**Noname manuscript No.**
(will be inserted by the editor)

# A globally convergent incremental Newton method

**M. Gürbüzbalaban** ·
**A. Ozdaglar** ·
**P. Parrilo**

Received: date / Accepted: date

**Abstract** Motivated by machine learning problems over large data sets and distributed optimization over networks, we develop and analyze a new method called incremental Newton method for minimizing the sum of a large number of strongly convex functions. We show that our method is globally convergent for a variable stepsize rule. We further show that under a gradient growth condition, convergence rate is linear for both variable and constant stepsize rules. By means of an example, we show that without the gradient growth condition, incremental Newton method cannot achieve linear convergence. Our analysis can be extended to study other incremental methods: in particular, we obtain a linear convergence rate result for the incremental Gauss-Newton algorithm under a variable stepsize rule.

**Keywords** Incremental methods · Convex optimization · Newton method · Gauss-Newton method · Strong convexity · EKF algorithm

## 1 Introduction

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

$$\begin{aligned} \text{minimize} \quad & f(x) = \sum_{i=1}^{m} f_i(x) \\ \text{subject to} \quad & x \in \mathbb{R}^n, \end{aligned} \tag{1.1}$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$ is a strongly convex and twice continuously differentiable function. This problem arises in many applications including least squares or more

M. Gürbüzbalaban · A. Ozdaglar · P. Parrilo
Laboratory for Information and Decision Systems,
Massachusetts Institute of Technology, Cambridge, MA, 02139.
E-mail: {mertg, asuman, parrilo}@mit.edu

* This work was supported by AFOSR MURI FA9550-09-1-0538 and ONR Basic Research Challenge No. N000141210997.

general parameter estimation problems (where $f_i(x)$ is the loss function representing the error between prediction of a parametric model obtained from data and the actual output), distributed optimization over networks (where $f_i(x)$ is the local objective function of agents connected through a network), dual formulation of problems with many constraints, 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) (see e.g., [4, 9, 10, 23, 24, 27, 33, 35]). An important feature of these problems is that the number of component functions $f_i$ is large and not all simultaneously available. One is therefore interested in optimization algorithms that can iteratively update the estimate for an optimal solution using partial information about component functions.

One widely studied approach is the incremental gradient method, which cycles through the component functions using a deterministic order and updates the iterates using the gradient of a single component function. This method typically outperforms non-incremental methods in numerical studies since each inner iteration makes reasonable progress. However, it typically has sublinear convergence rate as it requires the stepsize to go to zero to obtain convergence to the optimal solution of problem (1.1) (see [4]).

In this paper, we present an incremental Newton (IN) method that cycles deterministically through the component functions $f_i$ and uses the gradient of $f_i$ to determine the direction of motion and the Hessian of $f_i$ to construct the Hessian of the sum of component functions, $f$. Our main results can be summarized as follows:

First, we adopt a variable stepsize rule, which was introduced in Moriyama *et al.* [21] for the analysis of the incremental Gauss-Newton method with an adaptive stepsize rule. The stepsize measures the progress of the iterates over a cycle relative to the progress in the inner iterations and aims to dampen the oscillations associated with incremental methods in the "region of confusion" (i.e., the set over which the component functions have non-aligned gradients; see e.g. [3, Example 1.5.5]). We show that our IN algorithm is globally convergent with this variable stepsize rule.

Second, we identify a sufficient condition, which we refer to as the *gradient growth* condition, under which the normalized stepsize sequence (normalization of stepsize by the iteration number $k$ is used since the Hessians are accumulated at each step) remains bounded away from zero[1]. We also provide an explicit characterization of this bound in terms of problem parameters. Our analysis relies on viewing the IN method as an inexact perturbed Newton method. We use the lower and upper bounds on the stepsize sequence together with bounds on the Hessian of iterates to provide bounds on the Hessian error and the gradient error of the method. This allows us to use the convergence rate results on inexact perturbed Newton methods to show that IN method converges locally linearly to the optimal solution of problem (1.1). Under some additional assumptions, we show that IN method achieves asymptotically error-free curvature (or Hessian matrix of $f$) estimates which do not extend to many incremental quasi-Newton methods (see

[1] This condition requires that norms of gradients of $f_i$'s are bounded from above by a linear function of the norm of $f$, implying that each component function $f_i$ has the same (unique) minimizer. This is a strong condition, and is satisfied only in a limited number of applications (see Section 3.1). However, we show that it is indispensable for linear convergence of the incremental Newton methods.

Remark 3.6). However, our global convergence and linear convergence rate results admit extensions to incremental quasi-Newton methods. Our analysis can also be extended to study *incremental Gauss-Newton* method under a variable stepsize rule for solving least square problems, also known as the extended Kalman filter (EKF) method with variable stepsize or equivalently the EKF-S algorithm [21], and shows linear convergence rate for this method, thus answering a problem left open in Moriyama *et al.* [21, §7]. Note that the incremental Gauss-Newton method without the variable stepsize shows typically sublinear convergence behavior [1, 3, 13].

Third, we show that under the gradient growth condition, the IN method converges globally linearly to the optimal solution for a sufficiently small constant stepsize. The analysis of our algorithm under a constant stepsize rule uses bounds on the gradient errors, which may be of independent interest. We also provide an example that shows that without the gradient growth condition, IN method cannot converge faster than sublinear, thus highlighting the importance of gradient growth condition in the performance of the IN method.

The assumption that each component function is strongly convex in (1.1) can be relaxed while preserving global convergence of the IN method. In fact, it suffices that at least one of the component functions is strongly convex (see Remark 3.1). Furthermore, even if none of the component functions is strongly convex, global convergence can be guaranteed by a simple modification to IN which ensures that Hessian approximations are bounded away from zero (see Remark 3.2). Thus, IN method has a broader range of applicability beyond strongly convex setting.

Our work is related to the literature on incremental gradient (IG) methods (see [2–4, 34]). The randomized version of the IG method, also referred to as the *stochastic gradient descent* (SGD) [9,28,31], has been popular and used extensively for solving machine learning problems [8,9,38]. Many variants of the IG method are proposed to accelerate its convergence, including the *IG method with momentum* of Tseng [36] and Mangasarian *et al.* [19]. Tseng's approach with momentum [36] requires once in a while constructing the gradient of $f$, and can be hard to implement in problems where the entire objective function is not available. Another interesting class of methods includes the *incremental aggregated gradient* (IAG) method of Blatt *et al.* (see [6,37]) and closely-related stochastic methods including the *stochastic average gradient* (SAG) method [29], the SAGA method [14] and the MISO method [18]. These methods process a single component function at a time as in incremental methods, but keeps a memory of the most recent gradients of all component functions so that a full gradient step is taken at each iteration. They have been shown to have fast convergence properties but may require an excessive amount of memory when $m$ is large.

There has also been a recent interest in incremental and stochastic second-order methods (see [5, Chapter 2] for a survey), motivated by numerical evidence showing that second-order methods are faster than first-order methods in many practical problems [7,11,20,33]. In particular, Mokhtari *et al.* propose a stochastic BFGS algorithm with a $O(1/k)$ convergence result [20]. Byrd *et al.* [11] develop a stochastic quasi-Newton algorithm that avoids the potentially harmful effects of differencing stochastic gradients that can be noisy, although no convergence analysis is given. SGD-QN algorithm [7], AdaGrad algorithm [17], oBFGS and oLBFGS algorithms [31], SFO algorithm [33] are among other recent second-order stochastic methods that use quasi-Newton approaches. DANE algorithm [32] is a Newton-like method based on mirror-descent type updates with a linear convergence rate when the functions $f_i$ are quadratics, although to the best of our knowledge no convergence rate results are currently known beyond quadratic objective functions.

*Outline.* In Section 2, we motivate and introduce the IN method deriving key lemmas for its analysis. In Section 3, first we show its global convergence under a variable stepsize rule. Then, we introduce the gradient growth assumption and under this assumption we prove local linear convergence. We also discuss implications of our analysis to the incremental quasi-Newton methods and the EKF-S algorithm. In Section 4, we start with deriving upper bounds on the norm of the gradient error in our method for an arbitrary stepsize and then show global linear convergence with a constant stepsize under the gradient growth assumption. In Section 5, we give examples illustrating the possible sublinear convergence of the IN method in case this assumption does not hold. We conclude by Section 6 with a summary of our results.

## 2 The IN Method

Newton's method is an important classical method for solving smooth uncon-stained optimization problems of the form

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \mathbb{R}^n. \end{aligned}$$

The standard Newton iteration is

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k), \tag{2.1}$$

where $\nabla f(x)$ and $\nabla^2 f(x)$ denote the gradient and Hessian of $f$ at $x \in \mathbb{R}^n$ respectively. In this paper, we focus on problem (1.1) where

$$f(x) = \sum_{i=1}^{m} f_i(x)$$

with $m \geq 2$. Computing $\nabla f(x^k)$ and $\nabla^2 f(x^k)$ in (2.1) necessitates summing gradients and Hessian matrices of each of the component functions $f_i$ which may be costly when the number of the functions $m$ is large. For such problems, it may be more effective to use an incremental method, which cycles through each of the component functions $f_i$ and update $x^k$ based on the gradient $\nabla f_i$ and the Hessian $\nabla^2 f_i$ (see e.g. [3]). Our aim is to provide an incremental version of the Newton method. A natural idea would be to approximate $\nabla f$ with $\nabla f_i$, where $i \in \{1, 2, \ldots, m\}$ varies in a cyclic manner and to construct $\nabla^2 f$ by incrementally summing $\nabla^2 f_i$'s. These observations motivate the following algorithm, which we call the *incremental Newton* (IN) algorithm. Given an initial point $x_1^1 \in \mathbb{R}^n$, we consider the iterations

$$\text{(IN)} \qquad x_{i+1}^k := x_i^k - \alpha^k \left(H_i^k\right)^{-1} \nabla f_i(x_i^k), \quad i = 1, 2, \ldots, m, \tag{2.2}$$

$$x_1^{k+1} := x_{m+1}^k, \tag{2.3}$$

where $H_i^k$ is a symmetric matrix updated by

$$H_i^k := H_{i-1}^k + \nabla^2 f_i(x_i^k), \quad i = 1, 2, \dots, m, \tag{2.4}$$

$$H_0^{k+1} := H_m^k, \quad H_0^1 := 0, \tag{2.5}$$

and $\alpha^k > 0$ is the stepsize. The matrices $H_i^k$ accumulate and capture the second-order information at the iterates. For a fixed value of $k \geq 1$ and $i \in \{1, 2, \dots, m\}$, we refer to the update (2.2) as an *inner iteration*. Consecutive $m$ iterations starting with $i = 1$ will be denoted as a *cycle* of our algorithm.

