

# A CONVERGENT INCREMENTAL GRADIENT METHOD WITH A CONSTANT STEP SIZE

DORON BLATT<sup>†§</sup>, ALFRED HERO<sup>†§</sup>, AND HILLEL GAUCHMAN<sup>‡</sup>

**Abstract.** An incremental gradient method for minimizing a sum of continuously differentiable functions is presented. The method requires a single gradient evaluation per iteration and uses a constant step size. For the case that the gradient is bounded and Lipschitz continuous, we show that the method visits regions in which the gradient is small infinitely often. Under certain unimodality assumptions, global convergence is established. In the quadratic case, a global linear rate of convergence is shown. The method is applied to distributed optimization problems arising in wireless sensor networks, and numerical experiments compare the new method with the standard incremental gradient method.

**Key words.** incremental gradient method, convergence analysis, sensor networks, neural networks

**AMS subject classifications.** 90C30, 49M37, 65K05

**1. Introduction.** Consider the unconstrained optimization problem

$$(1.1) \quad \text{minimize} \quad f(x) = \sum_{l=1}^L f_l(x), \quad x \in \mathbb{R}^p,$$

where  $\mathbb{R}^p$  is the  $p$ -dimensional Euclidean space, and  $f_l : \mathbb{R}^p \rightarrow \mathbb{R}$  are continuously differentiable scalar functions on  $\mathbb{R}^p$ . Our interest in this problem stems from optimization problems arising in wireless sensor networks (see e.g. [9, 29, 32, 33]), in which  $f_l(x)$  corresponds to the data collected by the  $l$ th sensor in the network. This problem also arises in neural network training, in which  $f_l(x)$  corresponds to the  $l$ th training data set (see e.g. [7, 14, 15, 23, 24, 22]).

The iterative method proposed and analyzed in this paper for solving (1.1), which we call the *incremental aggregated gradient* (IAG) method, generates a sequence  $\{x^k\}_{k \geq 1}$  as follows. Given arbitrary  $L$  initial points  $x^1, x^2, \dots, x^L$ , an aggregated gradient, denoted by  $d^L$ , is defined as  $\sum_{l=1}^L \nabla f_l(x^l)$ . Possible initializations are discussed in §3. For  $k \geq L$ ,

$$(1.2) \quad x^{k+1} = x^k - \mu \frac{1}{L} d^k,$$

$$(1.3) \quad d^{k+1} = d^k - \nabla f_{(k+1)_L}(x^{k+1-L}) + \nabla f_{(k+1)_L}(x^{k+1}),$$

where  $\mu$  is a positive constant step size chosen small enough to ensure convergence,  $(k)_L$  denotes  $k$  modulo  $L$  with representative class  $\{1, 2, \dots, L\}$ , and the factor  $1/L$  is explicitly included to make the approximate descent direction  $\frac{1}{L}d^k$  comparable in magnitude to the one used in the standard incremental gradient method to be discussed below. Thus, at every iteration a new point  $x^{k+1}$  is generated according to the direction of the aggregated gradient  $d^k$ . Then, only one of the gradient summands

---

<sup>†</sup>Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 (`{blatt, hero}@eecs.umich.edu`).

<sup>‡</sup>Department of Mathematics and Computer Science, Eastern Illinois University, Charleston, IL 61920 (`cfhvg@eiu.edu`).

<sup>§</sup>This work was supported in part by DARPA-MURI grant ARO DAAD 19-02-1-0262 and by NSF contract CCR-0325571.


$\nabla f_{(k+1)L}(x^{k+1})$  is computed to replace the previously computed  $\nabla f_{(k+1)L}(x^{k+1-L})$ . Note that for  $k \geq L$  the IAG iteration (1.2)–(1.3) is equivalent to

$$(1.4) \quad x^{k+1} = x^k - \mu \frac{1}{L} \sum_{l=0}^{L-1} \nabla f_{(k-l)L}(x^{k-l}).$$

The IAG method is related to the large class of incremental gradient methods that has been studied extensively in the literature [8, 14, 15, 16, 18, 21, 22, 24, 38] (see also [19, 28] and references therein for incremental subgradient methods for nondifferentiable convex optimization). The standard incremental gradient method updates  $x^k$  according to

$$(1.5) \quad x^{k+1} = x^k - \mu(k) \nabla f_{(k)L}(x^k),$$

where  $\mu(k)$  is a positive step size, possibly depending on  $k$ . Therefore, it is seen that the principal difference between the two methods is that the standard incremental gradient method uses only one of the components in order to generate an approximate descent direction, whereby the IAG method uses the average of the  $L$  previously computed gradients. This property leads to convergence of the IAG method for fixed and sufficiently small positive step size  $\mu$ . This is as contrasted to the standard incremental gradient method, whose convergence requires that the step size sequence  $\mu(k)$  converge to zero.

Incremental gradient methods are based on the observation that when the iterates are far from the eventual limit, the evaluation of a single gradient component is sufficient for generating an approximate descent direction. Hence, these methods lead to a significant reduction in the amount of required computations per iteration (see e.g. [6] section 1.5.2 and the discussion in [5]). The drawback of these methods, when using a constant step size, is that the iterates converge to a limit cycle and oscillate around a stationary point [21], unless restrictions of the type  $\nabla f_l(x) = 0$ ,  $l = 1, \dots, L$  whenever  $\nabla f(x) = 0$  are imposed [38]. Convergence for a diminishing step size has been established by a number of authors under different conditions [8, 14, 15, 18, 21, 22, 24, 38]. However, a diminishing step size usually leads to slow convergence near the eventual limit and requires exhaustive experimentation to determine how rapidly the step size must decrease in order to prevent scenarios in which the step size becomes too small when the iterates are far from the eventual limit (e.g. determining the constants  $a$  and  $b$  in step sizes of the form  $\mu(k) = a/(k + b)$ ). 

A hybrid between the steepest descent method and the incremental gradient method was studied in [5]. The hybrid method starts as an incremental gradient method and gradually becomes the steepest descent. This method requires a tuning parameter, which controls the transition between the two methods, to gradually increase with  $k$  to ensure convergence. When the tuning parameter increases sufficiently fast with the number of iterations, it is shown that the rate of convergence is linear. However, the question of determining the rate of transition between the two methods still remains. For any fixed value of the tuning parameter, the hybrid method converges to a limit cycle, unless a diminishing step size is used, similar to the standard incremental gradient method.

The choice of the aggregated gradient  $d^k$  (1.3) for generating an approximate descent direction was mentioned in [15] in the context of adaptive step size methods, which require repeated evaluations of either the complete objective function  $f(x)$  or its gradient. This requirement renders the methods proposed in [15] inapplicable to

problems in sensor networks of interest to us or any other applications which require decentralized implementation, as will be explained in §3. In addition, as noted in [40], if  $\nabla f_l(x)$ ,  $l = 1, \dots, L$ , are not necessarily zero whenever  $\nabla f(x) = 0$ , the step size tends to zero, resulting in slow convergence.

The IAG method is closely related to Tseng's incremental gradient with momentum term [40], which is an incremental generalization of Polyak's heavy-ball method [30, p. 65] (also called the steepest descent with momentum term [7, p. 104]). Rewriting Tseng's method's update rule as

$$x^{k+1} = x^k - \mu(k) \sum_{l=0}^k \zeta^l \nabla f_{(k-l)_L}(x^{k-l}),$$

we see from (1.4) that the IAG method is a variation of this method with a truncated sum,  $\zeta = 1$ , and a constant step size. Similar to [15], the step size adaptation rule that leads to convergence in [40] requires repeated evaluations of the complete objective function  $f(x)$  and its gradient. Hence, this method cannot be implemented in a distributed manner either. Furthermore, a linear convergence rate is established only under a certain growth property on the functions' gradients, which requires  $\nabla f_l(x) = 0$ ,  $l = 1, \dots, L$ , whenever  $\nabla f(x) = 0$ .

In contrast to the available methods, the IAG method has all four of the following properties: (a) it evaluates a single gradient per iteration, (b) it uses a constant step size, (c) it is convergent (Proposition 2.7), and (d) it has global linear convergence rate for quadratic objective  $f(x)$  (Proposition 2.8).

Finally, we note that the IAG method is reminiscent of other methods in various optimization problems, such as the incremental version of the Gauss-Newton method or the extended Kalman filter [2, 4, 13, 26], the distributed EM algorithm for maximum likelihood estimation [27, 29], the ordered subset and incremental optimization transfer for image reconstruction [1, 3, 10], and the block iterative method for the convex feasibility problem [11].

**2. Convergence Analysis.** In this section we present convergence proofs for two different function classes: (I) restricted Lipschitz and (II) quadratic. Under a Lipschitz condition and a bounded gradient assumption on  $f_l(x)$ ,  $l = 1, \dots, L$  (Assumptions 1 and 2), we obtain an upper bound on the limit inferior of  $\|\nabla f(x^k)\|$ , which depends linearly on the step size  $\mu$ . By imposing additional restrictions on the function  $f(x)$  (Assumptions 3 and 4), we prove pointwise convergence of the method. There are many functions that satisfy Assumptions 1–4. However, one important case does not satisfy these assumptions. This is the case when  $f(x)$  and  $f_l(x)$  are quadratic functions on  $\mathbb{R}^p$ . For this important case we provide a completely different convergence proof and show in addition that the convergence rate is globally linear.

