bach, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, Adv. Neural Inf. Process. Syst. 24, Curran Associates, Red Hook, NY, 2011, pp. 451–459.

[28] A. Nedić and D. Bertsekas, *Convergence rate of incremental subgradient algorithms*, in Stochastic Optimization: Algorithms and Applications, Appl. Optim. 54, S. Uryasev and P. Pardalos, eds., Springer, New York, 2001, pp. 223–264.

[29] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.

[30] A. Nedić and A. Ozdaglar, *On the rate of convergence of distributed subgradient methods for multi-agent optimization*, in Proceedings of the 46th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2007, pp. 4711–4716.

[31] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Trans. Autom. Control, 54 (2009), pp. 48–61.

[32] A. Nedich, *Convergence rate of distributed averaging dynamics and optimization in networks*, Found. Trends Syst. Control, 2 (2015), pp. 1–100, https://doi.org/10.1561/2600000004.

[33] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

[34] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Springer, New York, 2004.

[35] E. Polak, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21–89.

[36] B. Polyak and A. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855, https://doi.org/10.1137/0330046.

[37] J. B. Predd, S. R. Kulkarni, and H. V. Poor, *A collaborative training algorithm for distributed learning*, IEEE Trans. Inf. Theory, 55 (2009), pp. 1856–1871, https://doi.org/10.1109/TIT.2009.2012992.

[38] A. Rakhlin, O. Shamir, and K. Sridharan, *Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization*, preprint, https://arxiv.org/abs/1109.5647, 2011.

[39] S. Ram, A. Nedić, and V. Veeravalli, *Stochastic incremental gradient descent for estimation in sensor networks*, in The Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers, IEEE Press, Piscataway, NJ, 2007, pp. 582–586, https://doi.org/10.1109/ACSSC.2007.4487280.

[40] B. Recht and C. Ré, *Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences*, in Proceedings of the 25th Annual Conference on Learning Theory, Proc. Mach. Learn. Res. 23, 2012, pp. 1–24; available at http://proceedings.mlr.press/v23/recht12.html.

[41] B. Recht and C. Ré, *Parallel stochastic gradient algorithms for large-scale matrix completion*, Math. Program. Computation, 5 (2013), pp. 201–226.

[42] W. Shi, Q. Ling, G. Wu, and W. Yin, *EXTRA: An exact first-order algorithm for decentralized consensus optimization*, SIAM J. Optim., 25 (2015), pp. 944–966, https://doi.org/10.1137/14096668X.

[43] J. Sohl-Dickstein, B. Poole, and S. Ganguli, *Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods*, in Proceedings of the International Conference on Machine Learning, E. P. Xing and T. Jebara, eds., Proc. Mach. Learn. Res. 32, 2014, pp. 604–612; available at http://proceedings.mlr.press/v32/.

[44] M. Solodov, *Incremental gradient algorithms with stepsizes bounded away from zero*, Comput. Optim. Appl., 11 (1998), pp. 23–35, https://doi.org/10.1023/A:1018366000512.

[45] E. Sparks, A. Talwalkar, V. Smith, J. Kottalam, P. Xinghao, J. Gonzalez, M. Franklin, M. Jordan, and T. Kraska, *MLI: An API for distributed machine learning*, in Proceedings of the 13th International Conference on Data Mining, IEEE Press, Piscataway, NJ,

2013, pp. 1187–1192.

[46] P. Tseng, *An incremental gradient(-projection) method with momentum term and adaptive stepsize rule*, SIAM J. Optim., 8 (1998), pp. 506–531, https://doi.org/10.1137/S1052623495294797.

[47] N. Vanli, M. Gürbüzbalaban, and A. Ozdaglar, *Global convergence rate of proximal incremental aggregated gradient methods*, SIAM J. Optim., 28 (2018), pp. 1282–1300.

[48] B. Widrow and M. E. Hoff, *Adaptive switching circuits*, in Proceedings of 1960 IRE WESCON, Institute of Radio Engineers, New York, 1960, pp. 96–104.

[49] K. Yuan, Q. Ling, W. Yin, and A. Ribeiro, *A linearized Bregman algorithm for decentralized basis pursuit*, in Proceedings of the 21st European Signal Processing Conference, IEEE Press, Piscataway, NJ, 2013, pp. 1–5.