Algorithm IN is reminiscent of the EKF algorithm (when $\alpha^k = 1$) [1] or the EKF algorithm with variable stepsize (EKF-S algorithm) [21], but there are major differences: EKF and EKF-S are Gauss-Newton based methods designed specifically for the least square problems using only first-order derivatives whereas Algorithm IN applies not only to least square problems, but also to problem (1.1) and is a Newton-based method that uses second-order derivatives in addition to first-order derivatives. When $\alpha^k = 1$, the IN iterations satisfy

$$x_{i+1}^k = \arg\min_{x \in \mathbb{R}^n} \sum_{\ell=1}^{k-1} \sum_{j=1}^{m} \hat{f}_j(x, x_j^\ell) + \sum_{j=1}^{i} \hat{f}_j(x, x_j^k), \quad i = 1, 2, \dots, m, \tag{2.6}$$

where $\hat{f}_j(x, x_j^k)$ is the standard quadratic approximation to $f_j$ around the point $x_j^k$ formed by the Taylor's series expansion given by

$$\hat{f}_j(x, x_j^k) = f_j(x_j^k) + \nabla f_j(x_j^k)^T (x - x_j^k) + \frac{1}{2}(x - x_j^k)^T \nabla^2 f_j(x_j^k)(x - x_j^k). \tag{2.7}$$

To see this, for $k \geq 1$, define recursively

$$z_{i+1}^k = \arg\min_{x \in \mathbb{R}^n} \sum_{\ell=1}^{k-1} \sum_{j=1}^{m} \hat{f}_j(x, z_j^\ell) + \sum_{j=1}^{i} \hat{f}_j(x, z_j^k), \quad i = 1, 2, \dots, m, \tag{2.8}$$

with the starting point $z_1^1 = x_1^1$ and the convention that $z_1^{k+1} = z_{m+1}^k$ and the double sum is zero when $k = 1$. Let $F_i^k(x)$ denote the total sum on the right-hand side of (2.8). It is easy to see that $F_i^k(x)$ is a quadratic function with a Hessian matrix equal to $H_i^k$ (see also (2.18)) and

$$\begin{aligned} F_i^k(x) &= F_i^k(z_{i+1}^k) + \frac{1}{2}(x - z_{i+1}^k)^T H_i^k (x - z_{i+1}^k) \\ &= F_{i-1}^k(x) + \hat{f}_i(x, z_i^k) = F_{i-1}^k(z_i^k) + \frac{1}{2}(x - z_i^k)^T H_{i-1}^k (x - z_i^k) + \hat{f}_i(x, z_i^k) \end{aligned}$$

which follows from the Taylor expansions of $F_i^k(x)$ and $F_{i-1}^k(x)$ around their minimizers $z_{i+1}^k$ and $z_i^k$ respectively. Using first-order optimality conditions for $z_{i+1}^k$, we obtain the recurrence

$$z_{i+1}^k = z_i^k - \left(H_i^k\right)^{-1} \nabla f_i(z_i^k)$$

which is exactly the same as the inner updates (2.2) when $\alpha^k = 1$. This implies that $z_i^k = x_i^k$, proving the identity (2.6). As a consequence, when each function $f_j$ is a quadratic, we have $\hat{f}_j = f_j$ for $j = 1, 2, \dots, m$ and it suffices to have only one cycle ($m$ inner iterations) of the IN method to reach out to the globally optimal solution. This is clearly much faster than first-order methods or the Newton-like methods such as the DANE method [32] which has only linear convergence for quadratic $f_i$'s. However, the trade-off for this accuracy in our method is increased memory requirement $O(n\times n)$ and the additional computation of the second-order derivatives.

We start with a lemma that provides a characterization for the evolution of inner iterations.

**Lemma 2.1** *Let* $\{x_1^k, x_2^k, \dots, x_m^k\}$ *be the iterates formed by the IN algorithm given by* (2.2)–(2.5)*. Then, for* $i = 1, 2, \dots, m$*, we have*

$$x_{i+1}^k = x_1^k - \alpha^k (H_i^k)^{-1} \sum_{j=1}^{i} \left( \nabla f_j(x_j^k) + \frac{1}{\alpha^k} \nabla^2 f_j(x_j^k)(x_1^k - x_j^k) \right). \tag{2.9}$$

*Proof* Let $x_1^k$ be given. The iterations (2.2) can be rewritten as

$$x_{i+1}^k = x_i^k - \big(H_i^k\big)^{-1} \big(\alpha^k \nabla f_i(x_i^k)\big), \quad i = 1, 2, \dots, m,$$

which is equivalent to

$$x_{i+1}^k = x_i^k + \big(H_i^k\big)^{-1} \big(C_i^k\big)^T (z_i^k - C_i^k x_i^k), \quad i = 1, 2, \cdots, m, \tag{2.10}$$

where $C_i^k$ is a positive definite matrix satisfying

$$(C_i^k)^T C_i^k = \nabla^2 f_i(x_i^k), \quad H_i^k = H_{i-1}^k + (C_i^k)^T C_i^k,$$

$$z_i^k = -\big(C_i^k\big)^{-T} \big(\alpha^k \nabla f_i(x_i^k) - \nabla^2 f_i(x_i^k) x_i^k\big).$$

(Such a matrix $C_i^k$ exists and can for instance be obtained by a Cholesky decomposition of the positive definite matrix $\nabla^2 f_i(x_i^k)$). Then, the update formula (2.10) is equivalent to a Kalman filter update that solves the incremental quadratic optimization problem

$$x_{i+1}^k = \arg\min_{x\in\mathbb{R}^n} \sum_{j=1}^{i} \|z_j^k - C_j^k x\|^2, \quad \text{for} \quad i = 1, 2, \cdots, m,$$

(see [3, Proposition 1.5.2]). Using existing results on Kalman filters, (in particular [3, Proposition 1.5.2]), we obtain

$$\begin{aligned} x_{i+1}^k &= x_1^k + \big(H_i^k\big)^{-1} \sum_{j=1}^{i} (C_j^k)^T (z_j^k - C_j^k x_1^k) \\ &= x_1^k - \big(H_i^k\big)^{-1} \left( \sum_{j=1}^{i} \alpha^k \nabla f_j(x_j^k) + \nabla^2 f_j(x_j^k)(x_1^k - x_j^k) \right) \end{aligned}$$

which is equivalent to (2.9). This completes the proof. □

Using Lemma 2.1, for $i = 1, 2, \dots, m$, we can write

$$x_{i+1}^k = x_1^k - \alpha^k D_i^k h_i^k \tag{2.11}$$

where

$$D_i^k = (H_i^k)^{-1}, \quad h_i^k = \sum_{j=1}^{i} \left( \nabla f_j(x_j^k) + \frac{1}{\alpha^k} \nabla^2 f_j(x_j^k)(x_1^k - x_j^k) \right). \tag{2.12}$$

We make the following two assumptions which have been used in a number of papers for analyzing incremental and stochastic methods (see e.g. [1], [21], [19, Theorem 3.1], [20, Assumption 1], [32], [14]).

**Assumption 2.1** ***(Boundedness)*** *The sequence* $\{x_1^k, x_2^k, \dots, x_m^k\}_{k=1,2,\dots}$ *generated by the IN iterations* (2.2)–(2.5) *is contained in a convex, compact set* $\mathcal{X} \subset \mathbb{R}^n$ *whose diameter is*

$$R := \max_{x,y \in \mathcal{X}} \|x - y\|. \tag{2.13}$$

**Assumption 2.2** ***(Hessian boundedness)*** *The functions* $f_i$, $i = 1, 2, \dots, m$ *are twice continuously differentiable and convex on* $\mathbb{R}^n$*, and there exists constants* $c > 0$ *and* $L > 0$ *such that*[2]

$$cI \preceq \nabla^2 f_i(x) \preceq LI, \tag{2.14}$$

*for all* $x \in \mathbb{R}^n$ *and* $i = 1, 2, \dots, m$.

A consequence of Assumption 2.2 is that the function $f$ is strongly convex with parameter $cm > 0$ as each $f_i$ is strongly convex with parameter $c > 0$. Thus, the optimization problem (1.1) admits a unique optimal solution. Another consequence is that the gradients have a Lipschitz constant $L$, i.e.,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad i = 1, 2, \dots, m, \tag{2.15}$$

for all $x, y \in \mathbb{R}^n$, where we use $\|\cdot\|$ to denote the 2-norm (Euclidean norm) of a vector or the 2-norm (spectral norm) of a matrix depending on the context throughout this paper. We note that, by (2.14), the ratio

$$Q := \frac{L}{c} \tag{2.16}$$

is an upper bound for the condition number of the Hessian matrices at the iterates. Therefore, we will refer to it as the *condition number* of problem (1.1).

As an example application, consider the following non-linear least square problem for which incremental methods can be order of magnitude faster than non-incremental methods as each inner iteration makes reasonable progress on average [4]: We want to minimize $\sum_{i=1}^m f_i(x)$ where

$$f_i(x) = \sigma_i(a_i^T x - b_i) + \frac{\lambda}{2}\|x\|^2, \tag{2.17}$$

[2] The assumption on the Hessian bounds (2.14) can be relaxed to hold only for all $x \in \mathcal{X}$ (instead of for all $x \in \mathbb{R}^n$) as long as $\mathcal{X}$ defined in Assumption 2.1 is large enough to contain a stationary point of $f$. Note that the existence of the upper bound $L$ (for the Hessian matrices) on the compact set $\mathcal{X}$ is an immediate implication of Assumption 2.1 as $f$ is twice continuously differentiable.

$\lambda > 0$ is a regularization parameter, $a_i$ are vectors, $b_i$ are scalars and each $\sigma_i : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable and convex function on $\mathbb{R}$ that is used to penalize the error between some data and the output of a parameteric model. For example $\sigma_i(x) = \|x\|^2$ corresponds to $\ell_2$-regularized linear regression. Due to strong convexity of $\sigma_i$ and $f_i$ and continuity of the second-derivatives of $f_i$, Assumption 2.2 holds whenever Assumption 2.1 is satisfied. In particular, we have

$$\nabla^2 f_i(x) = \Big(\nabla^2 \sigma_i(a_i^T x - b_i)\Big) a_i a_i^T + \lambda I \succeq \lambda I$$

which satisfies Assumption 2.2 with $c = \lambda$ on any compact set $\mathcal{X}$.

We now investigate the evolution of the Hessian matrices $H_i^k$ with $k$. It is straightforward to see from the Hessian update formula (2.4) that

$$H_i^k = \sum_{i=1}^{k-1} \sum_{j=1}^{m} \nabla^2 f_j(x_j^i) + \sum_{j=1}^{i} \nabla^2 f_j(x_j^k). \tag{2.18}$$

The next lemma shows that the matrices $\{H_i^k\}_{i,k\geq 1}$ have a norm (2-norm) growing linearly with $k$.

**Lemma 2.2** *Suppose that Assumptions 2.1 and 2.2 hold. Then, for any* $i = 1, 2, \ldots, m$ *and* $k \geq 1$*, we have*

$$ckI \preceq c\big((k-1)m + i\big)I \preceq H_i^k \preceq LmkI, \tag{2.19}$$

$$\frac{1}{Lmk} I \preceq D_i^k \preceq \frac{1}{c\big((k-1)m+i\big)} I \preceq \frac{1}{ck} I. \tag{2.20}$$

*It follows that, for any* $i = 1, 2, \ldots, m$ *and* $k \geq 2$,

$$\frac{ckm}{2} I \preceq H_i^k, \quad D_i^k \preceq \frac{2}{ckm} I. \tag{2.21}$$

*Proof* The first inequality on the left of (2.19) is a direct consequence of the inequality $k \leq \big((k-1)m + i\big)$ for $k \geq 1$ and $i \in \{1, 2, \ldots, m\}$. Other inequalities in (2.19) and (2.20) follow directly from applying the Hessian bounds (2.14) to the representation of $H_i^k$ given by the formula (2.18) and the fact that $D_i^k = (H_i^k)^{-1}$. Inequalities (2.21) follow from (2.19) and (2.20) using the fact that $k/(k-1) \leq 2$ for $k \geq 2$. □

## 3 Convergence with variable stepsize

In this section, we introduce a variable stepsize rule and study global convergence of the IN method under this stepsize. We use the variable stepsize that was proposed in Moriyama *et al.* [21] for showing convergence of the EKF-S method.