For later reference, it will be useful to write (1.4) in a form known as the “gradient method with errors” [8]:

$$\begin{aligned} x^{k+1} &= x^k - \mu \frac{1}{L} \left[ \sum_{l=0}^{L-1} \nabla f_{(k-l)_L}(x^k) + \sum_{l=0}^{L-1} \nabla f_{(k-l)_L}(x^{k-l}) - \sum_{l=0}^{L-1} \nabla f_{(k-l)_L}(x^k) \right] \\ &= x^k - \mu \frac{1}{L} [\nabla f(x^k) + h^k], \end{aligned}$$

where

$$h^k = \sum_{l=1}^{L-1} [\nabla f_{(k-l)_L}(x^{k-l}) - \nabla f_{(k-l)_L}(x^k)]$$

is the error term in the calculation of the gradient at  $x^k$ . Also note that for all  $k \geq 2L$  and  $1 \leq l \leq L$ ,

$$x^{k-l} - x^k = \mu \frac{1}{L} (d^{k-1} + d^{k-2} + \dots + d^{k-l}).$$

**2.1. Case I.** ASSUMPTION 1.  $\nabla f_l(x)$ ,  $l = 1, \dots, L$ , satisfy a Lipschitz condition in  $\mathbb{R}^p$ , i.e. there is a positive number  $M_1$  such that for all  $x, \bar{x} \in \mathbb{R}^p$ ,  $\|\nabla f_l(x) - \nabla f_l(\bar{x})\| \leq M_1 \|x - \bar{x}\|$ ,  $l = 1, \dots, L$ .

Assumption 1 implies that  $\nabla f(x)$  also satisfies a Lipschitz condition, that is, for all  $x, \bar{x} \in \mathbb{R}^p$ ,  $\|\nabla f(x) - \nabla f(\bar{x})\| \leq M_2 \|x - \bar{x}\|$ , where  $M_2 = LM_1$ .

ASSUMPTION 2. There exists a positive number  $M_3$  such that for all  $x \in \mathbb{R}^p$ ,  $\|\nabla f_l(x)\| \leq M_3$ ,  $l = 1, \dots, L$ .

Assumption 2 implies that for all  $x \in \mathbb{R}^p$ ,  $\|\nabla f(x)\| \leq M_4$ , where  $M_4 = LM_3$ .

LEMMA 2.1. Let  $\{s_k\}_{k \geq 1}$  be a sequence of non-negative real numbers satisfying for some fixed integer  $L > 1$  and all  $k \geq L$

$$s_k \leq cQ(s_{k-1}, s_{k-2}, \dots, s_{k-L+1}) + M,$$

where  $0 < c < 1$ ,  $M$  is nonnegative, and  $Q(s_{k-1}, s_{k-2}, \dots, s_{k-L+1})$  is a linear form in the variables  $s_{k-1}, s_{k-2}, \dots, s_{k-L+1}$ , whose coefficients are non-negative and the sum of the coefficients equals one. Then,  $\limsup_{k \rightarrow \infty} s_k \leq \frac{M}{1-c}$ .

*Proof.* Define the sequence  $\{w_k\}_{k \geq 1}$  by  $w_k = s_k$  for  $1 \leq k \leq L-1$  and

$$w_k = cQ(w_{k-1}, w_{k-2}, \dots, w_{k-L+1}) + M,$$

for  $k \geq L$ . Since  $s_k \leq w_k$  for all  $k$ , if  $\lim_{k \rightarrow \infty} w_k = \frac{M}{1-c}$  then

$$\limsup_{k \rightarrow \infty} s_k \leq \limsup_{k \rightarrow \infty} w_k = \lim_{k \rightarrow \infty} w_k = \frac{M}{1-c}.$$

To show that  $\lim_{k \rightarrow \infty} w_k = \frac{M}{1-c}$ , define the sequence  $\{v_k\}_{k \geq 1}$  by  $v_k = s_k - \frac{M}{1-c}$  for  $1 \leq k \leq L-1$  and

$$v_k = cQ(v_{k-1}, v_{k-2}, \dots, v_{k-L+1}),$$

for  $k \geq L$ . By this construction,

$$\begin{aligned} w_L &= cQ\left(\frac{M}{1-c} + v_{L-1}, \frac{M}{1-c} + v_{L-2}, \dots, \frac{M}{1-c} + v_1\right) + M \\ &= c\frac{M}{1-c} + cQ(v_{L-1}, v_{L-2}, \dots, v_1) + M = \frac{M}{1-c} + v_L, \end{aligned}$$

and, by induction,  $w_k = \frac{M}{1-c} + v_k$  for all  $k > L$ . Therefore, if  $\lim_{k \rightarrow \infty} v_k = 0$  then  $\lim_{k \rightarrow \infty} w_k = \frac{M}{1-c}$ . To show that  $\lim_{k \rightarrow \infty} v_k = 0$ , set  $A = \max\{|v_1|, |v_2|, \dots, |v_{L-1}|\}$ . Hence,

$$|v_L| = c|Q(v_{L-1}, v_{L-2}, \dots, v_1)| \leq cQ(|v_{L-1}|, |v_{L-2}|, \dots, |v_1|) \leq cA.$$

Similarly,  $|v_{L+1}| \leq cA$ , and in general  $|v_k| \leq cA$  for all  $k \geq L$ . Consider now  $v_{2L}$ . Since  $\max\{|v_{2L-1}|, |v_{2L-2}|, \dots, |v_{L+1}|\} \leq cA$ , we have

$$|v_{2L}| = c|Q(v_{2L-1}, v_{2L-2}, \dots, v_{L+1})| \leq cQ(|v_{2L-1}|, |v_{2L-2}|, \dots, |v_{L+1}|) \leq c^2A,$$

and in general  $|v_k| \leq c^2A$  for all  $k \geq 2L$ . Similarly, we obtain  $|v_k| \leq c^nL$  for all  $k \geq nL$ . Since  $0 < c < 1$ , we have  $\lim_{n \rightarrow \infty} c^n = 0$ , and therefore  $\lim_{k \rightarrow \infty} v_k = 0$ .  $\square$

REMARK 1. *Lemma 2.1 can also be proven using concepts from dynamical systems. The sequence  $w_k$  is the output of an autoregressive linear system*

$$w_k = c \sum_{l=1}^{L-1} \alpha_k w_{k-l} + Mu(k-L),$$

where  $u(k)$  is the unit step function which equals one when  $k \geq 0$  and zero otherwise, with initial condition  $w_k = s_k$  for  $1 \leq k \leq L-1$ . Since the coefficients of the linear form are all positive and sum to one, and  $0 < c < 1$ , it is possible to show that the system is stable (bounded input bounded output) and the steady state response is  $\frac{M}{1-c}$  [31], i.e.,  $\lim_{k \rightarrow \infty} w_k = \frac{M}{1-c}$ .

LEMMA 2.2. *Under Assumption 1, if  $\|\nabla f(x^k)\| > \frac{\|h^k\|}{1-2\mu M_1}$ , and  $0 < 1 - 2\mu M_1 < 1$ , then  $f(x^k) > f(x^{k+1})$ .*

*Proof.* Assume that  $\|\nabla f(x^k)\| > \frac{\|h^k\|}{1-2\mu M_1}$ . Then

$$\begin{aligned} \|d^k\|^2 &= \|\nabla f(x^k) + h^k\|^2 \leq 2\|\nabla f(x^k)\|^2 + 2\|h^k\|^2 \\ &< 2\|\nabla f(x^k)\|^2 + 2\frac{\|h^k\|^2}{1-2\mu M_1} < 4\|\nabla f(x^k)\|^2. \end{aligned}$$

By [6, Prop. A.24], if Assumption 1 holds, then

$$f(x+y) - f(x) \leq y' \nabla f(x) + \frac{1}{2} M_2 \|y\|^2.$$

Hence

$$\begin{aligned} f(x^k) - f(x^{k+1}) &= f(x^k) - f(x^k - \mu \frac{1}{L} d^k) \\ &\geq \mu \frac{1}{L} d^{k'} \nabla f(x^k) - \frac{1}{2} M_2 \mu^2 \frac{1}{L^2} \|d^k\|^2 \\ &> \mu \frac{1}{L} (\nabla f(x^k) + h^k)' \nabla f(x^k) - \frac{1}{2} M_2 \mu^2 \frac{1}{L^2} 4\|\nabla f(x^k)\|^2 \\ &= \mu \frac{1}{L} \|\nabla f(x^k)\|^2 + \mu \frac{1}{L} h^{k'} \nabla f(x^k) - 2M_2 \mu^2 \frac{1}{L^2} \|\nabla f(x^k)\|^2 \\ &\geq \mu \frac{1}{L} \|\nabla f(x^k)\|^2 - \mu \frac{1}{L} \|h^k\| \cdot \|\nabla f(x^k)\| - 2M_2 \mu^2 \frac{1}{L^2} \|\nabla f(x^k)\|^2 \\ &= \frac{\mu}{L} \|\nabla f(x^k)\| (1 - 2\mu M_1) \left( \|\nabla f(x^k)\| - \frac{\|h^k\|}{1 - 2\mu M_1} \right) \\ &> 0. \quad \square \end{aligned}$$

Set  $\delta_0 = \mu M_2 M_3$ .