**Assumption 3.1** ***(Variable stepsize)*** *The stepsize used in the Algorithm IN defined by* (2.2)–(2.5) *satisfies*

$$1 \leq \alpha^k \leq \max(1, \alpha_*^k), \quad k = 1, 2, \ldots,$$

*where*

$$\alpha_*^k = \begin{cases} \frac{1-\eta}{L} \frac{(x_1^{k+1}-x_1^k)^T H_m^k (x_1^{k+1}-x_1^k)}{\|x_1^{k+1}-x_1^k\| \sum_{i=2}^m \|x_i^k-x_1^k\| + \frac{m}{2}\|x_1^{k+1}-x_1^k\|^2}, & \textit{if } x_1^k \neq x_1^{k+1}, \\ 0, & \textit{otherwise}, \end{cases} \tag{3.1}$$

*for some* $\eta \in (0,1)$*, which we refer to as the* stepsize control parameter.

The form of this stepsize can be motivated as follows: The representation formulae for the inner iterates (2.11)–(2.12) show that if the norm of $\nabla f$ is very small compared to the norm of $\nabla f_i$ for some $i$, unless the stepsize $\alpha^k$ is small, we could be in a situation where the total distance traveled during one cycle of the iteration $\|x_1^{k+1} - x_1^k\|$ is very small compared to that of the inner iterations $\|x_i^k - x_1^k\|$, resulting in large oscillations. This suggests to have a variable stepsize that takes smaller steps if the ratio $\|x_1^{k+1} - x_1^k\| / \sum_i \|x_i^k - x_1^k\|$ gets smaller as in Assumption 3.1. Such a variable stepsize would kill undesired oscillations, enabling moving towards the optimal solution in a more efficient way. This stepsize can also be motivated by a descent condition, which leads to the monotonic decrease of the function values $\{f(x_1^k)\}_k$ asymptotically (when $k$ is large enough) as long as the iterates stay bounded (see [21, Lemma 4.1]). Furthermore, it is easy to implement by a simple adaptive stepsize algorithm (see Remark 3.4). For efficient implementation techniques and promising practical performance of this variable stepsize with incremental Gauss-Newton methods, we also refer the reader to [21].

By Assumptions 2.1 and 2.2 on the boundedness of the iterates, gradients and Hessian matrices, we see that $h_m^k$ defined by (2.12) is bounded. Hence, by Lemma 2.2 that provides bounds for the matrices $D_m^k$ and (2.11) on the evolution of inner iterates, it follows that the distance between the consecutive iterates at the beginning of each cycle satisfies

$$\|x_1^{k+1} - x_1^k\| = \alpha^k \|D_m^k h_m^k\| = O(\alpha^k/k).$$

Thus, the *normalized stepsize*

$$\gamma^k = \alpha^k/k \tag{3.2}$$

can be thought as the effective stepsize whose behavior determines the convergence rate. If $\alpha^k$ is bounded, then $\gamma^k \to 0$ in which case we would expect sublinear convergence in general as Example 5.1 in Section 5 shows. For faster convergence, we would need $\alpha^k$ and (hence $\alpha_*^k$ by Assumption 3.1) to grow with $k$. This motivates us to define

$$\gamma_*^k = \alpha_*^k/k, \quad \underline{\gamma} = \liminf_{k\to\infty} \gamma_*^k \quad \text{and} \quad \overline{\gamma} = \limsup_{k\to\infty} \gamma_*^k, \tag{3.3}$$

requiring a lower bound on the growth rate $\underline{\gamma}$. For linear convergence, we would also typically need an upper bound on the stepsize, because even the simplest methods (such as the steepest descent method) with constant stepsize require the stepsize to be small enough in order to be able to guarantee linear convergence [26, Section 1.4.2]. This motivates the next result which provides an upper bound for $\overline{\gamma}$.

**Lemma 3.1** *Suppose that Assumptions 2.1, 2.2 and 3.1 hold. Then, we have*

$$\gamma_*^k \leq \phi \quad \textit{for all} \quad k = 1, 2, \ldots, \tag{3.4}$$

*with*

$$\phi = 2(1-\eta)Q > 0 \tag{3.5}$$

*where $\eta$ is the stepsize control parameter as in* (3.1) *and $Q$ is the condition number of the problem* (1.1) *defined by* (2.16). *It follows that $\overline{\gamma} \le \phi$.*

*Proof* If $h_m^k \neq 0$, then

$$\begin{aligned}
\alpha_*^k &= \frac{1-\eta}{L} \frac{(h_m^k)^T D_m^k h_m^k}{\|D_m^k h_m^k\| \sum_{i=2}^m \|D_{i-1}^k h_{i-1}^k\| + \frac{m}{2}\|D_m^k h_m^k\|^2} \\
&\le \frac{1-\eta}{L} \frac{\frac{1}{cmk}\|h_m^k\|^2}{(1/Lmk)^2(\|h_m^k\| \sum_{i=2}^m \|h_{i-1}^k\| + \frac{m}{2}\|h_m^k\|^2)} \\
&= \phi \frac{1}{(2/m)\sum_{i=2}^m \|h_{i-1}^k\|/\|h_m^k\| + 1} k \\
&\le \phi k
\end{aligned}$$

where the first equality follows from (2.11), the first inequality is obtained by using the lower and upper bounds (2.20) on $D_i^k$ for $i = 1, 2, \ldots, m$ and the second inequality follows since the term $(2/m)\sum_{i=2}^m \|h_{i-1}^k\|/\|h_m^k\|$ is non-negative. This implies (3.4). Otherwise, if $h_m^k = 0$, then $\alpha_*^k = \gamma_*^k = 0$ satisfying (3.4) clearly. □

The next theorem shows the global convergence of the iterates generated by the IN method to the unique optimal solution of problem (1.1). The proof uses a similar line of argument as in the proof of [21, Theorem 4.1]; so we skip it here due to space considerations.

**Theorem 3.1** ***(Global convergence with variable stepsize)*** *Suppose that Assumptions 2.1, 2.2 and 3.1 hold. Then the iterates $\{x_1^k\}_{k=1}^\infty$ generated by the IN method* (2.2)–(2.5) *satisfy*

$$\lim_{k\to\infty} \|\nabla f(x_1^k)\| = 0$$

*and converge to the unique optimal solution of the optimization problem* (1.1).

*Remark 3.1* **(Global convergence with (at least) one strongly convex function)** The assumption in Theorem 3.1 on the strong convexity of each component function $f_i$, $i = 1, 2, \ldots, m$ (see Assumption 2.2) leads to an $O(1/k)$ positive definite lower and upper bounds on $D_i^k$ in Lemma 2.2 and this plays a key role in the proof of Theorem 3.1. However, this assumption can be relaxed. As long as one of the component functions is strongly convex and $H_0^1$ is initialized to a positive definite matrix (for instance, the identity matrix), Theorem 3.1 is still valid. This is because with these assumptions, the matrices $H_i^k$ stay positive-definite with upper and lower bounds that grow linearly in $k$ and Lemma 2.2 would still be true with minor modifications to the constants.

*Remark 3.2* **(Global convergence without strong convexity)** If none of the component functions is strongly convex, as long as each $f_i$ is convex and twice continuously differentiable, one can add a small positive definite quadratic term $f_\varepsilon(x) = \frac{\varepsilon}{2}\|x - x^0\|^2$ to the first function $f_1(x)$ to make it strongly convex and then minimize $f(x) + f_\varepsilon(x)$ with IN method (see also Remark 3.1) and get global convergence. Another possibility is to fix a real positive constant $\varepsilon > 0$ and use the positive definite matrix $\varepsilon I$ instead of $\nabla^2 f_1(x_1^k)$ at the beginning of each cycle whenever $\nabla^2 f_1(x_1^k) \preceq \varepsilon I$ for some $k$ (This is similar in spirit to the regularization of Newton's method in lack of strong convexity). Then, if Assumptions 2.1 and 3.1 hold and if each $f_i$ is bounded below, convex and twice continuously differentiable on $\mathcal{X}$ (without being strongly convex), a similar line of reasoning to Theorem 3.1 shows that we also have asymptotic stationarity, i.e.

$$\lim_{k\to\infty} \|\nabla f(x_1^k)\| = 0.$$

## 3.1 Linear Convergence

We use the following assumption which was also adopted in [21, 30, 34, 36] for analyzing stochastic and incremental gradient methods.

**Assumption 3.2** ***(Gradient growth condition)*** *There exists a positive constant $M$ such that*

$$\|\nabla f_i(x)\| \leq M\|\nabla f(x)\|$$

*for all $i = 1, 2, \ldots, m$.*

Assumption 3.2 states that the norm of $\nabla f_1, \nabla f_2, \ldots, \nabla f_m$ is bounded by a linear function of the norm of $\nabla f$. Thus, it limits the oscillations that might arise due to an imbalance between the norm of $\nabla f_i$ (for some $i$) and the norm of $\nabla f$ which led us previously to adopt a variable stepsize that gets smaller when such oscillations arise (see the paragraph after Assumption 3.1). Indeed, we show in Theorem 3.2 that this assumption, by limiting such oscillations, can avoid the variable stepsize rule of Assumption 3.1 getting too small (keeping the normalized stepsize bounded away from zero).

Note that Assumption 3.2 requires $\nabla f_1(x) = \nabla f_2(x) = \cdots = \nabla f_m(x) = 0$ at a stationary point $x$ of $f$. This requirement, although restrictive, is not completely unrealistic for certain applications such as neural network training problems or least square problems when each component function is strongly convex (for example; linear regression, $\ell_2$-regularized logistic regression [29] or the example (2.17) from Section 2) and the residual error is zero [30, 36]. Under this assumption,

- Tseng [36] shows that his incremental gradient method is either linearly convergent or the stepsize is bounded away from zero in which case convergence rate is not known. It is not clear whether his method can achieve linear convergence under stronger conditions. This method is not applicable to our setting as it requires constructing and evaluating the full gradient, $\nabla f$.
- Solodov [34] shows global convergence of the IG method with a constant stepsize, although no convergence rate results are given.
- Moriyama *et al.* [21] show that EKF-S method is globally convergent with a stepsize $\alpha^k$ that grows linearly with $k$ but do not provide an explicit lower bound on the growth rate or any convergence rate results.
- Schmidt [30] proves that the SGD method is linearly convergent in expectation in a stochastic setting but the analysis and results are not deterministic and do not apply directly to our method.

In this paper, we show the linear convergence of our method under Assumption 3.2 when the stepsize control parameter $\eta$ in Assumption 3.1 is appropriately chosen. As a by-product, our analysis also implies the linear convergence of the EKF-S algorithm (see Corollary 3.1) which was left open in [21, §7].

By adapting a result from Moriyama *et al.* [21, Theorem 4.2], it is not hard to show that $\underline{\gamma}$ is positive under Assumption 3.2. However, Moriyama *et al.* do not provide an explicit lower bound on $\underline{\gamma}$ (in terms of the problem constants $m$, $c$, $L$ and $M$). For estimating an explicit lower bound, we will need the following lemma.

**Lemma 3.2** *Suppose that Assumptions 2.1, 2.2, 3.1 and 3.2 hold. Let the iterates* $\{x_1^k\}_{k=1}^{\infty}$ *be generated by the IN method* (2.2)–(2.5)*. Then, for each* $i \in \{1, 2, \dots, m\}$*, we have*

$$\|x_i^k - x_1^k\| \leq \gamma^k B_i^k(\phi)\|\nabla f(x_1^k)\|, \quad \textit{for all} \quad k \geq 2, \tag{3.6}$$

*where* $B_i^k(\phi)$ *is given by the recursion*

$$B_{i+1}^k(\phi) = \left(1 + \frac{2Q}{m}\max(1/k, \phi)\right)B_i^k(\phi) + \frac{2M}{cm} \quad \textit{and} \quad B_1^k = 0, \tag{3.7}$$

*and the limits* $B_i(\phi) := \lim_{k\to\infty} B_i^k(\phi)$ *satisfy*

$$B_{i+1}(\phi) = \left(1 + \frac{2Q}{m}\phi\right)B_i(\phi) + \frac{2M}{cm} \quad \textit{and} \quad B_1(\phi) = 0, \tag{3.8}$$

*where* $\phi$ *is given by* (3.5)*.*

*Proof* Fix $k$. We will proceed by induction on $i$. For $i = 1$, the left-hand side of (3.6) is zero, so the result holds. For simplicity of the notation, we will write $B_i^k$ for $B_i^k(\phi)$. Suppose that (3.6) holds for all $1 \leq i \leq j \leq m$. Then,

$$\begin{aligned}
\|x_{j+1}^k - x_1^k\| &\leq B_j^k\gamma^k\|\nabla f(x_1^k)\| + \|x_{j+1}^k - x_j^k\| \\
&\leq B_j^k\gamma^k\|\nabla f(x_1^k)\| + \frac{2\gamma^k}{cm}\|\nabla f_j(x_j^k)\| \\
&\leq B_j^k\gamma^k\|\nabla f(x_1^k)\| + \frac{2\gamma^k}{cm}\Big(M\|\nabla f(x_1^k)\| + L\|x_j^k - x_1^k\|\Big) \\
&\leq B_j^k\gamma^k\|\nabla f(x_1^k)\| + \frac{2\gamma^k}{cm}\Big(M\|\nabla f(x_1^k)\| + LB_j^k\gamma^k\|\nabla f(x_1^k)\|\Big) \\
&= \left(B_j^k(1 + \frac{2Q}{m}\gamma^k) + \frac{2M}{cm}\right)\gamma^k\|\nabla f(x_1^k)\|,
\end{aligned} \tag{3.9}$$

where we used the induction hypothesis in the first and the fourth inequality, the inner update equation (2.2) for relating the distance between inner iterates to gradients and Lemma 2.2 for bounding the norm of $(H_j^k)^{-1}$ in the second inequality, and the third inequality follows from Assumption 3.2 on the gradient growth, the triangle inequality over the gradients and (2.15) on the Lipschitzness of the gradients. Using Assumption 3.1 on the variable stepsize and the bound on the normalized stepsize from Lemma 3.1, we have for any $k \geq 1$,

$$\gamma^k \leq \max(1/k, \gamma_*^k) \leq \max(1/k, \phi),$$

which, once combined with (3.9), implies that the inequality (3.6) is true for $i = j+1$. This completes the induction-based proof of the equality (3.7). Then, (3.8) follows directly from (3.7) by taking the limit as $k \to \infty$ and using the fact that $\phi > 0$. □

We use the preceding result to provide a lower bound on the asymptotic behavior of the normalized stepsize.

**Theorem 3.2** ***(Asymptotic stepsize behavior)*** *Suppose that Assumptions 2.1, 2.2, 3.1 and 3.2 hold. Then, there exists a constant $\kappa$ such that*

$$0 < \kappa \leq \underline{\gamma}. \tag{3.10}$$

*Furthermore, if $\phi < \frac{1}{B(\phi)L}$ where $\phi$ is defined by* (3.5),

$$B(\phi) := \sum_{j=2}^{m} B_j(\phi) \tag{3.11}$$

*and $B_j(\phi)$ is given by* (3.7)*, then a choice of*

$$\kappa = \phi \frac{1}{Q^2} \frac{1}{\frac{2B(\phi)L}{1-B(\phi)L\phi} + 1} \tag{3.12}$$

*satisfies* (3.10)*.*

*Proof* The existence of such $\kappa$ follows by a reasoning along the lines of [21, Theorem 4.2]. To prove the second part, suppose that $\phi < \frac{1}{B(\phi)L}$ and Assumptions 2.1, 2.2, 3.1 and 3.2 hold. Using the definition of $h_m^k$ from (2.12),

$$\begin{aligned} \|\nabla f(x_1^k) - h_m^k\| &\leq \sum_{j=1}^{m} \left( \|\nabla f_j(x_j^k) - \nabla f_j(x_1^k)\| + \frac{1}{\alpha^k} \|\nabla^2 f_j(x_j^k)(x_1^k - x_j^k)\| \right) \\ &\leq L(1 + 1/\alpha^k) \sum_{j=2}^{m} \|x_j^k - x_1^k\|, \end{aligned} \tag{3.13}$$

where we used (2.14) and (2.15) in the second inequality. Let $k \geq 2$. Using Lemma 3.2, we have

$$\sum_{j=2}^{m} \|x_j^k - x_1^k\| \leq \gamma^k B^k(\phi) \|\nabla f(x_1^k)\| \tag{3.14}$$

where

$$B^k(\phi) := \sum_{j=2}^{m} B_i^k(\phi); \qquad \lim_{k \to \infty} B^k(\phi) = B(\phi). \tag{3.15}$$

Combining (3.13) and (3.14),

$$\begin{aligned} \|h_m^k\| &\geq \|\nabla f(x_1^k)\| - \|\nabla f(x_1^k) - h_m^k\| \\ &\geq L\left( \frac{1}{B^k(\phi) L \gamma^k} - (1 + 1/\alpha^k) \right) \sum_{j=2}^{m} \|x_j^k - x_1^k\| \\ &= L\left( \frac{1}{B^k(\phi) L \gamma^k} - \big(1 + \frac{1}{k\gamma^k}\big) \right) \sum_{j=2}^{m} \|x_j^k - x_1^k\|, \end{aligned} \tag{3.16}$$

where $\gamma^k$ is the normalized stepsize given by (3.2). By assumption $B(\phi)L\phi < 1$. Using Lemma 3.1, this implies that $B(\phi)L\overline{\gamma} < 1$. Since $\underline{\gamma} > 0$ by the first part, the right hand side of (3.16) stays positive for $k$ large enough. Then, by re-arranging (3.16), there exists $\bar{k}$ such that for $k \geq \bar{k}$,

$$I_k := \frac{\sum_{j=2}^m \|x_j^k - x_1^k\|}{\|h_m^k\|} \leq \frac{B^k(\phi)\gamma^k}{1 - B^k(\phi)L\gamma^k - B^k(\phi)L/k}.$$

Thus,

$$\limsup_{k\to\infty} \frac{I_k}{\gamma^k} \leq \frac{B(\phi)}{1 - B(\phi)L\overline{\gamma}} \leq \frac{B(\phi)}{1 - B(\phi)L\phi}, \tag{3.17}$$

where we used Lemma 3.1 to bound $\overline{\gamma}$ and (3.15) for taking the limit superior of the sequence $\{B^k(\phi)\}$. We also have

$$\begin{aligned} \alpha_*^k &\geq \frac{1-\eta}{L} \frac{cmk(\alpha^k)^2\|D_m^k h_m^k\|^2}{\alpha^k\|D_m^k h_m^k\| \sum_{j=2}^m \|x_j^k - x_1^k\| + (m/2)(\alpha^k)^2\|D_m^k h_m^k\|^2} \\ &\geq \frac{(1-\eta)}{L} \frac{cmk}{I_k/(\alpha^k/Lmk) + m/2} = \left(2\frac{(1-\eta)}{Q} \frac{1}{2L(I_k/\gamma^k)+1}\right)k, \end{aligned} \tag{3.18}$$

where we used Lemma 2.2 to bound $H_m^k$ and $D_m^k$ in the first and second inequalities respectively. Combining (3.17) and (3.18) and letting $k \to \infty$,

$$\underline{\gamma} \geq 2\frac{(1-\eta)}{Q} \frac{1}{2L \limsup_{k\to\infty}(I_k/\gamma^k) + 1} \geq \phi \frac{1}{Q^2} \frac{1}{\frac{2B(\phi)L}{1-B(\phi)L\phi} + 1} > 0,$$

which is the desired result. □

*Remark 3.3* Since $\eta \in (0, 1)$, we have $\phi \in (0, 2Q)$ by the definition (3.5) of $\phi$. Taking limits in (3.8), we obtain

$$\lim_{\phi\to 0} B_j(\phi) = \frac{2M}{cm}(j-1), \quad \text{for} \quad j = 1, 2, \ldots, m.$$

Then, as $B_j(\phi)$ is a monotonically non-decreasing function of $\phi$ for every $j$ (see (3.8)), $B(\phi)$ defined by (3.11) is also non-decreasing in $\phi$ satisfying,

$$B_{min} := \inf_{\phi\in(0,2Q)} B(\phi) = \lim_{\phi\to 0} B(\phi) = \sum_{j=2}^m \frac{2M}{cm}(j-1) = \frac{M(m-1)}{c} > 0.$$

This shows that $B(\phi)L$ can never vanish so that the condition $\phi < \frac{1}{B(\phi)L}$ in Theorem 3.2 is well-defined. To see that this condition is always satisfied when $\phi$ is positive and small enough, note that the monotonicity of $B(\phi)$ leads to

$$B_{max} := \sup_{\phi\in(0,2Q)} B(\phi) = \lim_{\phi\to 2Q} B(\phi) < \infty$$

as well. Hence, $\phi \in (0, \frac{1}{B_{max}L})$ always satisfies this condition.

We now analyze Algorithm IN as an inexact perturbed Newton method. Using the representation (2.11) and the formula (2.18) for the Hessian matrices at the iterates, we can express IN iterations as

$$x_1^{k+1} = x_1^k - \gamma^k (\bar{H}_k)^{-1} \big(\nabla f(x_1^k) + e^k\big) \tag{3.19}$$

where

$$\bar{H}_k := \frac{H_m^k}{k} = \frac{\sum_{i=1}^k \big(\sum_{j=1}^m \nabla^2 f_j(x_j^i)\big)}{k} = \frac{\sum_{i=1}^k \nabla^2 f(x_1^i)}{k} + \widehat{e}^k \tag{3.20}$$

is the average of the Hessian of $f$ at the previous iterates up to an error term

$$\widehat{e}^k = \frac{\sum_{i=1}^k \sum_{j=1}^m \Big(\nabla^2 f_j(x_j^i) - \nabla^2 f_j(x_1^i)\Big)}{k}, \tag{3.21}$$

and the gradient error is

$$e^k = \sum_{j=1}^m \Big(\nabla f_j(x_j^k) - \nabla f_j(x_1^k) + \frac{1}{\alpha^k} \nabla^2 f_j(x_j^k)(x_1^k - x_j^k)\Big). \tag{3.22}$$

Applying (2.14) on the Hessian bounds to the first equality in (3.20), the Hessian term $\bar{H}_k$ satisfies

$$cmI \preceq \bar{H}_k \preceq LmI, \tag{3.23}$$

where $I$ is the $n \times n$ identity matrix and the Hessian error $\widehat{e}^k$ admits the simple bound

$$\|\widehat{e}^k\| \le (L-c)m = cm(Q-1). \tag{3.24}$$

We can also bound the gradient error in terms of the norm of the gradient for $k \ge 2$ as

$$\begin{aligned}\|e^k\| &\le \sum_{j=1}^m \Big(\|\nabla f_j(x_j^k) - \nabla f_j(x_1^k)\| + \frac{1}{\alpha^k}\|\nabla^2 f_j(x_j^k)\|\|x_1^k - x_j^k\|\Big) \\ &\le (L + L/\alpha^k)\gamma^k \sum_{j=1}^m B_j^k(\phi)\|\nabla f(x_1^k)\| = L(\gamma^k + 1/k)B^k(\phi)\|\nabla f(x_1^k)\|,\end{aligned}$$

where we used (2.14) on the boundedness of the Hessian matrices, (2.15) on the Lipschitzness of the gradients and Lemma 3.2 on the distance between the iterates in the second inequality, the (last) equality holds by the definitions (3.2) and (3.15). This leads to

$$\limsup_{k\to\infty} \Big(\|e^k\|/\|\nabla f(x_1^k)\|\Big) \le LB(\phi)\overline{\gamma} \le LB(\phi)\phi, \tag{3.25}$$

by Lemma 3.1.