LEMMA 2.3. *Under Assumptions 1 and 2, if  $\mu M_2 < 1$ , there exists  $K$  such that for all  $k > K$ ,  $\|h^k\| < \delta_0$ .*

*Proof.*

$$\begin{aligned}
\|h^k\| &\leq \sum_{l=1}^{L-1} \|\nabla f_{(k-l)_L}(x^{k-l}) - \nabla f_{(k-l)_L}(x^k)\| \\
&\leq M_1 \sum_{l=1}^{L-1} \|x^{k-l} - x^k\| \\
&= \mu M_1 \frac{1}{L} \sum_{l=1}^{L-1} \|d^{k-1} + d^{k-2} + \dots + d^{k-l}\| \\
&\leq \mu M_1 \frac{1}{L} \sum_{l=1}^{L-1} (\|d^{k-1}\| + \|d^{k-2}\| + \dots + \|d^{k-l}\|) \\
&= \mu M_1 \frac{1}{L} [(L-1)\|d^{k-1}\| + (L-2)\|d^{k-2}\| + \dots + \|d^{k-L+1}\|] \\
&= \mu M_1 \frac{1}{L} \frac{L(L-1)}{2} \left[ \frac{(L-1)\|d^{k-1}\| + (L-2)\|d^{k-2}\| + \dots + \|d^{k-L+1}\|}{L(L-1)/2} \right] \\
&= \mu M_1 \frac{L-1}{2} Q(\|d^{k-1}\|, \|d^{k-2}\|, \dots, \|d^{k-L+1}\|),
\end{aligned}$$

where  $Q(\|d^{k-1}\|, \|d^{k-2}\|, \dots, \|d^{k-L+1}\|)$  is a linear form in the variables  $\|d^{k-1}\|, \|d^{k-2}\|, \dots, \|d^{k-L+1}\|$  whose coefficients,  $\frac{L-1}{L(L-1)/2}, \frac{L-2}{L(L-1)/2}, \dots, \frac{1}{L(L-1)/2}$ , sum to one. Next we use  $\|d^k\| = \|\nabla f(x^k) + h^k\| \leq \|\nabla f(x^k)\| + \|h^k\|$  to obtain

$$\begin{aligned}
\|h^k\| &\leq \mu M_1 \frac{L-1}{2} Q(\|h^{k-1}\|, \|h^{k-2}\|, \dots, \|h^{k-L+1}\|) \\
&\quad + \mu M_1 \frac{L-1}{2} Q(\|\nabla f(x^{k-1})\|, \|\nabla f(x^{k-2})\|, \dots, \|\nabla f(x^{k-L+1})\|) \\
&\leq \mu M_1 \frac{L-1}{2} Q(\|h^{k-1}\|, \|h^{k-2}\|, \dots, \|h^{k-L+1}\|) + \mu M_1 \frac{L-1}{2} M_3 \\
&< \mu \frac{M_2}{2} Q(\|h^{k-1}\|, \|h^{k-2}\|, \dots, \|h^{k-L+1}\|) + \mu \frac{M_2}{2} M_3,
\end{aligned}$$

where Assumption 2 was used in the second to last inequality. Hence, by Lemma 2.1, since  $0 < \mu \frac{M_2}{2} < 1/2$ ,  $\limsup_{k \rightarrow \infty} \|h^k\| \leq \frac{\mu \frac{M_2}{2} M_3}{1 - \mu \frac{M_2}{2}}$ . By using  $\mu \frac{M_2}{2} < 1/2$ , we obtain  $\limsup_{k \rightarrow \infty} \|h^k\| < \mu M_2 M_3$  and the lemma follows.  $\square$

PROPOSITION 2.4. *Under Assumptions 1 and 2, if  $f$  is bounded from below and  $\mu \max\{2M_1, M_2\} < 1$  then,*

$$\liminf_{k \rightarrow \infty} \|\nabla f(x^k)\| \leq \frac{2\delta_0}{1 - 2\mu M_1}.$$

*Proof.* Assume the contrary; that is  $\liminf_{k \rightarrow \infty} \|\nabla f(x^k)\| > \frac{2\delta_0}{1 - 2\mu M_1}$ . Then there exists  $K_1$  such that  $\|\nabla f(x^k)\| > \frac{2\delta_0}{1 - 2\mu M_1}$  for all  $k > K_1$ . By Lemma 2.3, there exists  $K_2$  such that  $\|h^k\| < \delta_0$  for all  $k > K_2$ . Therefore, at all iterations for which  $k > \max\{K_1, K_2\}$ ,  $\|\nabla f(x^k)\| > \frac{2\|h^k\|}{1 - 2\mu M_1} \geq \frac{\|h^k\|}{1 - 2\mu M_1}$ . By Lemma 2.2, the sequence  $\{f(x^k)\}_{k=n_1}^\infty$  is decreasing. Since it is bounded from below, there exists  $\lim_{k \rightarrow \infty} f(x^k)$ . In the proof of Lemma 2.2 we showed that

$$f(x^k) - f(x^{k+1}) \geq \frac{\mu}{L} (1 - 2\mu M_1) \|\nabla f(x^k)\| \left[ \|\nabla f(x^k)\| - \frac{\|h^k\|}{1 - 2\mu M_1} \right].$$

Taking limit when  $k \rightarrow \infty$  of both parts of this inequality, we obtain that

$$\lim_{k \rightarrow \infty} \|\nabla f(x^k)\| \left[ \|\nabla f(x^k)\| - \frac{\|h^k\|}{1 - 2\mu M_1} \right] = 0.$$

But this is impossible since  $\|\nabla f(x^k)\| > \frac{2\delta_0}{1 - 2\mu M_1}$  and

$$\|\nabla f(x^k)\| - \frac{\|h^k\|}{1 - 2\mu M_1} > \frac{2\delta_0}{1 - 2\mu M_1} - \frac{\delta_0}{1 - 2\mu M_1} = \frac{\delta_0}{1 - 2\mu M_1},$$

for all  $k \geq \max\{K_1, K_2\}$ .  $\square$

Proposition 2.4 asserts that when  $\mu$  is sufficiently small, the method is guaranteed to visit regions in which  $\|\nabla f(x)\|$  is small (proportional to  $\mu$ ) infinitely often. This type of result has been established for the incremental gradient method with a step size converging to a positive limit [38, Th. 2.1], and for the incremental subgradient method with a constant step size, in the case where  $f_l(x)$ ,  $l = 1, \dots, L$ , are not differentiable but convex [28, Prop. 2.1(b)]. Next, by imposing two additional assumptions, we prove that the IAG method converges with a constant step size to the minimum point of  $f$ .

**ASSUMPTION 3.**  *$f(x)$  has a unique global minimum at  $x^*$ . The Hessian  $\nabla^2 f(x)$  is continuous and positive definite at  $x^*$ .*

**ASSUMPTION 4.** *For any sequence  $\{t^k\}_{k=1}^\infty$  in  $\mathbb{R}^p$ , if  $\lim_{k \rightarrow \infty} f(t^k) = f(x^*)$  or  $\lim_{k \rightarrow \infty} \|\nabla f(t^k)\| = 0$ , then  $\lim_{k \rightarrow \infty} t^k = x^*$ .*

There is an equivalent form of Assumption 4: For each neighborhood  $\mathcal{U}$  of  $x^*$  there exists  $\eta > 0$  such that if  $f(x) - f(x^*) < \eta$  or  $\|\nabla f(x)\| < \eta$ , then  $x \in \mathcal{U}$ .

**REMARK 2.** *Assumptions 3 and 4 are stronger than the assumptions usually made on  $f(x)$  in the literature (see [8] for a summary of the available convergence proofs and the assumptions they require). However, our results hold for a constant step size and do not require that  $\nabla f_l(x) = 0$ ,  $l = 1, \dots, L$ , whenever  $\nabla f(x) = 0$ . In addition, note that there are non-convex functions that satisfy Assumption 4. However, if  $f$  is strictly convex, then Assumption 4 is automatically satisfied. In fact, the implication  $\lim_{k \rightarrow \infty} f(t^k) = f(x^*) \Rightarrow \lim_{k \rightarrow \infty} t^k = x^*$  is the statement of Corollary 27.2.2 from [36]. The implication  $\lim_{k \rightarrow \infty} \|\nabla f(t^k)\| = 0 \Rightarrow \lim_{k \rightarrow \infty} t^k = x^*$  can be obtained as follows: Consider the function  $\nabla f : \mathbb{R}^p \rightarrow \mathbb{R}^p$ . The derivative  $(\nabla f)'$  of this function is the Hessian  $\nabla^2 f$ . Since  $f$  is strictly convex,  $\det(\nabla f)' \neq 0$ . Therefore, by the Inverse Function Theorem, there are open neighborhoods  $V$  of  $x^* \in \mathbb{R}^p$  and  $W$  of  $0 \in \mathbb{R}^p$  such that  $\nabla f : V \rightarrow W$  has a continuous inverse  $\gamma : W \rightarrow V$ . Let  $\{t\}_{k=1}^\infty$  be a sequence such that  $\lim_{k \rightarrow \infty} \|\nabla f(t^k)\| = 0$ . Then there exists  $k_0$  such that  $\nabla f(t^k) \in W$  for all  $k \geq k_0$ . By Theorem B on page 99 in [35], since  $f$  is strictly convex,  $\nabla f$  is one-to-one, i.e. if  $x \neq y$ , then  $\nabla f(x) \neq \nabla f(y)$ . It follows that  $t^k \in V$  for all  $k \geq k_0$ . Now we have*

$$\begin{aligned} \lim_{k \rightarrow \infty} t^k &= \lim_{k \rightarrow \infty} \gamma(\nabla f(t^k)) \\ &= \gamma\left(\lim_{k \rightarrow \infty} \nabla f(t^k)\right) \\ &= \gamma(0) = x^*. \end{aligned}$$

**LEMMA 2.5.** *Under Assumption 3, there exists a neighborhood  $\mathcal{U}$  of  $x^*$  and positive constants  $A_1, A_2, B_1, B_2$  such that for all  $x \in \mathcal{U}$ ,*

$$(2.1) \quad A_1 \|x - x^*\|^2 \leq f(x) - f(x^*) \leq B_1 \|x - x^*\|^2,$$

$$(2.2) \quad A_2 \|x - x^*\|^2 \leq \|\nabla f(x)\|^2 \leq B_2 \|x - x^*\|^2.$$

*Proof.* Consider a Taylor expansion of  $f(x)$  around  $x^*$ . Since  $\nabla f(x^*) = 0$ , we obtain

$$f(x) = f(x^*) + \frac{1}{2} (x - x^*)' \nabla^2 f(\bar{x}) (x - x^*),$$

where  $\bar{x}$  depends on  $x$ . By the well known extremal property of eigenvalues,

$$\lambda_{\min}(x) \|x - x^*\|^2 \leq (x - x^*)' \nabla^2 f(\bar{x}) (x - x^*) \leq \lambda_{\max}(x) \|x - x^*\|^2,$$

where  $\lambda_{\min}(x)$  and  $\lambda_{\max}(x)$  are the smallest and largest eigenvalues of  $\nabla^2 f(\bar{x})$ , which depend on  $x$  through  $\bar{x}$ . Therefore,

$$\frac{1}{2} \lambda_{\min}(x) \|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2} \lambda_{\max}(x) \|x - x^*\|^2.$$

Since  $\nabla^2 f(x^*) > 0$ ,  $\lambda_{\min}(x^*) > 0$  and  $\lambda_{\max}(x^*) > 0$ . Since  $\lambda_{\min}(x)$  and  $\lambda_{\max}(x)$  are continuous, there is a neighborhood  $\mathcal{U}_1$  of  $x^*$  such that  $\lambda_{\min}(x) \geq 1/2 \lambda_{\min}(x^*)$  and  $\lambda_{\max}(x) \leq 2 \lambda_{\max}(x^*)$  for  $x \in \mathcal{U}_1$ . Denoting  $1/4 \lambda_{\min}(x^*)$  by  $A_1$  and  $\lambda_{\max}(x^*)$  by  $B_1$ , we obtain inequality (2.1) for all  $x \in \mathcal{U}_1$ .

Similarly, considering a Taylor expansion of  $\nabla f(x)$ ,

$$\nabla f(x) = \nabla f(x^*) + \nabla^2 f(\bar{x}) (x - x^*) = \nabla^2 f(\bar{x}) (x - x^*),$$

where each row of  $\nabla^2 f(\bar{x})$  depends on a different  $\bar{x}$ , and by the fact that  $[\nabla^2 f(x^*)]^2$  is also positive definite, we obtain inequality (2.2) for all  $x$  in some neighborhood  $\mathcal{U}_2$  of  $x^*$ . Clearly both inequalities (2.1) and (2.2) are satisfied for all  $x \in \mathcal{U} = \mathcal{U}_1 \cap \mathcal{U}_2$ .  $\square$

Let  $\mathcal{U}$  be a neighborhood of  $x^*$  for which inequalities (2.1) and (2.2) hold. By assumption 4 there exists  $\eta > 0$  such that  $x \in \mathcal{U}$  if  $f(x) - f(x^*) < \eta$  or  $\|\nabla f(x)\| < \eta$ . Set  $M_5 = \max\{3\sqrt{\frac{B_1 B_2}{A_1 A_2}}, \frac{2}{1-2\mu M_1}\}$  and  $\lambda = \mu M_2 M_5$ .

LEMMA 2.6. *Under Assumptions 1, 3, and 4, if there exist positive numbers  $n_1$  and  $\delta$  such that  $\|h^k\| < \delta$  for every  $k \geq n_1$ ,  $3\delta < \eta$ ,  $\frac{9B_1}{A_2} \delta^2 < \eta$ , and  $9\mu M_1 < 1$ , then*

(i) *there exists a number  $k_1$  such that  $\|\nabla f(x^k)\| < M_5 \delta$  and  $\|d^k\| < 2M_5 \delta$  for every  $k \geq k_1$ , and*

(ii) *there exists a number  $n_2$  such that  $\|h^k\| < \lambda \delta$  for every  $k \geq n_2$ .*

*Proof.* First we show that there exists  $k$  such that  $k \geq n_1$  and  $\|\nabla f(x^k)\| < \frac{2\delta}{1-2\mu M_1}$ . In fact, if  $\|\nabla f(x^k)\| \geq \frac{2\delta}{1-2\mu M_1}$  for all  $k \geq n_1$ , then  $\|\nabla f(x^k)\| > \frac{2\|h^k\|}{1-2\mu M_1} \geq \frac{\|h^k\|}{1-2\mu M_1}$  for all  $k \geq n_1$ . By Lemma 2.2, the sequence  $\{f(x^k)\}_{k=n_1}^\infty$  is decreasing. Since it is bounded from below by  $f(x^*)$ , there exists  $\lim_{k \rightarrow \infty} f(x^k)$ . By replacing  $\delta_0$  with  $\delta$  and  $\max\{K_1, K_2\}$  with  $n_1$  at the last argument of the proof of Proposition 2.4, we obtain a contradiction.

Let  $k_1$  be the smallest natural number such that  $k_1 \geq n_1$  and  $\|\nabla f(x^{k_1})\| \leq \frac{2\delta}{1-2\mu M_1}$ . Let  $k_2$  be the smallest natural number such that  $k_2 > k_1$  and  $\|\nabla f(x^{k_2})\| > \frac{2\delta}{1-2\mu M_1}$ . (If  $k_2$  does not exist, i.e. if  $\|\nabla f(x^k)\| \leq \frac{2\delta}{1-2\mu M_1}$  for all  $k \geq k_1$ , the proof of Lemma 2.6 still holds.) Let  $k_3$  be the smallest natural number such that  $k_3 > k_2$  and  $\|\nabla f(x^{k_3})\| \leq \frac{2\delta}{1-2\mu M_1}$ . Let  $k_4$  be the smallest natural number such that  $k_4 > k_3$  and  $\|\nabla f(x^{k_4})\| > \frac{2\delta}{1-2\mu M_1}$ . We define  $k_5, k_6, \dots$  in a similar manner.



For every natural  $m$ ,

$$\|d^{k_{2m}-1}\| \leq \|\nabla f(x^{k_{2m}-1})\| + \|h^{k_{2m}-1}\| \leq \frac{2\delta}{1-2\mu M_1} + \delta \leq \frac{3\delta}{1-2\mu M_1},$$

$$\|x^{k_{2m}} - x^{k_{2m}-1}\| = \mu \frac{1}{L} \|d^{k_{2m}-1}\| \leq \frac{3\mu/L}{1-2\mu M_1} \delta,$$

and

$$\begin{aligned} \|\nabla f(x^{k_{2m}})\| &\leq \|\nabla f(x^{k_{2m}}) - \nabla f(x^{k_{2m}-1})\| + \|\nabla f(x^{k_{2m}-1})\| \\ &\leq M_2 \|x^{k_{2m}} - x^{k_{2m}-1}\| + \frac{2\delta}{1-2\mu M_1} \\ &\leq M_2 \frac{3\mu/L}{1-2\mu M_1} \delta + \frac{2}{1-2\mu M_1} \delta \\ &= \frac{2+3\mu M_1}{1-2\mu M_1} \delta < 3\delta, \end{aligned}$$

where we used  $\mu < \frac{1}{9M_1}$  to obtain the last inequality.

Since  $\|\nabla f(x^{k_{2m}})\| < 3\delta < \eta$ ,  $x^{k_{2m}} \in \mathcal{U}$  and we can use Lemma 2.5. We obtain

$$f(x^{k_{2m}}) - f(x^*) \leq B_1 \|x^{k_{2m}} - x^*\| \leq \frac{B_1}{A_2} \|\nabla f(x^{k_{2m}})\|^2 < \frac{B_1}{A_2} 9\delta^2.$$

Let  $k$  be such that  $k_{2m} \leq k < k_{2m+1}$ . Then, by Lemma 2.2,

$$f(x^k) - f(x^*) < f(x^{k_{2m}}) - f(x^*) < 9 \frac{B_1}{A_2} \delta^2.$$

Since  $f(x^k) - f(x^*) < 9 \frac{B_1}{A_2} \delta^2 < \eta$ ,  $x^k \in \mathcal{U}$ , and we can use Lemma 2.5. We obtain

$$\|\nabla f(x^k)\|^2 \leq B_2 \|x^k - x^*\|^2 \leq \frac{B_2}{A_1} [f(x^k) - f(x^*)] < 9 \frac{B_1 B_2}{A_1 A_2} \delta^2.$$

Thus, if  $k$  satisfies  $k_{2m} \leq k < k_{2m+1}$ , we have  $\|\nabla f(x^k)\| < 3\sqrt{\frac{B_1 B_2}{A_1 A_2}} \delta$ . If  $k$  satisfies  $k_{2m-1} \leq k < k_{2m}$ , we have  $\|\nabla f(x^k)\| < \frac{2}{1-2\mu M_1} \delta$ . Therefore for each  $k \geq k_1$ ,  $\|\nabla f(x^k)\| < M_5 \delta$  and therefore:

$$\|d^k\| \leq \|\nabla f(x^k)\| + \|h^k\| \leq M_5 \delta + \delta < 2M_5 \delta.$$

Thus, if  $k \geq k_1$ , we have

$$(2.3) \quad \begin{aligned} \|\nabla f(x^k)\| &< M_5 \delta \\ \|d^k\| &< 2M_5 \delta. \end{aligned}$$

This proves the first part of the Lemma.