We prove our rate results using the next theorem regarding sufficient conditions for linear convergence of the inexact perturbed Newton methods of the form

$$(F'(y^k) + \Delta_k)s^k = -F(y^k) + \delta_k \tag{3.26}$$

$$y^{k+1} = y^k + s^k, \quad y^1 \in \mathbb{R}^n, \tag{3.27}$$

where the map $F : \mathbb{R}^n \to \mathbb{R}^n$ and $F'$ denotes the Jacobian matrix of $F$, $\delta^k$ is the perturbation to $F$ and $\Delta^k$ is the perturbation to the Jacobian matrix $F'$. The local convergence of such iterates to a solution $y^*$ satisfying $F(y^*) = 0$ is well-studied. Under the following conditions,

– there exists $y^*$ such that $F(y^*) = 0$, (C1)
– The Jacobian matrix $F'(y^*)$ is invertible, (C2)
– $F$ is differentiable on a neighborhood of $y^*$ and $F'$ is continuous at $y^*$, (C3)
– $\Delta_k$ are such that $F'(y^k) + \Delta_k$ are non-singular for all $k = 1, 2, \ldots,$ (C4)

the following local linear convergence result is known in the literature.

**Theorem 3.3** ***( [12, Theorem 2.2], based on [15])*** *Assume conditions* (C1)–(C4) *are satisfied. Given* $0 \le \xi_k \le \bar{\xi} < t < 1$, $k = 1, 2, \ldots,$ *there exists* $\epsilon > 0$ *such that if* $\|y^1 - y^*\| \le \epsilon$ *and if the iterates* $\{y^k\}$ *generated by* (3.26)–(3.27) *satisfy*

$$\left\| \Delta_k \left( F'(y^k) + \Delta_k \right)^{-1} F(y^k) + \left( I - \Delta_k (F'(y^k) + \Delta_k)^{-1} \right) \delta_k \right\| \le \xi_k \|F(y^k)\|,$$

*then the convergence is linear in the sense that*

$$\|y^{k+1} - y^*\|_* \le t \|y^k - y^*\|_*, \quad k = 1, 2, \ldots.$$

*where* $\|z\|_* := \|F'(y^*) z\|$.

When $\Delta_k = 0$, Theorem 3.3 says roughly that as long as the perturbation $\delta_k$ is a *relative error* in the right-hand side of (3.26), i.e. $\|\delta_k\|/\|F(y^k)\| \le \xi_k < \bar{\xi} < 1$, we have local linear convergence. This result was first proven by Dembo *et al.* [15] and was later extended to the $\Delta_k \neq 0$ case in [12] by noticing that (3.26) can be re-arranged and put into the form

$$F'(y^k) s^k = -F(y^k) + \bar{\delta}_k$$

with

$$\bar{\delta}_k = \Delta_k \left( F'(y^k) + \Delta_k \right)^{-1} F(y^k) + \left( I - \Delta_k (F'(y^k) + \Delta_k)^{-1} \right) \delta_k$$

(see [12, Theorem 2.2]). Then, the result of Dembo *et al.* for the $\Delta_k = 0$ case is directly applicable, leading to a proof of Theorem 3.3.

In order to be able to apply Theorem 3.3, we rewrite the IN iteration given by (3.19) in the form of (3.26)–(3.27) by setting

$$\begin{aligned} &y^k = x_1^k, \quad F(\cdot) = \nabla f(\cdot), \quad \delta_k = -e^k, \quad y^* = x^*, \\ &F'(\cdot) = \nabla^2 f(\cdot), \quad \Delta_k = (\bar{H}_k/\gamma^k) - \nabla^2 f(x_1^k). \end{aligned}$$

In this formulation, the perturbation $\delta^k$ and the gradient error $e^k$ are equal up to a sign factor. The additive Hessian error $\widehat{e}^k$ is similar to $\Delta^k$ but is not exactly the same because $\Delta^k$ has also a division by the normalized stepsize in it.

Note that the previous lower bound on $\kappa$ obtained in Theorem 3.2 for the growth rate of $\alpha_*^k$ is achieved in the limit as $k$ goes to infinity. But, $\alpha_*^k$ can achieve any growth rate strictly less than $\kappa$, say $\nu k$ for some $\nu \in (0, 1)$, in finitely many steps. To capture such growth in stepsize for some finite $k$, we define the following stepsize rule that satisfies Assumption 3.1.

**Assumption 3.3** ***(Variable stepsize with linear growth)*** *The stepsize $\alpha^k$ depends on the parameters $(\widehat{\eta}, \widehat{\nu}, \widehat{\kappa})$ and satisfies*

$$\alpha^k = \alpha^k(\widehat{\eta}, \widehat{\nu}, \widehat{\kappa}) = \begin{cases} (\widehat{\nu}\widehat{\kappa})k, & \textit{if } 1 \le (\widehat{\nu}\widehat{\kappa})k \le \max\big(1, \alpha_*^k\big) \\ 1, & \textit{otherwise,} \end{cases} \tag{3.28}$$

*for some $\widehat{\kappa} > 0$ and $\widehat{\eta}, \widehat{\nu} \in (0,1)$ where $\alpha_*^k$ is defined by* (3.1) *with stepsize control parameter $\eta$ equal to $\widehat{\eta}$.*

We argue now how this choice of stepsize with linear growth can lead to linear convergence if parameters $(\widehat{\eta}, \widehat{\nu}, \widehat{\kappa})$ are chosen appropriately. Suppose that Assumptions 2.1 and 2.2 on the boundedness of iterates and Hessian matrices, Assumption 3.3 on the variable stepsize with parameters $(\eta, \nu, \kappa)$ where $\eta, \nu \in (0,1)$ and $\kappa$ is given by (3.12) and Assumption 3.2 on the gradient growth hold. Suppose also that[3]

$$\phi < \min\left(\frac{1}{Q}, \frac{1}{B(\phi)L}\right) \tag{3.29}$$

where $B(\phi)$ is given by (3.11). Using Lemma 3.1 and Theorem 3.2, $\alpha_*^k$ grows linearly with an asymptotic rate bounded from below by $\kappa$ and bounded from above by $\phi$ so that the stepsize defined by (3.28) with parameters $(\eta, \nu, \kappa)$ satisfies

$$\lim_{k\to\infty} \gamma^k = \lim_{k\to\infty} \frac{\alpha^k}{k} = \nu\kappa < \limsup_{k\to\infty} \frac{\alpha_*^k}{k} \le \phi. \tag{3.30}$$

Note that by (2.14) and (3.23) on the boundedness of the Hessian matrices $\nabla^2 f(y^k)$ and the averaged Hessian $\bar{H}_k$, we have

$$\frac{1}{Q} I \preceq \nabla^2 f(y^k) \bar{H}_k^{-1} \preceq QI. \tag{3.31}$$

As $\phi < 1/Q$ by the condition (3.29), the inequalities (3.30) and (3.31) imply that there exists a positive integer $\bar{k}$ such that for $k \ge \bar{k}$, we have $\gamma^k < 1/Q$ and

$$0 \prec \Big(1 - \gamma^k Q\Big) I \preceq \Big(I - \gamma^k \nabla^2 f(y^k) \bar{H}_k^{-1}\Big) \preceq \Big(1 - \frac{\gamma^k}{Q}\Big) I. \tag{3.32}$$

Combining (3.31) and (3.32) leads to

$$\begin{aligned} \limsup_{k\to\infty} \left\| \varDelta_k \Big(F'(y^k) + \varDelta_k\Big)^{-1} \right\| &= \limsup_{k\to\infty} \left\| I - \gamma^k \nabla^2 f(y^k) \bar{H}_k^{-1} \right\| \\ &\le 1 - \frac{\liminf_{k\to\infty} \gamma^k}{Q} = 1 - \frac{\nu\kappa}{Q} \end{aligned} \tag{3.33}$$

and similarly to

$$\begin{aligned} \limsup_{k\to\infty} \left\| \Big(I - \varDelta_k \Big(F'(y^k) + \varDelta_k\Big)^{-1}\Big) \right\| &= \limsup_{k\to\infty} \left\| \gamma^k \nabla^2 f(y^k) \bar{H}_k^{-1} \right\| \\ &\le \phi Q \end{aligned} \tag{3.34}$$

[3] Remark 3.3 shows that this condition is always satisfied when $\phi$ is small enough. It is adopted both to use the $\kappa$ bound from Theorem 3.2 (by having $\phi < 1/\big(B(\phi)L\big)$ and also to keep the normalized stepsize small enough (by having $\phi < 1/Q$ which implies $\gamma^k < 1/Q$ by Lemma 3.1) to control the norm of the perturbations in the estimates (3.33) and (3.34).

where we used (3.30) for bounding the limit superior of $\gamma^k$ and (3.31) to bound the norm of the matrix products. Hence,

$$\begin{aligned}\limsup_{k\to\infty}\frac{\left\|\left(I-\Delta_k\left(F'(y^k)+\Delta_k\right)^{-1}\right)\delta_k\right\|}{\|F(y^k)\|} &\le \phi Q\limsup_{k\to\infty}(\|\delta^k\|/\|F(y^k)\|)\\ &\le \phi^2 QB(\phi)L \qquad (3.35)\end{aligned}$$

where we used (3.34) in the first inequality and (3.25) in the second inequality. Combining (3.33) and (3.35),

$$\begin{aligned}\eta^\infty &:= \limsup_{k\to\infty}\frac{\left\|\Delta_k\left(F'(y^k)+\Delta_k\right)^{-1}F(y^k)+\left(I-\Delta_k(F'(x_k)+\Delta_k)^{-1}\right)\delta_k\right\|}{\|F(y^k)\|}\\ &\le 1-\frac{\nu\kappa}{Q}+\phi^2 QB(\phi)L.\\ &= 1-\phi\frac{\nu}{Q^3}\frac{1}{\frac{2B(\phi)L}{1-B(\phi)L\phi}+1}+\phi^2 QB(\phi)L := r_\nu(\phi), \qquad (3.36)\end{aligned}$$

where we used the definition of $\kappa$ from (3.12) in the last equality. By Assumption 2.2 on the strong convexity and the regularity of $f$, the conditions (C1)–(C4) are satisfied for $F(\cdot)=\nabla f(\cdot)$. Hence, we can apply Theorem 3.3, which says that it suffices to have

$$r_\nu(\phi) < 1 \qquad (3.37)$$

for local linear convergence. It is straightforward to see from (3.36) that this condition is satisfied for $\phi$ positive and around zero as $r_\nu(0)=1$ and the derivative $r'_\nu(0)<0$. Remembering the assumption (3.29), we conclude that there exists a positive constant $\overline{\phi}_\nu$ such that we have linear convergence when

$$0<\phi<\overline{\phi}_\nu \le \min\left(\frac{1}{Q},\frac{1}{B(\phi)L}\right). \qquad (3.38)$$

We discuss how $\overline{\phi}_\nu$ can be determined later in Remark 3.7. This condition for linear convergence (3.38) is satisfied if

$$0<1-\eta<\frac{\overline{\phi}_\nu}{2Q}, \qquad (3.39)$$

by the definition of $\phi$ in (3.5). Thus, by choosing the stepsize control parameter $\eta\in(0,1)$ close enough to 1, we can satisfy (3.39) and hence guarantee local linear convergence of the IN algorithm. These findings provide a proof of the following linear convergence result.

**Theorem 3.4** ***(Linear convergence with variable stepsize)*** *Suppose that Assumptions 2.1, 2.2 and 3.2 hold. Let $\nu\in(0,1)$ be given and the stepsize control parameter $\eta\in(0,1)$ satisfy the inequality* (3.39)*. Then, the IN method with the stepsize rule* (3.28) *defined by Assumption 3.3 with parameters $(\eta,\nu,\kappa)$ where $\kappa$ is given by* (3.12) *is locally linearly convergent with rate $t$ where $t$ is any number satisfying*

$$0\le r_\nu(\phi)<t<1,$$

*$\phi$ is defined by* (3.5) *and $r_\nu(\phi)$ is given by* (3.36)*.*

Our arguments and proof techniques apply also to the EKF method with variable stepsize rule (EKF-S algorithm) which also uses the stepsize defined by Assumption 3.1 and makes similar assumptions such as the boundedness of iterates and Lipschitzness of the gradients for achieving global convergence [21]. Our reasoning with minor modifications leads to the following linear convergence result whose proof is skipped due to space considerations.

**Corollary 3.1** ***(Linear convergence of EKF-S)*** *Consider the problem* (1.1) *with* $f_i = \frac{1}{2}g_i^2$ *where each* $g_i : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable for* $i = 1, 2, \ldots, m$ *(non-linear least squares). Consider the EKF-S algorithm of Moriyama et al. with variable stepsize* $\alpha^k$ *[21]. Let* $\nu \in (0,1)$ *be given and* $\eta \in (0,1)$*. Suppose that Assumptions 3.1–3.2, 3.4–3.5 and 4.1 from Moriyama et al. [21] hold. It follows that there exists constants* $\tilde{\kappa} > 0$ *and* $\tilde{\phi}_\nu > 0$ *such that if*

$$0 < 1 - \eta < \tilde{\phi}_\nu,$$

*then the EKF-S algorithm with the variable stepsize rule* (3.28) *with parameters* $(\eta, \nu, \tilde{\kappa})$ *given in Assumption 3.3 is locally linearly convergent.*

We continue by several remarks about satisfying the variable stepsize rules we introduced by a simple adaptive stepsize algorithm, improving the convergence rate results with additional regularity assumptions, extensions to incremental quasi-Newton methods, some advantages of an incremental Newton approach over incremental quasi-Newton approaches in our framework and determining the constant $\overline{\phi}_\nu$.

*Remark 3.4* By definition, the exact value of $\alpha_*^k$ can only be computed at the end of the $k$-th iteration as it requires the knowledge of $H_m^k$. However, it is possible to guarantee that Assumption 3.1 on the variable stepsize holds by a simple bisection-type adaptive algorithm as follows:

*Adaptive stepsize with bisection:*

1. At the start of the $k$-th cycle, i.e. right after $x_1^k$ is available, set $\alpha^k$ to an initial value, say $\alpha^k(j)$ with $j = 1$.
2. Compute $\alpha_*^k$ (depending on $\alpha^k(j)$) by running one cycle of the algorithm.
3. If Assumption 3.1 is satisfied, accept the step, set $\alpha^k = \alpha^k(j)$ and exit. Else, bisect by setting $\alpha^k(j+1) = \max\big(1, \tau\alpha^k(j)\big)$ for some $\tau \in (0,1)$, increment $j$ by 1 and go to step 2.

There is no risk of an infinite loop during the bisections, as the step $\alpha^k = 1$ is always accepted. Theorem 3.2 shows that when Assumption 3.2 on the gradient growth holds, by setting $\alpha^k(1) = (\nu\kappa)k$ in the above iteration as in the stepsize rule (3.28) with $\nu \in (0,1)$ and $\kappa$ as in (3.12), $\alpha^k(1)$ will be immediately accepted requiring no bisection steps, except for finitely many $k$.

*Remark 3.5* Main estimates used for proving Theorem 3.4 are the inequalities (3.33) and (3.34) which require only (3.23) on the boundedness of the averaged Hessian. If one replaces actual Hessians $\nabla^2 f_i(x_i^k)$ with approximate Hessians $\nabla^2 \tilde{f}_i(x_i^k)$ as long as the eigenvalues of $\nabla^2 \tilde{f}_i(x_i^k)$ are bounded (with the same constant for all $i$ and $k$) by a lower bound $\tilde{c} > 0$ and an upper bound $\tilde{L} > 0$, all these inequalities as well as Theorem 3.4 would still hold with $c$ and $L$ replaced by $\tilde{c}$ and $\tilde{L}$. Thus, the IN method admits straightforward generalizations to the incremental quasi-Newton methods while preserving its global convergence and linear convergence results.

*Remark 3.6* In the setting of Theorem 3.4, if we assume slightly more regularity than the continuity of the Hessian of $f$ implied by Assumption 2.2, the Hessian error upper bound (3.24) and convergence rates can be improved as follows: Assume that the Hessian of $f$ is not only continuous but also Hölder continuous on the compact set $\mathcal{X}$ defined in Assumption 2.1 with some Hölder exponent $\lambda$ and Hölder coefficient $L_\lambda$ satisfying $0 < \lambda \leq 1$ and $L_\lambda < \infty$ (reduces to the Lipschitz condition if $\lambda = 1$), then the Hessian error bound $\widehat{e}^k$ defined in (3.21) satisfies

$$\begin{aligned}\|\widehat{e}^k\| &\leq \frac{\sum_{i=1}^k \sum_{j=1}^m L_\lambda \|x_j^i - x_1^i\|^\lambda}{k} \\ &\leq L_\lambda \frac{\sum_{i=1}^k \sum_{j=1}^m \big(1 + B_j^i(\phi)\big)(1+\gamma^i)\|\nabla f(x_1^i)\|^\lambda}{k} \\ &\leq L_\lambda \frac{\sum_{i=1}^k \sum_{j=1}^m \big(1 + B^i(\phi)\big)(1+\gamma^i)\|\nabla f(x_1^i)\|^\lambda}{k} \\ &= L_\lambda m \frac{\sum_{i=1}^k \big(1 + B^i(\phi)\big)(1+\gamma^i)\|\nabla f(x_1^i)\|^\lambda}{k}, \end{aligned} \tag{3.40}$$

where we used the definition of Hölder continuity in the first inequality, Lemma 3.2 that provide bounds on the distances between inner iterates together with the fact that $z^\lambda \leq 1 + z$ for $z \geq 0$ and $0 < \lambda \leq 1$ in the second inequality and the upper bound $B_j^i(\phi) \leq \sum_{j=1}^m B_j^i(\phi) = B^i(\phi)$ for all $j$ (by the non-negativity of $B_j^i(\phi)$) in the third inequality. Note that the summand term in (3.40) satisfies

$$\lim_{i\to\infty} \big(1 + B^i(\phi)\big)(1+\gamma^i)\|\nabla f(x_1^i)\|^\lambda = 0 \tag{3.41}$$

because the gradient term $\nabla f(x_1^i)$ goes to zero by Theorem 3.1 on the global convergence, the sequence $\{B^i(\phi)\}_{i\geq 1}$ is bounded admitting $B(\phi) < \infty$ as a limit by (3.15) and the normalized stepsize $\gamma^i$ is bounded for any of the stepsize rules we discuss, including Assumptions 3.1 and 3.3 (see Lemma 3.1). Combining (3.40) and (3.41) shows that the Hessian error $\widehat{e}^k$ goes to zero, i.e.

$$\|\widehat{e}^k\| \to 0, \tag{3.42}$$

improving the trivial bound (3.24). Using (3.20) and the global convergence of $\{x_1^k\}$ to the optimal solution $x^*$ by Theorem 3.1, this implies that

$$\nabla^2 f_i(x_1^k) \to \nabla^2 f_i(x^*), \quad \bar{H}_k \to \nabla^2 f(x^*), \quad \nabla^2 f(y^k)\bar{H}_k^{-1} \to I, \tag{3.43}$$

with $y^k = x_1^k$. This allows us to replace the upper bounds $\big(1 - \frac{\nu\kappa}{Q}\big)$ in (3.33) and $\phi Q$ in (3.34) with $(1 - \nu\kappa)$ and $\phi$ respectively, eliminating some of the $Q$ terms in the condition $r_\nu(\phi) < 1$ for linear convergence (see (3.36) and (3.37)) and leading to the less restrictive condition

$$1 - \phi\frac{\nu}{Q^2}\frac{1}{\frac{2B(\phi)L}{1-B(\phi)L\phi} + 1} + \phi^2 B(\phi)L := \hat{r}_\nu(\phi) < 1$$

for linear convergence as $\hat{r}_\nu(\phi) \leq r_\nu(\phi)$ (with equality only in the special case $Q = 1$). This relaxed condition would not extend to many classes of incremental quasi-Newton methods (such as DFP and its variants) that uses approximations $\nabla^2 \tilde{f}_i$ instead of the true Hessian $\nabla^2 f_i$ because such methods do not lead to an asymptotically correct Hessian approximation, i.e. it is possible that $\nabla^2 \tilde{f} \not\to \nabla^2 f$ (see [16, Section 4]) so that (3.43) does not always hold. In this sense, using an incremental Newton approach instead of an incremental quasi-Newton approach in our algorithm allows us to get stronger convergence results.

*Remark 3.7* It is possible to compute $\overline{\phi}_\nu$ as follows. By a straightforward computation, the condition on linear convergence $r_\nu(\phi) < 1$ (see (3.37)) where $r_\nu(\phi)$ is given by (3.36) is equivalent to

$$\begin{aligned} r_\nu(\phi) < 1 \iff & 0 < \phi < \frac{1}{B(\phi)L}\frac{\nu}{Q^4}\frac{1}{\frac{2B(\phi)L}{1-B(\phi)L\phi}+1} && (3.44)\\ & \text{and } \phi < \min\left(\frac{1}{Q}, \frac{1}{B(\phi)L}\right) && (3.45)\\ \iff & 0 < B(\phi)^2L^2\phi^2 - \big(2B(\phi)L + 1 + \psi\big)B(\phi)L\phi + \psi =: p_1(\phi),\\ & 0 < \phi := p_2(\phi), \quad 0 < \frac{1}{Q} - \phi := p_3(\phi) \quad \text{and}\\ & 0 < 1 - \phi B(\phi)L =: p_4(\phi), \end{aligned}$$

with $\psi = \frac{\nu}{Q^4}$. It is straightforward to check that these four inequalities $\{0 < p_i(\phi)\}_{i=1}^4$ are always satisfied for $\phi$ positive and around zero. Furthermore, they are all polynomial inequalities as $B(\phi)$ is a polynomial in $\phi$. Thus, by a standard root-finding algorithm, we can compute all the roots of these four polynomial equations $\{0 = p_i(\phi)\}_{i=1}^4$ accurately and set $\overline{\phi}_\nu$ to the smallest positive root among all the roots. With this choice of $\overline{\phi}_\nu$, the inequalities (3.44)–(3.45) are satisfied for $0 < \phi < \overline{\phi}_\nu$ leading to local linear convergence.

## 4 Linear convergence with constant stepsize

Consider the IN iteration (3.19). In this section, we will analyze the case when $\gamma^k$ is equal to a constant $\gamma$ without requiring the variable stepsize rule (Assumption 3.1) to hold. We start with two lemmas that provide bounds on the norm of the difference of gradients at inner iterates $x_j^k$ and $x_1^k$, and also on the overall gradient error defined in (3.22). Note that both these lemmas hold for arbitrary (normalized) stepsizes $\gamma^k$. The first lemma is inspired by Solodov [34].