To prove the second part, we take  $n_2 = k_1 + L - 1$ . If  $k \geq n_2$ , then not only  $x^k$  but also  $L - 1$  previous terms of the sequence  $\{x^k\}$  satisfy inequalities (2.3). Therefore,

by following the steps in the proof of Proposition 2.4, we have for  $k \geq n_2$

$$\begin{aligned} \|h^k\| &\leq \mu M_1 \frac{1}{L} \sum_{l=1}^{L-1} (\|d^{k-1}\| + \|d^{k-2}\| + \dots + \|d^{k-l}\|) \\ &< \mu M_1 \frac{1}{L} 2M_5 \delta \sum_{l=1}^{L-1} \sum_{m=1}^l 1 = \mu M_1 \frac{1}{L} 2M_5 \delta \frac{L(L-1)}{2} \\ &< \mu M_2 M_5 \delta = \lambda \delta. \end{aligned}$$

Thus  $\|h^k\| < \lambda \delta$ . This proves the second part of Lemma 2.6.  $\square$

PROPOSITION 2.7. *Under Assumptions 1, 2, 3, and 4,*

*if  $\mu < \min\{\frac{1}{9M_1}, \frac{1}{M_2M_5}, \frac{\eta}{3M_1M_3}, \frac{1}{3M_2M_3} \sqrt{\frac{A_2\eta}{B_1}}\}$ , then  $\lim_{k \rightarrow \infty} x^k = x^*$ .*

*Proof.* We prove Proposition 2.7 by repeated use of Lemma 2.6. We start with  $\delta = \delta_0$ . By applying Lemma 2.3, there exists  $K$  such that for all  $k > K$ ,  $\|h^k\| < \delta_0$ . After applying Lemma 2.6  $r$  times we get a number  $n_r$  such that  $\|h^k\| < \delta_0 \lambda^r$ ,  $\|\nabla f(x^k)\| < M_5 \delta_0 \lambda^r$ , and  $\|d^k\| < 2M_5 \delta_0 \lambda^r$ , for  $k \geq n_r$ . The inequality  $\mu < \frac{1}{M_2M_5}$  is equivalent to  $0 < \lambda < 1$ . Hence,  $\lim_{k \rightarrow \infty} \|h^k\| = 0$ ,  $\lim_{k \rightarrow \infty} \|d^k\| = 0$ , and  $\lim_{k \rightarrow \infty} \|\nabla f(x^k)\| = 0$ , and by Assumption 4,  $\lim_{k \rightarrow \infty} x^k = x^*$ .

Note that the inequality  $\mu < \frac{1}{9M_1}$  was used in the proof of Lemma 2.6, and the inequalities  $\mu < \frac{\eta}{3M_2M_3}$  and  $\mu < \frac{1}{3M_2M_3} \sqrt{\frac{A_2\eta}{B_1}}$  are equivalent to  $3\delta_0 < \eta$  and  $\frac{9B_1}{A_2} \delta_0^2 < \eta$ , respectively.  $\square$

**2.2. Case II: Quadratic Case.** Suppose that the functions  $f_l$ ,  $l = 1, \dots, L$ , have the following form

$$(2.4) \quad f_l(x) = \frac{1}{2} x' Q_l x - c_l' x, \quad l = 1, \dots, L,$$

where  $Q_l$  are given symmetric matrices,  $c_l$  are given vectors, and  $\sum_{l=1}^L Q_l$  is positive definite. Under this assumption, the function  $f(x) = \sum_{l=1}^L f_l(x)$  is strictly convex, has its minimum point at

$$(2.5) \quad x^* = \left( \sum_{l=1}^L Q_l \right)^{-1} \sum_{l=1}^L c_l,$$

and  $x^*$  is the only stationary point of  $f(x)$ .

PROPOSITION 2.8. *For sufficiently small  $\mu$ ,  $\lim_{k \rightarrow \infty} x^k = x^*$  and the rate of convergence of the IAG method (1.4) is linear.*

*Proof.* Plugging (2.4) in (1.4), the IAG method becomes

$$x^{k+1} = x^k - \mu \left[ \sum_{l=0}^{L-1} Q_{(k-l)_L} x^{k-l} - c_{(k-l)_L} \right] = x^k - \mu \sum_{l=0}^{L-1} Q_{(k-l)_L} x^{k-l} + \mu c,$$

where  $c = \sum_{l=1}^L c_l$ , and the factor  $\frac{1}{L}$  was absorbed into  $\mu$  to simplify the notation. Subtracting  $x^*$  (2.5) from both sides and adding and subtracting  $x^*$  inside the parentheses, we obtain

$$x^{k+1} - x^* = x^k - x^* - \mu \sum_{l=0}^{L-1} Q_{(k-l)_L} (x^{k-l} - x^* + x^*) + \mu c.$$

Denoting the error at the  $k$ th iteration by  $e^k = x^k - x^*$  and the substitution of (2.5) for  $x^*$  lead to the following error form

$$e^{k+1} = e^k - \mu \sum_{l=0}^{L-1} Q_{(k-l)L} e^{k-l}.$$

This relation between a new error and the previous errors can be seen as a periodically time varying linear system. To analyze its stability, which will lead to the convergence result, it is useful to consider  $L$  iterations as one iteration [25]. This can be seen as down-sampling the original system by a factor of  $L$ , which leads to a time invariant system of a lower sampling rate. Without loss of generality, consider the case where  $k = NL$  for some integer  $N$ , i.e.  $k+1$  corresponds to the first iteration of a new cycle. In this case we have

$$\begin{aligned} e^{k+1} &= e^k - \mu \sum_{l=0}^{L-1} Q_{(k-l)L} e^{k-l} = e^k - \mu \begin{bmatrix} Q_L & Q_{L-1} & Q_{L-2} & \dots & Q_1 \end{bmatrix} \bar{e}^k \\ &= \begin{bmatrix} I_p - \mu Q_L & -\mu Q_{L-1} & -\mu Q_{L-2} & \dots & -\mu Q_1 \end{bmatrix} \bar{e}^k, \end{aligned}$$

where  $I_p$  is the  $p \times p$  identity matrix and

$$\bar{e}^k = \begin{bmatrix} e^k \\ e^{k-1} \\ \vdots \\ e^{k-L+1} \end{bmatrix}.$$

Similarly,

$$\begin{aligned} e^{k+2} &= e^{k+1} - \mu \sum_{l=0}^{L-1} Q_{(k+1-l)L} e^{k+1-l} \\ &= e^{k+1} - \mu \begin{bmatrix} Q_1 & Q_L & Q_{L-1} & \dots & Q_2 \end{bmatrix} \bar{e}^{k+1} \\ &= \begin{bmatrix} I_p - \mu Q_1 & -\mu Q_L & -\mu Q_{L-1} & \dots & -\mu Q_2 \end{bmatrix} \bar{e}^{k+1}, \end{aligned}$$

and finally

$$\begin{aligned} e^{k+L} &= e^{k+L-1} - \mu \sum_{l=0}^{L-1} Q_{(k+L-1-l)L} e^{k+L-1-l} \\ &= e^{k+L-1} - \mu \begin{bmatrix} Q_{L-1} & Q_{L-2} & Q_{L-3} & \dots & Q_L \end{bmatrix} \bar{e}^{k+L-1} \\ &= \begin{bmatrix} I_p - \mu Q_{L-1} & -\mu Q_{L-2} & -\mu Q_{L-3} & \dots & -\mu Q_L \end{bmatrix} \bar{e}^{k+L-1}. \end{aligned}$$

This leads to the relation

$$\bar{e}^{k+L} = M_L \bar{e}^{k+L-1},$$

where

$$M_L = \begin{bmatrix} I_p - \mu Q_{L-1} & -\mu Q_{L-2} & \dots & -\mu Q_1 & -\mu Q_L \\ I_p & 0_p & \dots & 0_p & 0_p \\ 0_p & I_p & \dots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \dots & I_p & 0_p \end{bmatrix},$$

where  $0_p$  denotes the  $p \times p$  zero matrix. Taking another step we have

$$\bar{e}^{k+L} = M_L M_{L-1} \bar{e}^{k+L-2},$$

where

$$M_{L-1} = \begin{bmatrix} I_p - \mu Q_{L-2} & -\mu Q_{L-3} & \cdots & -\mu Q_L & -\mu Q_{L-1} \\ I_p & 0_p & \cdots & 0_p & 0_p \\ 0_p & I_p & \cdots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \cdots & I_p & 0_p \end{bmatrix},$$

and finally, by induction,

$$\bar{e}^{k+L} = M_L M_{L-1} \cdots M_1 \bar{e}^k,$$

where

$$M_1 = \begin{bmatrix} I_p - \mu Q_L & -\mu Q_{L-1} & \cdots & -\mu Q_2 & -\mu Q_1 \\ I_p & 0_p & \cdots & 0_p & 0_p \\ 0_p & I_p & \cdots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \cdots & I_p & 0_p \end{bmatrix}.$$

Denoting  $M = M_L M_{L-1} \cdots M_1$ , we have  $\bar{e}^{k+L} = M \bar{e}^k$ , and in general  $\bar{e}^{k+nL} = M^n \bar{e}^k$ . Therefore, if for sufficiently small  $\mu > 0$  the eigenvalues of  $M$  are inside the unit circle, then  $\lim_{n \rightarrow \infty} \bar{e}^{k+nL} = 0_{pL \times 1}$ , where  $0_{pL \times 1}$  is a  $pL \times 1$  zero vector, i.e. the method converges to the minimum of the function  $f(x)$  and the convergence rate is linear.

To prove that the eigenvalues of  $M$  are inside the unit circle, set

$$A = \begin{bmatrix} I_p & 0_p & \cdots & 0_p & 0_p \\ I_p & 0_p & \cdots & 0_p & 0_p \\ 0_p & I_p & \cdots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \cdots & I_p & 0_p \end{bmatrix},$$

and

$$B_k = \begin{bmatrix} Q_{(k-1)L} & Q_{(k-2)L} & \cdots & Q_{(k+1)L} & Q_k \\ 0_p & 0_p & \cdots & 0_p & 0_p \\ 0_p & 0_p & \cdots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_p & 0_p & \cdots & 0_p & 0_p \end{bmatrix}, \quad k = 1, \dots, L,$$

so that  $M_k = A - \mu B_k$  and  $M = (A - \mu B_L)(A - \mu B_{L-1}) \cdots (A - \mu B_1)$ . Hence,

$$M = A^L - \mu(B_L A^{L-1} + A B_{L-1} A^{L-2} + A^2 B_{L-2} A^{L-3} + \cdots + A^{L-2} B_2 A + A^{L-1} B_1) + \mu^2 C(\mu),$$

where  $C(\mu)$  is a  $Lp \times Lp$  matrix whose elements are polynomials in  $\mu$ .