**Lemma 4.1** *Suppose that Assumptions 2.1 and 2.2 hold. Let* $\{x_1^k, x_2^k, \dots, x_m^k\}_{k=1}^\infty$ *be the IN iterates generated by* (2.2)–(2.5)*. For any given* $k \geq 1$*, let*

$$\delta_j^k := \|\nabla f_j(x_j^k) - \nabla f_j(x_1^k)\|, \quad j = 1, 2, \dots, m. \tag{4.1}$$

*Then,*

$$\delta_j^k \leq r^k \sum_{i=1}^{j-1}(1 + r^k)^{j-1-i}\|\nabla f_i(x_1^k)\| \quad \textit{for all} \quad k \geq 2, \tag{4.2}$$

*with the convention that the right-hand side of* (4.2) *is zero for* $j = 1$*, where*

$$r^k = \frac{2Q}{m}\gamma^k \tag{4.3}$$

*and* $\gamma^k$ *is the normalized stepsize defined in* (3.2)*.*

*Proof* Let $k \geq 2$ be given. When $j = 1$, $\delta_1^k = 0$ so the statement is clearly true. For $j = 2$,

$$\delta_2^k \leq L\|x_2^k - x_1^k\| = L\alpha^k \big\|D_1^k \nabla f_1(x_1^k)\big\| \leq r^k \big\|\nabla f_1(x_1^k)\big\|$$

where we used (2.15) on the Lipschitzness of the gradients in the (first) inequality, the representation of inner iterates by the formula (2.12) in the first equality and Lemma 2.2 to bound $D_1^k$ in the second inequality. Thus, the statement is also true for $j = 2$. We will proceed with an induction argument. Suppose (4.2) is true for $j = 1, 2, \ldots, \ell$ with $\ell < m$. Then, we have the upper bounds

$$\begin{aligned} \|\nabla f_j(x_j^k)\| &\leq \|\nabla f_j(x_1^k)\| + \delta_j^k \\ &\leq \|\nabla f_j(x_1^k)\| + r^k \sum_{i=1}^{j-1} (1 + r^k)^{j-1-i} \|\nabla f_i(x_1^k)\|, \end{aligned} \tag{4.4}$$

for $j = 1, 2, \ldots, \ell$. We will show that (4.2) is also true for $j = \ell + 1$. Using similar estimates as before,

$$\begin{aligned} \delta_{\ell+1}^k &\leq L\|x_{\ell+1}^k - x_1^k\| & (4.5) \\ &\leq L \sum_{i=1}^{l} \|x_{i+1}^k - x_i^k\| = L \sum_{j=1}^{l} \alpha^k \|D_j^k \nabla f_j(x_j^k)\| \\ &\leq r^k \sum_{j=1}^{\ell} \|\nabla f_j(x_j^k)\| \\ &\leq r^k \sum_{j=1}^{\ell} \left( \|\nabla f_j(x_1^k)\| + r^k \sum_{i=1}^{j-1} (1 + r^k)^{j-1-i} \|\nabla f_i(x_1^k)\| \right) \\ &= r^k \sum_{j=1}^{\ell} (1 + r^k)^{\ell-j} \|\nabla f_j(x_1^k)\| & (4.6) \end{aligned}$$

where we used (2.15) on the Lipschitzness of the gradients in the first inequality, Lemma 2.2 to bound $D_j^k$ terms in the third inequality and (4.4) to bound gradients in the fourth inequality. Thus, the inequality (4.2) is also true for $j = \ell + 1$. This completes the proof. □

The next result gives an upper bound on the norm of the gradient errors.

**Lemma 4.2** *Suppose that Assumptions 2.1 and 2.2 hold. The gradient error $e^k$ defined by* (3.22) *satisfies*

$$\|e^k\| \leq \left( r^k + \frac{2Q}{km} \right) \sum_{j=2}^{m} \sum_{i=1}^{j-1} (1 + r^k)^{j-1-i} \|\nabla f_i(x_1^k)\| \quad \textit{for all} \quad k \geq 2, \tag{4.7}$$

*where $r^k$ is defined by* (4.3)*.*

*Proof* Using triangle inequality and the upper bound (2.14) on the Hessian, the gradient error given by (3.22) admits the bound

$$\|e^k\| \leq \sum_{j=2}^{m} \left( \delta_j^k + \frac{1}{k\gamma^k} L \|x_j^k - x_1^k\| \right) \tag{4.8}$$

where $\gamma^k$ is the normalized stepsize defined by (3.2). By the estimates (4.5)–(4.6), we have also

$$\delta_{l+1}^k \leq L\|x_{\ell+1}^k - x_1^k\| \leq r^k \sum_{j=1}^{\ell} (1+r^k)^{\ell-j} \|\nabla f_j(x_1^k)\| \tag{4.9}$$

for any $\ell = 1, 2, \ldots, m-1$. Combining (4.8) and (4.9),

$$\begin{aligned} \|e^k\| &\leq (1 + \frac{1}{k\gamma^k}) r^k \sum_{j=2}^{m} \sum_{i=1}^{j-1} (1+r^k)^{\ell-i} \|\nabla f_i(x_1^k)\| \\ &= \left( r^k + \frac{2Q}{km} \right) \sum_{j=2}^{m} \sum_{i=1}^{j-1} (1+r^k)^{j-1-i} \|\nabla f_i(x_1^k)\| \end{aligned}$$

as desired. □

Lemma 4.2 shows that under Assumptions 2.1 and 2.2 on the boundedness of the iterates, gradients and Hessian matrices, the gradient error $e^k$ is bounded as long as $\gamma^k$ is bounded and $e^k \to 0$ as $k \to \infty$ if $\gamma^k \to 0$. Then, by [26, Theorem 1, Section 4.2.2], the iterates $\{x_1^k\}_{k=1}^{\infty}$ generated by (3.19) with a constant stepsize $\gamma^k = \gamma$ converge to an $\varepsilon$-neighborhood of the optimal solution linearly for some $\varepsilon > 0$ and the size of the neighborhood shrinks down as the stepsize is decreased, i.e. $\varepsilon \to 0$ as $\gamma \to 0$. This type of result was also achieved for the subgradient method [22] and the incremental gradient method for least square problems [3, Section 1.5.2]. In this paper, our focus will be the optimal solution rather than approximate solutions.

The next theorem shows that under Assumption 3.2 on the gradient growth, we have global linear convergence with a constant stepsize rule if the stepsize is small enough. This is stronger than the local linear convergence obtained for the variable stepsize rule, however unlike the variable stepsize rule, if Assumption 3.2 does not hold, the constant stepsize rule does not guarantee convergence to the optimal solution but only convergence to an $\varepsilon$-neighborhood of the optimal solution for some $\varepsilon > 0$. We first prove a lemma.

**Lemma 4.3** *Consider the IN iterates* $\{x_1^k\}_{k=1}^{\infty}$ *generated by* (3.19) *with a constant stepsize* $\gamma^k = \gamma$. *Suppose that Assumptions 2.1 and 2.2 hold and there exists a positive integer* $\hat{k}$ *such that the gradient error* $e^k$ *defined by* (3.22) *satisfies*

$$\|e^k\| \leq \bar{\xi} \|\nabla f(x_1^k)\| \quad \text{for all} \quad k \geq \hat{k}, \quad 0 \leq \bar{\xi} < \frac{1}{Q}. \tag{4.10}$$

*It follows that there exists constants* $\hat{\gamma} > 0$, $\hat{A} > 0$ *and* $0 < \hat{\rho} < 1$ *such that if* $0 < \gamma < \hat{\gamma}$, *then*

$$\|x_1^k - x^*\| \leq \hat{A} \hat{\rho}^k \quad \text{for all} \quad k \geq \hat{k}, \tag{4.11}$$

*where* $x^*$ *is the unique optimal solution of the problem* (1.1).

*Proof* Take $V(x) = f(x) - f(x^*)$ as a Lyapunov function, following the proof of [26, Theorem 2, Section 4.2.3] closely. The iteration (3.19) is equivalent to

$$x_1^{k+1} = x_1^k - \gamma s^k, \quad s^k = \bar{D}_k(\nabla f(x_1^k) + e^k), \quad \bar{D}^k := (\bar{H}_k)^{-1},$$

where

$$\frac{1}{Lm} I \preceq \bar{D}^k \preceq \frac{1}{cm} I \tag{4.12}$$

by taking the inverse of the bounds for $\bar{H}_k$ given in (3.23). Let $\langle \cdot, \cdot \rangle$ denote the Euclidean dot product on $\mathbb{R}^n$. We compute

$$\begin{aligned}\langle \nabla V(x_1^k), s^k \rangle &= \langle \nabla f(x_1^k), \bar{D}_k \nabla f(x_1^k) + \bar{D}_k e^k \rangle \\ &\geq \frac{1}{Lm} \|\nabla f(x_1^k)\|^2 - \frac{1}{cm} \bar{\xi} \|\nabla f(x_1^k)\|^2 = \frac{1}{Lm}(1 - \bar{\xi} Q) \|\nabla f(x_1^k)\|^2 \\ &\geq \frac{2}{Q}(1 - \bar{\xi} Q) V(x_1^k) \geq 0,\end{aligned}$$

where we used (4.12) for bounding $\bar{D}_k$ from below in the first inequality and the strong convexity with constant $cm$ of $f$ implied by Assumption 2.2 in the second inequality. Similarly, from (2.15), it follows that the gradients of $f$ are Lipschitz with constant $Lm$, leading to, for $k \geq \hat{k}$,

$$\|s^k\|^2 \leq \|\bar{D}_k(\nabla f(x_1^k) + e^k)\|^2 \leq \frac{(1 + \bar{\xi})^2}{(cm)^2} \|\nabla f(x_1^k)\|^2 \leq 2Lm \frac{(1 + \bar{\xi})^2}{(cm)^2} V(x_1^k),$$

where we used (4.12) to bound $\bar{D}_k$ from above together with the bound (4.10) on the gradient error in the second inequality. Then, by [26, Theorem 4, Section 2.2], there exists constants $\hat{\gamma} > 0$ and $\rho \in (0, 1)$ such that for any $0 < \gamma < \hat{\gamma}$, the iterations are linearly convergent after the $\hat{k}$–th step, satisfying

$$f(x_1^k) - f(x^*) \leq \big(f(x_1^{\hat{k}}) - f(x^*)\big) \rho^{k - \hat{k}} \quad \text{for all} \quad k \geq \hat{k}.$$

Using the bounds (2.14) on the Hessian of $f_i$, we have $cmI \preceq \nabla^2 f(x) \preceq LmI$ for all $x \in \mathbb{R}^n$. This implies the following strong convexity-based inequalities, for all $k \geq \hat{k}$,

$$\frac{cm}{2} \|x_1^k - x^*\|^2 \leq f(x_1^k) - f(x^*) \leq \frac{Lm}{2} \|x_1^{\hat{k}} - x^*\|^2 \rho^{k - \hat{k}} \leq \frac{Lm}{2} R^2 \rho^{-\hat{k}} \rho^k,$$

where $R$ is the diameter of $\mathcal{X}$ defined by (2.13). Hence, (4.11) holds with $\hat{\rho} = \rho^{1/2} > 0$ and $\hat{A} = (QR^2 \rho^{-\hat{k}})^{1/2} > 0$. This completes the proof. □

**Theorem 4.1** ***(Linear convergence with constant stepsize)*** *Consider the iterates $\{x_1^k\}_{k=1}^\infty$ generated by (3.19) with a constant stepsize $\gamma^k = \gamma$. Suppose that Assumptions 2.1, 2.2 and 3.2 hold. Then, there exists a constant $\widetilde{\gamma}$ (depending on $M, c, L$ and $m$) such that if $0 < \gamma < \widetilde{\gamma}$, the iterates are globally linearly convergent, i.e.,*

$$\|x_1^k - x^*\| \leq A \rho^k, \quad \textit{for all} \quad k = 1, 2, \ldots, \tag{4.13}$$

*for some constants $A > 0$ and $\rho < 1$.*

*Proof* Under Assumption 3.2 on the gradient growth, the bound (4.7) implies that the gradient error admits the bound

$$\begin{aligned}\|e^k\| &\le M\left(r^k + \frac{2Q}{km}\right)\sum_{j=2}^{m}\sum_{i=1}^{j-1}(1+r^k)^{j-1-i}\|\nabla f(x_1^k)\| \\ &\le M\left(r^k + \frac{2Q}{km}\right)m(1+r^k)^{m-2}\|\nabla f(x_1^k)\|.\end{aligned}$$