Note that pre-multiplying a matrix by  $A$  will duplicate the first row of  $p \times p$  matrices and will shift the rest of the rows down, discarding the last  $p$  rows. Post-multiplying by  $A$  will add the second column of  $p \times p$  matrices to the first one and will shift the rest of the columns to the left, inserting a block of  $p \times p$  zero matrices to the last column. It follows that

$$A^L = \begin{bmatrix} I_p & 0_p & \cdots & 0_p & 0_p \\ I_p & 0_p & \cdots & 0_p & 0_p \\ I_p & 0_p & \cdots & 0_p & 0_p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ I_p & 0_p & \cdots & 0_p & 0_p \end{bmatrix},$$

and

$$A^{L-k} B^k A^{k-1} = \begin{bmatrix} W_1(k) & 0_{(L-k+1)p \times (k-1)p} \\ 0_{(k-1)p \times (L-k+1)p} & 0_{(k-1)p \times (k-1)p} \end{bmatrix},$$

where  $W_1(k)$  is a  $(L-k+1)p \times (L-k+1)p$  matrix whose elements are

$$W_1(k) = \begin{bmatrix} \sum_{l=0}^{k-1} Q_{(l)_L} & Q_{L-1} & \cdots & Q_k \\ \vdots & \vdots & & \vdots \\ \sum_{l=0}^{k-1} Q_{(l)_L} & Q_{L-1} & \cdots & Q_k \end{bmatrix}.$$

Therefore, the characteristic polynomial  $F(\mu, \lambda)$  of  $M$  is

$$F(\mu, \lambda) = \det(M - \lambda I_{Lp}) = \det \left( A^L - \mu \sum_{k=1}^L A^{L-k} B^k A^{k-1} - \lambda I_{Lp} + \mu^2 C(\mu) \right).$$

The first  $p$  columns of  $\left( A^L - \mu \sum_{k=1}^L A^{L-k} B^k A^{k-1} - \lambda I_{Lp} + \mu^2 C(\mu) \right)$  are

$$\begin{bmatrix} (1-\lambda)I_p - \mu[LQ_L + (L-1)Q_1 + \cdots + Q_{L-1}] + \mu^2 C_{11} \\ I_p - \mu[(L-1)Q_L + (L-2)Q_1 + \cdots + Q_{L-2}] + \mu^2 C_{21} \\ I_p - \mu[(L-2)Q_L + (L-3)Q_1 + \cdots + Q_{L-3}] + \mu^2 C_{31} \\ \vdots \\ I_p - \mu(2Q_L + Q_1) + \mu^2 C_{L-1,1} \\ I_p - \mu Q_L + \mu^2 C_{L1} \end{bmatrix},$$

the second  $p$  columns are

$$\begin{bmatrix} -(L-1)\mu Q_{L-1} + \mu^2 C_{12} \\ -(L-1)\mu Q_{L-1} - \lambda I_p + \mu^2 C_{22} \\ -(L-2)\mu Q_{L-1} + \mu^2 C_{32} \\ \vdots \\ -2\mu Q_{L-1} + \mu^2 C_{L-1,2} \\ -\mu Q_{L-1} + \mu^2 C_{L2} \end{bmatrix},$$

the next  $(L-3)p$  columns are

$$\begin{bmatrix} -(L-2)\mu Q_{L-2} + \mu^2 C_{13} & \dots & -2\mu Q_2 + \mu^2 C_{1\ L-1} \\ -(L-2)\mu Q_{L-2} + \mu^2 C_{23} & \dots & -2\mu Q_2 + \mu^2 C_{2\ L-1} \\ -(L-2)\mu Q_{L-2} - \lambda I_p + \mu^2 C_{33} & \dots & -2\mu Q_2 + \mu^2 C_{3\ L-1} \\ \vdots & & \vdots \\ -2\mu Q_{L-2} + \mu^2 C_{L-1\ 3} & \dots & -2\mu Q_2 - \lambda I_p + \mu^2 C_{L-1\ L-1} \\ -\mu Q_{L-2} + \mu^2 C_{L3} & \dots & -\mu Q_2 + \mu^2 C_{L\ L-1} \end{bmatrix},$$

and the last  $p$  columns are

$$\begin{bmatrix} -\mu Q_1 + \mu^2 C_{1L} \\ -\mu Q_1 + \mu^2 C_{2L} \\ -\mu Q_1 + \mu^2 C_{3L} \\ \vdots \\ -\mu Q_1 + \mu^2 C_{L-1\ L} \\ -\mu Q_1 - \lambda I_p + \mu^2 C_{LL} \end{bmatrix},$$

where  $C_{ij}$ ,  $i, j = 1, \dots, L$  are  $p \times p$  matrices whose entrees are polynomials in  $\mu$ .

It is easy to see that if  $\mu = 0$ , then  $F(0, \lambda) = (-1)^{Lp} \lambda^{Lp-p} (\lambda - 1)^p$ . Hence, if  $\mu = 0$ , we have an eigenvalue 0 of multiplicity  $Lp - p$  and an eigenvalue 1 of multiplicity  $p$ . If  $\mu$  is close enough to zero, the 0-eigenvalues will be close to the origin and therefore inside the unit circle. We need to prove that for sufficiently small positive  $\mu$ , all the 1-eigenvalues will be inside the unit circle. Let  $\lambda = \lambda(\mu)$  be a smooth function expressing the dependence of one of the 1-eigenvalues on  $\mu$ . We will prove that  $\frac{d\lambda}{d\mu}(0^+) < 0$ . It will be enough for our purposes since it will show that the trajectory  $\lambda = \lambda(\mu)$  is entering the unit circle, and hence  $\lambda(\mu)$  is inside the unit circle for sufficiently small positive  $\mu$ .

By the definition of  $\lambda(\mu)$ ,  $\lambda(0^+) = 1$  and  $F(\mu, \lambda(\mu)) = 0$  for all  $\mu$ . It follows that

$$(2.6) \quad \frac{d^p F(\mu, \lambda(\mu))}{d\mu^p} = 0.$$

To calculate the left side of (2.6), we use the formula for the derivative of a determinant [20]. Note that substituting  $\mu = 0$  and  $\lambda = 1$  into each of the first  $p$  rows of the matrix  $M - \lambda I_{Lp}$  leads to a row in which all of the entrees are zeros and therefore the determinant has a zero value. Therefore the only non-zero terms in  $\frac{d^p F(\mu, \lambda(\mu))}{d\mu^p}$  after substituting  $\mu = 0$  and  $\lambda = 1$  (more precisely, taking  $\mu \rightarrow 0^+$ ) are the terms with the first derivatives in the first  $p$  rows (there are  $p!$  such terms). Hence taking the  $p$ th derivative is reduced to taking the first derivative of each of the first  $p$  rows. Substituting  $\lambda = 1$  and  $\mu \rightarrow 0^+$  we obtain

$$\frac{d^p F(\mu, \lambda(\mu))}{d\mu^p} = p! \det \begin{bmatrix} W_2 & W_3 \\ W_4 & -I_{(L-1)p \times (L-1)p} \end{bmatrix} = 0,$$

where  $W_2 = -\lambda'(0^+) I_p - \sum_{k=0}^{L-1} (L-k) Q_{(k)_L}$ ,

$$W_3 = \begin{bmatrix} -(L-1)Q_{L-1} & -(L-2)Q_{L-2} & \dots & -2Q_2 & -Q_1 \end{bmatrix},$$

and  $W_4 = [I_p \ I_p \ \dots \ I_p]^T$ . Add all columns of  $p \times p$  matrices to the first column of  $p \times p$  matrices to obtain

$$\det \begin{bmatrix} W_5 & W_3 \\ 0_{(L-1)p \times p} & -I_{(L-1)p \times (L-1)p} \end{bmatrix} = 0,$$

where  $W_5 = -\lambda'(0^+)I_p - L \sum_{k=1}^L Q_k$ . Calculating the last determinant gives

$$\det \left[ L \sum_{k=1}^L Q_k + \lambda'(0^+)I_p \right] = 0.$$

The last equation shows that  $-\lambda'(0^+)$  is an eigenvalue of the matrix  $L \sum_{k=1}^L Q_k$ . Since  $L \sum_{k=1}^L Q_k$  is positive definite,  $-\lambda'(0^+) > 0$  and therefore  $\lambda'(0^+) < 0$ . This proves that for sufficiently small  $\mu > 0$  the eigenvalues of the matrix  $M$  are strictly inside the unit circle and hence the sequence  $x^k$  converges to  $x^*$  and the convergence rate is linear.  $\square$

**3. Initialization and Distributed Implementation.** As mentioned in §1, the IAG method is initiated with  $L$  points,  $x^1, x^2, \dots, x^L$ . Possible initialization strategies include setting  $x^1 = x^2 = \dots = x^L$  or generating the initial points using a single cycle of the standard incremental gradient method (1.5). Another possibility is the following. Given  $x^1$ , compute  $d^1 = \nabla f_1(x^1)$ . Then, for  $1 \leq k \leq L-1$ ,

$$(3.1) \quad \begin{aligned} x^{k+1} &= x^k - \mu \frac{1}{k} d^k, \\ d^{k+1} &= d^k + \nabla f_{(k+1)_L}(x^{k+1}). \end{aligned}$$

Therefore, after  $L-1$  iterations we obtain  $x^1, \dots, x^L$  and  $d^L = \sum_{l=1}^L \nabla f_l(x^l)$ .