Let $\hat{k}$ be the smallest positive integer greater than $12MQ^2$. Assume $\gamma^k = \gamma < \frac{1}{12MQ^2}$ so that $r^k < \frac{1}{6MmQ}$ where $r^k$ is defined by (4.3). Then, for $k \ge \hat{k}$, we have

$$\|e^k\| < \frac{1}{3}\frac{1}{Q}(1+r^k)^{m-2}\|\nabla f(x_1^k)\| \le \frac{2}{3}\frac{1}{Q}\|\nabla f(x_1^k)\|, \tag{4.14}$$

if $r^k \le \sqrt[m]{2} - 1$ or equivalently if $\gamma < \frac{m}{2Q}(\sqrt[m]{2} - 1)$ by the definition of $r^k$ (see (4.3)). Combining this with Lemma 4.3, we conclude that there exists $\widetilde{\gamma} > 0$ such that when $0 < \gamma < \widetilde{\gamma}$, the error bound (4.14) is satisfied for $k \ge \hat{k}$ and there exists $\hat{A} > 0$ and $\hat{\rho} < 1$ such that

$$\|x_1^k - x^*\| \le \hat{A}\hat{\rho}^k \quad \text{for all} \quad k \ge \hat{k}.$$

Then, a choice of $A = \max(\hat{A}, R)/\hat{\rho}^{\hat{k}}$ where $R$ is as in (2.13) and $\rho = \hat{\rho}$ satisfies (4.13) which completes the proof. □

## 5 An example with sublinear convergence

In the following simple example, we show that the normalized stepsize $\gamma^k$ has to go to zero if Assumption 3.2 on the gradient growth does not hold, illustrating the sublinear convergence behavior that might arise without Assumption 3.2.

*Example 5.1* Let $f_1 = 1000x + \varepsilon x^2$ and $f_2 = -1000x + \varepsilon x^2$ for a fixed $\varepsilon > 0$ with $m = 2$ and $n = 1$. This leads to a quadratic function $f = 2\varepsilon x^2$ with a unique optimal solution, $x^* = 0$ and condition number $Q = 1$. We have

$$\nabla f_1(x) = 1000 + 2\varepsilon x\ , \quad \nabla f_2(x) = -1000 + 2\varepsilon x, \tag{5.1}$$

$$\nabla^2 f_1(x) = 2\varepsilon, \quad \nabla^2 f_2(x) = 2\varepsilon\ , \quad H_1^k = 2\varepsilon(2k-1), \quad H_2^k = 4\varepsilon k. \tag{5.2}$$

Assumption 3.2 is clearly not satisfied, as the gradients of $f_1$ and $f_2$ do not vanish at the optimal solution 0. Rewriting the IN iterations as an inexact perturbed Newton method as in (3.19), we find that

$$x_1^{k+1} = x_1^k - \alpha^k \frac{1}{4\varepsilon k}(\nabla f(x_1^k) + e^k) \tag{5.3}$$

with the gradient error $e^k$ given by the formula (3.22) reducing to

$$\begin{aligned}e^k &= \nabla f_2(x_2^k) - \nabla f_2(x_1^k) + \frac{1}{\alpha^k}\nabla^2 f_2(x_2^k)(x_1^k - x_2^k) \\ &= -\frac{\alpha^k - 1}{2k-1}\nabla f_1(x_1^k) = -\left(\frac{\gamma^k}{2} - \frac{1}{2k}\right)\frac{2k}{2k-1}\nabla f_1(x_1^k)\end{aligned} \tag{5.4}$$

where we used formulas (5.1)–(5.2), the inner update equation (2.2) and the definition of the normalized stepsize $\gamma^k$ in (3.2).

For global convergence, a necessary condition is to have gradient error $e^k \to 0$. From (5.4) and the fact that $\nabla f_1(x_1^k)$ is bounded away from zero around the optimal solution 0, we see that this requires $\gamma^k \to 0$. Hence, we assume $\gamma^k \to 0$. In the special case, if $\alpha^k = 1$ for some $k$, then $e^k = 0$ and the IN iterations (5.3) converges in one cycle, as the quadratic approximations $\tilde{f}_j$ to $f_j$ defined by (2.7) become exact. Assume otherwise that $\alpha^k > 1$ for any $k$, we will show sublinear convergence by a simple classical analysis (the case $\alpha^k < 1$ for all $k$ can be handled similarly). Combining (5.3) and (5.4) and plugging in the formula (5.1) for the gradient of $f_1$, we can express the IN iteration as

$$x_1^{k+1} = \left(1 - \frac{\alpha^k}{2k}\right)\left(1 - \frac{\alpha^k}{2k-1}\right)x_1^k + \frac{1000}{2\varepsilon}\frac{\alpha^k}{2k}\frac{\alpha^k - 1}{2k-1}. \tag{5.5}$$

As $\gamma^k = \alpha^k/k \to 0$, there exists a positive integer $\hat{k}$ such that

$$1 \geq m_k := \left(1 - \frac{\alpha^k}{2k}\right)\left(1 - \frac{\alpha^k}{2k-1}\right) > 0, \quad \text{for all} \quad k \geq \hat{k}. \tag{5.6}$$

Then, from (5.5) and (5.6), for $x_1^{\hat{k}} > 0$ and $k \geq \hat{k}$, we have the lower bounds

$$x_1^{k+1} > (\prod_{j=\bar{k}}^{k} m_j)x_1^{\hat{k}} > 0, \quad x_1^{k+1} > \frac{1000}{2\varepsilon}\frac{\alpha^k}{2k}\frac{\alpha^k - 1}{2k-1} > 0. \tag{5.7}$$

For global convergence, by (5.7), we need $\prod_{k=1}^{\infty} m_k = 0$ because otherwise we would have $\limsup_{k\to\infty}\big(\prod_{j=1}^{k} m_j\big) > \delta$ for some $\delta > 0$ and any $x_1^{\bar{k}}$ satisfying $x_1^{\hat{k}} \geq 1/\delta$ would lead to $\limsup_{k\to\infty} x_1^k \geq 1$ which would be a contradiction with global convergence. Note that

$$\prod_k m_k = 0 \iff \sum_k -\log(m_k) = \infty \iff \sum_k \gamma^k = \infty.$$

where we used $2z \geq -\log(1-z) \geq z$ for $z \geq 0$ around zero and the definition (3.2) of $\gamma^k$. Thus, the sequence $\{\gamma^k\}$, having an infinite sum, cannot decay faster than $1/k^{1+\mu}$ for any $\mu > 0$ and by the lower bound (5.7), convergence to the optimal solution 0 cannot be faster than $O\big(1/k^{2(1+\mu)}\big)$ for any $\mu > 0$ and is thus sublinear.

## 6 Conclusion

We developed and analyzed an incremental version of the Newton method, proving its global convergence with alternative variable stepsize rules under some assumptions.

Furthermore, under a gradient growth assumption, we show that it can achieve linear convergence both under a constant stepsize and a variable stepsize. A by-product of our analysis is the linear convergence of the EKF-S method of [21] under similar assumptions. Our results admit straightforward extensions to incremental quasi-Newton methods and shed light into their convergence properties as well.

## References

1. D. Bertsekas. Incremental least squares methods and the extended Kalman filter. *SIAM Journal on Optimization*, 6(3):807–822, 1996.
2. D. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.
3. D. Bertsekas. *Nonlinear programming.* Athena Scientific, 1999.
4. D. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. *Optimization for Machine Learning*, 2010:1–38, 2011.
5. D. Bertsekas. *Convex Optimization Algorithms.* Athena Scientific, 2015.
6. D. Blatt, A. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.
7. A. Bordes, L. Bottou, and P. Gallinari. SGD-QN: Careful quasi-Newton stochastic gradient descent. *The Journal of Machine Learning Research*, 10:1737–1754, 2009.
8. L. Bottou. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, 2010.
9. L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.
10. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
11. R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A Stochastic Quasi-Newton Method for Large-Scale Optimization. *arXiv preprint arXiv:1401.7020*, 2014.
12. E. Cătinaş. Inexact perturbed Newton methods and applications to a class of Krylov solvers. *Journal of Optimization Theory and Applications*, 108(3):543–570, 2001.
13. W.C. Davidon. New least-square algorithms. *Journal of Optimization Theory and Applications*, 18(2):187–197, 1976.
14. A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA:A fast incremental gradient method with support for non-strongly convex composite objectives. *arXiv preprint arXiv:1407.0202*, 2014.
15. R. Dembo, S. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM Journal on Numerical Analysis*, 19(2):400–408, 1982.
16. J. E. Dennis and J. J. Moré. A characterization of superlinear convergence and its application to quasi-newton methods. *Mathematics of Computation*, 28(126):549–560, 1974.
17. J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.
18. Julien Mairal. Optimization with First-Order Surrogate Functions. In *ICML*, volume 28 of *JMLR Proceedings*, pages 783–791, Atlanta, United States, 2013.
19. O.L. Mangasarian and M.V. Solodov. Serial and parallel backpropagation convergence via nonmonotone perturbed minimization. *Optimization Methods and Software*, 4(2):103–116, 1994.
20. A. Mokhtari and A. Ribeiro. Res: Regularized Stochastic BFGS algorithm. *arXiv preprint arXiv:1401.7625*, 2014.
21. H. Moriyama, N. Yamashita, and M. Fukushima. The incremental Gauss-Newton algorithm with adaptive stepsize rule. *Computational Optimization and Applications*, 26(2):107–141, 2003.
22. A. Nedić and D. Bertsekas. Convergence rate of incremental subgradient algorithms. In S. Uryasev and P.M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, volume 54 of *Applied Optimization*, pages 223–264. Springer US, 2001.
23. A. Nedić and A. Ozdaglar. On the rate of convergence of distributed subgradient methods for multi-agent optimization. In *Proceedings of IEEE CDC*, pages 4711–4716, 2007.
24. A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
25. Y. Nesterov. *Introductory lectures on convex optimization: a basic course.* Applied Optimization. Springer, Boston, 2004.
26. B. T. Polyak. *Introduction to optimization.* Translations series in mathematics and engineering. Optimization Software, Publications Division, New York, 1987.
27. S.S. Ram, A. Nedic, and V.V. Veeravalli. Stochastic incremental gradient descent for estimation in sensor networks. In *Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on*, pages 582–586, 2007.

28. H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
29. N. L. Roux, M. Schmidt, and F.R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc., 2012.
30. M. Schmidt and N.L. Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.
31. N. Schraudolph, J. Yu, and S. Günter. A stochastic quasi-Newton method for online convex optimization. In *Proceedings of the 11th International Conference Artificial Intelligence and Statistics (AISTATS)*, pages 433–440, 2007.
32. O. Shamir, N. Srebro, and T. Zhang. Communication efficient distributed optimization using an approximate Newton-type method. *ICML*, 32(1):1000–1008, 2014.
33. J. Sohl-Dickstein, B. Poole, and S. Ganguli. Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods. In T. Jebara and E. P. Xing, editors, *ICML*, pages 604–612. JMLR Workshop and Conference Proceedings, 2014.
34. M.V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.
35. E.R. Sparks, A. Talwalkar, V. Smith, J. Kottalam, P. Xinghao, J. Gonzalez, M.J. Franklin, M.I Jordan, and T. Kraska. MLI: An API for distributed machine learning. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 1187–1192, 2013.
36. P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.
37. P. Tseng and S. Yun. Incrementally updated gradient methods for constrained and regularized optimization. *Journal of Optimization Theory and Applications*, 160(3):832–853, 2014.
38. T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML, pages 116–, New York, NY, USA, 2004. ACM.