The key feature of the IAG method that makes it suitable for wireless sensor networks applications is that it can be implemented in a distributed manner. Consider a distributed system of  $L$  processors enumerated over  $1, 2, \dots, L$ , each of which has access to one of the functions  $f_l(x)$ . The initialization (3.1) begins with  $x^1$  at processor 1. Then, processor 1 sets  $d^1 = \nabla f_1(x^1)$  and transmits  $x^1$  and  $d^1$  to processor 2. Upon receiving  $x^{k-1}$  and  $d^{k-1}$  from processor  $k-1$ , processor  $k$  calculates  $x^k$  and  $d^k$  according to (3.1) and transmits them to processor  $k+1$ . The initialization phase is completed when processor  $L$ , upon receiving  $x^{L-1}$  and  $d^{L-1}$  from processor  $L-1$ , computes  $x^L$  and  $d^L$  according to (3.1) and transmits them to processor 1.

Once the initialization phase is completed, the algorithm progresses in a cyclic manner. Upon receiving  $x^{k-1}$  and  $d^{k-1}$  from processor  $(k-1)_L$ , processor  $(k)_L$  computes  $x^k$  and  $d^k$  according to (1.2) and (1.3), respectively, and transmits them to processor  $(k+1)_L$ . Note that  $\nabla f_{(k)_L}(x^{k-L})$  in (1.3) is available at processor  $(k)_L$ , since it was the last gradient computed at that processor. Therefore, the only gradient computation at processor  $(k)_L$  is  $\nabla f_{(k)_L}(x^k)$ . At no phase of the algorithm do the processors share information regarding the complete function  $f(x)$  or its gradient  $\nabla f(x)$ .

**4. Application to Wireless Sensor Networks.** There are two motivations to use the IAG method: (a) reduced computational burden due to the evaluation of a single gradient per iteration compared to  $L$  gradients required for the steepest descent method; and (b) the possibility of a distributed implementation of the method in which each component has access to one of the functions  $f_l(x)$ . The second item has been shown to be very useful in the context of wireless sensor networks [32, 33]. Wireless sensor networks provide means for efficient large scale monitoring of large areas [39]. Often the ultimate goal is to estimate certain parameters based on measurements that the sensors collect, giving rise to an optimization problem. If measurements from distinct sensors are modelled as statistically independent, the estimation problem

takes the form of (1.1), where  $f_l(x)$  is indexed by the measurements available at sensor  $l$  (see e.g. [9, 29, 32, 33] and references therein). When transmitting the complete set of data to a central processor is impractical due to bandwidth and power constraints, the IAG method can be implemented in a distributed manner as described in §3. In the following sections we consider two such estimation problems.

**4.1. Robust Estimation.** One of the benefits of a wireless sensor network is the ability to deploy a large number of low cost sensors to densely monitor a certain area. Because low cost sensors have limited reliability, the system must be designed to be robust to the possibility of individual sensor failures. In estimation tasks, this means that some of the sensors will contribute unreliable measurements, namely outliers. In [32] the authors suggest the use of robust statistics to alleviate the influence of outliers in the data (see [17] or, specifically in the context of optimization, see [30, p. 347]). The robust statistics framework uses objective functions that give less weight to outliers. A common objective function used to this end is the function “Fair” [34, p. 110], given by

$$(4.1) \quad g(x) = c^2 \left[ \frac{|x|}{c} - \log \left( 1 + \frac{|x|}{c} \right) \right].$$

Following [32] we simulate a sensor network for measuring pollution levels and assume that a certain percentage of the sensors are damaged and provide unreliable measurements. Each sensor collects a single noisy measurement of the pollution level and the estimate of the average pollution level is found by minimizing the objective function defined by

$$(4.2) \quad f(x) = \sum_{l=1}^L f_l(x),$$

where  $x \in \mathbb{R}$ , and

$$f_l(x) = \frac{1}{L} g(x - y_l),$$

where  $y_l$  is the measurement collected by sensor  $l$ . There were  $L = 50$  sensors in the simulation. To reflect the possibility of faulty sensors, half of the samples were generated according to a Gaussian distribution with mean  $m_1 = 10$  and unit variance ( $\sigma_1^2 = 1$ ) and the other half were generated according to a Gaussian distribution with mean  $m_2 = 10$  and ten times higher variance ( $\sigma_2^2 = 10$ ). The coefficient  $c$  in (4.1) was chosen to be 10.

The first derivative of  $g$  is  $\frac{x}{1+|x|/c}$  whose magnitude is bounded by  $c$ . The second derivative of  $g$  is  $\frac{1}{(1+|x|/c)^2}$  which is bounded by 1. Hence both Assumptions 1 and 2 hold. In addition, since  $\frac{1}{(1+|x|/c)^2}$  is strictly positive,  $g$  is strictly convex, and therefore Assumptions 3 and 4 hold as well.

Both the standard incremental gradient method (1.5) with a constant step size  $\mu(k) = \mu$  (abbreviated as “IG” in the figures) and the IAG method with the initialization (3.1) were implemented with several choices of step size  $\mu$ . The initial point  $x^1$  was set to 0. In Fig. 4.1 the trajectories of the two methods are presented. The solid straight line corresponds to the minimum point  $x^*$ . It is seen that when the step size is sufficiently small, IAG increases more rapidly towards  $x^*$  than the standard incremental gradient in the early iterations. Furthermore, as predicted by the theory,



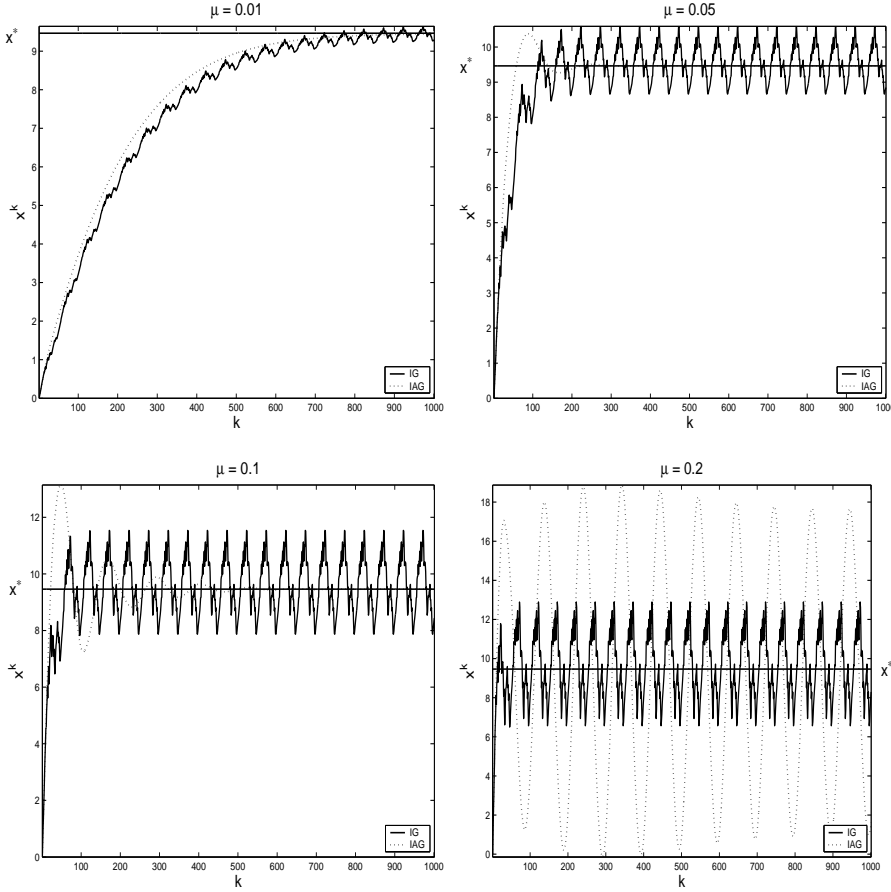


FIG. 4.1. Trajectories taken by the IG and IAG methods for the robust “Fair” estimation problem.

IAG converges to the true limit, whereas incremental gradient method converges to a limit cycle. For a larger step size the IAG method overshoots due to its heavy ball characteristic (1.4). When the step size is too large, the IAG method no longer converges but the incremental gradient method still converges to a limit cycle. We have observed this behavior for other values of the parameters  $m_1$ ,  $m_2$ ,  $\sigma_1^2$ ,  $\sigma_2^2$ ,  $c$  as well.

**4.2. Source Localization.** This section presents a simulation of a sensor network for localizing a source that emits acoustic waves.  $L$  sensors are distributed on the perimeter of a field at known spatial locations, denoted  $r_l$ ,  $l = 1, \dots, L$ , where  $r_l \in \mathbb{R}^2$ . Each sensor collects a noisy measurement of the acoustic signal transmitted by the source, denoted  $y_l$ , at an unknown location  $x$ . Based on a far-field assumption and an isotropic acoustic wave propagation model [12, 32, 37], the problem of estimation of source location can be formulated as a non-linear least squares problem. The objective function is again of the form (4.2), but now

$$(4.3) \quad f_l(x) = (y_l - g(\|r_l - x\|^2))^2,$$

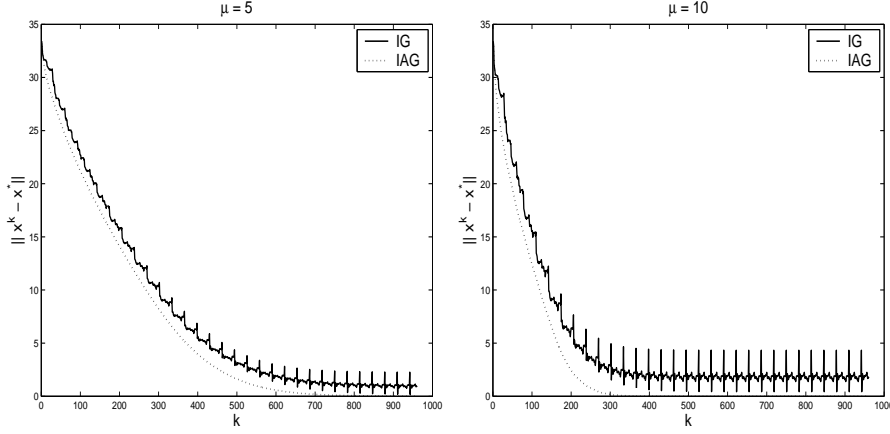


FIG. 4.2. Distance of IG and IAG iterates to the optimal solution  $x^*$  for source localization problem.

$x \in \mathbb{R}^2$ , and

$$(4.4) \quad g(z) = \begin{cases} A/z & : z \geq A/\epsilon \\ 2\epsilon - \epsilon^2 z/A & : z < A/\epsilon \end{cases}.$$

In (4.3)  $g(\cdot)$  models the received signal strength as a function of the squared distance. In (4.4)  $A$  is a known constant characterizing the source's signal strength. For  $z \geq A/\epsilon$  (far-field source), the source's signal strength has isotropic attenuation as an inverse function of the squared distance, while for  $z < A/\epsilon$  (near-field source), the attenuation is linear in the squared distance. It is easy to see that Assumptions 1 and 2 are satisfied and therefore, Proposition 2.4 holds. Clearly, since  $f(x)$  is multi-modal in this case, Assumptions 3 and 4 cannot hold. However, it was observed in our experiments that when the source is sufficiently distant from the sensors, the objective function has a single minimum inside the observed field (See Fig. 4.3 for a contour plot of the objective function) and, when initiated not too far from the minimum point, the IAG method has good convergence properties. This suggests the possible application of the IAG method under weaker assumptions than those considered in this paper, and motivates further investigation into its properties.

In the numerical experiment,  $L = 32$  sensors are distributed equidistantly on the perimeter of a  $100 \times 100$  field. The source is located at the point  $[60, 60]$  and emits a signal with strength  $A = 1000$ . The sensors' noisy measurements were generated according to a Gaussian distribution with a mean equal to the true signal power and unit variance. Both the incremental gradient method with a constant step size and the IAG method with the initialization (3.1) were initiated at the point  $[40, 40]$ . The error term  $\|x^k - x^*\|$  as a function of the iteration number is presented in Fig. 4.2 for two choices of step size. The actual path taken by the methods for step size  $\mu = 10$  is presented in Fig. 4.3, where the asterisk denotes the true minimum point of the objective function. It is seen that, as the theory predicts, the incremental gradient method exhibits oscillations near the eventual limit, whereas the IAG method converges to the minimum. In this scenario, the IAG method outperforms the IG method at early iterations as well.

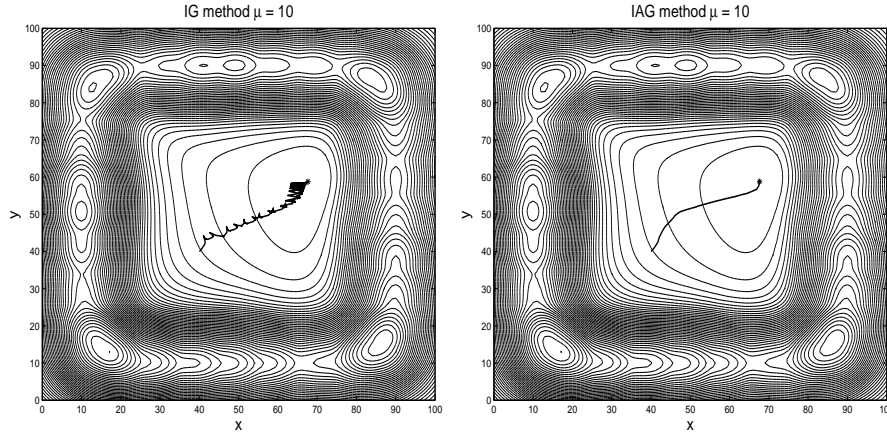


FIG. 4.3. Path taken by the IG and IAG methods for source localization problem.

## REFERENCES

- [1] S. AHN, J. FESSLER, D. BLATT, AND A. HERO, *Incremental optimization transfer algorithms: application to tomography*, Submitted to: IEEE Trans. Image Process., (2004).
- [2] B. M. BELL, *The iterated kalman smoother as a Gauss-Newton method*, SIAM J. Optim., 4 (1994), pp. 626–636.
- [3] A. BEN-TAL, T. MARGALIT, AND A. NEMIROVSKI, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM J. Optim., 12 (2001), pp. 79–108.
- [4] D. P. BERTSEKAS, *Incremental least squares methods and the extended Kalman filter*, SIAM J. Optim., 6 (1996), pp. 807–822.
- [5] ———, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926.
- [6] ———, *Nonlinear programming: second edition*, Athena Scientific, Belmont, MA, 1999.
- [7] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.
- [8] ———, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.
- [9] D. BLATT AND A. HERO, *Distributed maximum likelihood for sensor networks*, in Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, May 2004, pp. 929 – 932.
- [10] C. BYRNE, *Choosing parameters in block-iterative or ordered subset reconstruction algorithms*, IEEE Trans. Image Process., (2004). to appear.
- [11] Y. CENSOR AND G. T. HERMAN, *Block-iterative algorithms with underrelaxed Bregman projections*, SIAM J. Optim., 13 (2002), pp. 283–297.
- [12] J. C. CHEN, K. YAO, AND R. E. HUDSON, *Source localization and beamforming*, IEEE Signal Processing Magazine, 19 (2002), pp. 30–39.
- [13] W. C. DAVIDON, *New least-square algorithms*, J. Optim. Theory Appl., 18 (1976), pp. 187–197.
- [14] A. A. GAIVORONSKI, *Convergence analysis of parallel backpropagation algorithm for neural networks*, Optim. Methods Softw., 4 (1994), pp. 117–134.
- [15] L. GRIPPO, *A class of unconstrained minimization methods for neural networks training*, Optim. Methods Softw., 4 (1994), pp. 135–150.
- [16] ———, *Convergent on-line algorithms for supervised learning in neural networks*, IEEE Trans. Neural Networks, 11 (2000), pp. 1284–1299.
- [17] P. HUBER, *Robust Statistics*, John Wiley & Sons, New York, 1981.
- [18] V. M. KIBARDIN, *Decomposition into functions in the minimization problem*, Autom. Remote Control, 40 (1980), pp. 1311–1321.
- [19] K. C. KIWIEL, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM J. Optim., 14 (2004), pp. 807–840.
- [20] E. KREYSZIC, *Advanced engineering mathematics*, John Wiley & Sons, New York, 1988.
- [21] Z. Q. LUO, *On the convergence of the LMS algorithm with adaptive learning rate for linear*

- feedforward networks*, Neural Comput., 3 (1991), pp. 226–245.
- [22] Z. Q. LUO AND P. TSENG, *Analysis of an approximate gradient projection method with applications to the backpropagation algorithm*, Optim. Methods Softw., 4 (1994), pp. 85–101.
  - [23] O. L. MANGASARIAN, *Mathematical programming in neural networks*, ORSA J. Comput., 5 (1993), pp. 349–360.
  - [24] O. L. MANGASARIAN AND M. V. SOLODOV, *Serial and parallel backpropagation convergence via nonmonotone perturbed minimization*, Optim. Methods Softw., 4 (1994), pp. 103–116.
  - [25] R. MEYER AND C. BURRUS, *A unified analysis of multirate and periodically time-varying digital filters*, IEEE Trans. Circuits Systems, 22 (1975), pp. 162–168.
  - [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.
  - [27] R. M. NEAL AND G. E. HINTON, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, in Learning in Graphical Models, M. I. Jordan, ed., Kluwer Academic Publishers, Dordrecht, 1994, pp. 355–368.
  - [28] A. NEDIC AND D. P. BERTSEKAS, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.
  - [29] R. D. NOWAK, *Distributed EM algorithms for density estimation and clustering in sensor networks*, IEEE Trans. Signal Process., 51 (2003), pp. 2245–2253.
  - [30] B. T. POLYAK, *Introduction to Optimization*, Optimization Software, Inc., New York, 1987.
  - [31] J. G. PROAKIS AND D. G. MANOLAKIS, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1996.
  - [32] M. G. RABBAT AND R. D. NOWAK, *Decentralized source localization and tracking*, in Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, May 2004, pp. 921 – 924.
  - [33] ———, *Distributed optimization in sensor networks*, in Proceedings of the Third International Symposium on Information Processing in Sensor Networks, Berkeley, California, April 2004, ACM Press, New York, pp. 20–27.
  - [34] W. J. J. REY, *Introduction to Robust and Quasi-Robust Statistical Methods*, Springer-Verlag, Berlin, 1983.
  - [35] A. W. ROBERTS AND D. E. VARBERG, *Convex Functions*, Academic Press, New York, 1973.
  - [36] R. T. ROCKAFELLER, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
  - [37] X. SHENG AND Y. H. HU, *Energy based acoustic source localization*, in Information Processing in Sensor Networks, Second International Workshop, IPSN 2003, Z. Feng and G. Leonidas, eds., vol. 2634 of Lecture Notes in Computer Science, Palo Alto, California, April 2003, Springer-Verlag, New York, pp. 285–300.
  - [38] M. V. SOLODOV, *Incremental gradient algorithms with stepsizes bounded away from zero*, Comput. Optim. Appl., 11 (1998), pp. 23–35.
  - [39] R. SZEWCZYK, E. OSTERWEIL, J. POLASTRE, M. HAMILTON, A. MAINWARING, AND D. ESTRIN, *Habitat monitoring with sensor networks*, Commun. ACM, 47 (2004), pp. 34–40.
  - [40] P. TSENG, *An incremental gradient(-projection) method with momentum term and adaptive stepsize rule*, SIAM J. Optim., 8 (1998), pp. 506–531.