Pergamon

Computers Math. Applic. Vol. 30, No. 1, pp. 79–90, 1995

0898-1221/95 $9.50 + 0.00

0898-1221(95)00069-0

Monotone Discrete Newton Iterations and Elimination

J. P. MILASZEWICZ
Departamento de Matemática
Facultad de Ciencias Exactas y Naturales
Ciudad Universitaria, 1428 Buenos Aires, Argentina

(Received September 1994; accepted October 1994)

Abstract—The improvement in convergence by means of accurate functional elimination in the context of the monotone Newton theorem is further analyzed and extended to discrete approximations of the Newton method.

Keywords—Nonlinear systems, Order convex functions, Discretized Newton method, Functional elimination.

1. INTRODUCTION

In the discretization of mildly nonlinear elliptic problems, it is usually necessary to solve nonlinear systems that satisfy the hypotheses in the monotone Newton theorem; such systems also arise in the discretization of one-dimensional nonlinear boundary value problems. The theorem essentially states that, with two convenient starting points, it is possible to generate two monotone sequences, enclosing a root of the system with both converging quadratically to it [1]. One of the hypotheses in the monotone Newton theorem is that the Jacobian matrix is an M-matrix. For systems involving such matrices, it was shown that partial elimination is a useful preconditioner for Jacobi and Gauss-Seidel iterations [2]. A similar conclusion holds for fixed point linear equations for which the iteration function has a nonnegative Jacobian [3]. Here, the counterpart of linear elimination is functional substitution. More interestingly, in the framework of the monotone Newton theorem, it has been shown that accurate partial functional elimination (i.e., an unknown is eliminated by means of the equation with the same index) produces a reduced system that inherits the properties of the original one. Moreover, the corresponding monotone sequences with projected starting points converge termwise faster than those in the original sequences. This improvement is extended to the eliminated coordinate via functional evaluation [3]. The process can be repeated, inductively, in order to eliminate a set of unknowns by means of the equations with the same set of indexes. In this note, extensions of these results for a derivative free version of the monotone Newton theorem are proven and illustrated with a numerical example. They are based on a thorough analysis of the results in [3].

2. A DISCRETE MONOTONE NEWTON THEOREM

Consider a continuously differentiable function $F : D \subset \Re^n \longrightarrow \Re^n$ and the equation

$$Fx = 0. \tag{2.1}$$

It will be assumed throughout that we have $x^0 \leq y^0$; i.e., $x_i^0 \leq y_i^0$, for $1 \leq i \leq n$, such that

$$\langle x^0, y^0 \rangle := \left\{ \frac{x \in \Re^n}{x^0 \leq x \leq y^0} \right\} \subset D, \quad \text{and} \quad Fx^0 \leq 0 \leq Fy^0.$$

It will also be assumed that F' is isotone; i.e.,

$$x \leq y \qquad \text{implies } F'(x) \leq F'(y),$$

and that it is Lipschitz continuous, namely

$$\|F'(x) - F'(y)\| \leq \gamma \|x - y\|, \qquad \forall\, x, y \in D.$$

Recall that if F' is isotone, then F is order convex; i.e.,

$$F(\lambda x + (1-\lambda)y) \leq \lambda Fx + (1-\lambda)Fy,$$

for $x \leq y$ or $y \leq x$ and $\lambda \in (0,1)$. Notice that for $n > 1$, order convexity is a weaker notion than convexity.

Now consider mappings $P_k, Q_k : \langle x^0, y^0 \rangle \longrightarrow L(\Re^n)$, such that for each $y \in \langle x^0, y^0 \rangle$, $P_k(y)$ and $Q_k(y)$ are nonnegative subinverses of $F'(y)$. The following theorem can be proven [1].

THEOREM 2.1. *The sequence*

$$y^{k+1} := y^k - P_k(y^k)Fy^k, \qquad k = 0, 1, \ldots \tag{2.2}$$

is well defined and it satisfies $y^k \downarrow y^* \in \langle x^0, y^0 \rangle$ *as* $k \to \infty$ *and* y^* *is the unique solution of* (2.1) *in* $\langle x^0, y^0 \rangle$. *Moreover, the subsidiary sequence*

$$x^{k+1} := x^k - Q_k(y^k)Fx^k, \qquad k = 0, 1, \ldots \tag{2.3}$$

is also well defined and it satisfies $x^k \uparrow y^*$ *as* $k \to \infty$.
The iterates satisfy

$$Fx^k \leq 0 \leq Fy^k, \qquad k = 0, 1, \ldots. \tag{2.4}$$

Finally, if for all $y \in D$, $F'(y)$ *is nonsingular with nonnegative inverse, and it is defined*

$$Q_k\left(y^k\right) := P_k\left(y^k\right) := F'\left(y^k\right)^{-1}, \tag{2.5}$$

then there exists a constant c *such that*

$$\left\|y^{k+1} - x^{k+1}\right\| \leq c \left\|x^k - y^k\right\|^2, \qquad k = 0, 1, \ldots. \tag{2.6}$$

REMARK 2.2. It has been Fourier who, for $n = 1$, realized that the iterates in (2.3) with the choice (2.5) generate a complementary sequence to the one defined by (2.2) [4]. Theorem 2.1, with the choice (2.5) and exclusive of (2.6), has been obtained by Baluev [5], while the general version presented here, together with (2.6), has been proven by Ortega and Rheinboldt [6]. It is this general version which we shall apply in this paper; the hypotheses in Theorem 2.1 will be assumed throughout. There are three more possible variants of Theorem 2.1. They are obtained by interchanging the roles of x^0 and y^0, and also by supposing that F' is antitone (i.e., $x \leq y$ implies $F'(x) \geq F'(y)$), which implies that F is order concave (see [1, Table 13.1]). The terms generated by (2.3) and (2.2) with the choice (2.5), while setting $y_N^0 := y^0$ and $x_N^0 := x^0$, will be denoted respectively y_N^k, x_N^k, and referred to as the Newton-Fourier (N-F) iterates.

In the context given by Theorem 2.1, it may be desirable to avoid the calculation of $F'(y)$, while retaining monotone and superlinear convergence for the corresponding iterates. It will now

be shown that convenient approximations of the Jacobian matrix provide a satisfactory solution. F' will be assumed to be continuous.

Consider $h \in \Re$, $h > 0$, such that

$$y + h\, e \in D, \qquad \forall\, y \in \langle x^0, y^0 \rangle, \qquad e := (1, \ldots, 1)^t,$$

and define

$$[J(y,h)]_{i,j} := \frac{1}{h}\left[f_i(y + h\, e^j) - f_i(y)\right], \qquad 1 \le i, j \le n, \tag{2.7}$$

where e^j is the j^{th} unit coordinate vector. Note that there exists $h_0 > 0$ such that for $h \le h_0$, $J(y,h)$ is well defined and nonsingular whenever $y \in \langle x^0, y^0 \rangle$. It will throughout be assumed that $y \in \langle x^0, y^0 \rangle$ and that $0 < h \le h_0$.

LEMMA 2.3. *The following propositions hold:*

(i) $F'(y) \le J(y,h) \le F'(y + h\, e), \quad \forall\, y \in \langle x^0, y^0 \rangle.$
(ii) $J(x,h) \le J(y,h), \quad \forall x, y \in \langle x^0, y^0 \rangle, \quad x \le y.$

PROOF. It follows from the isotonocity of F', when applied to

$$[J(y,h)]_{i,j} = \int_0^1 \partial_j f_i\left(y + th\, e^j\right) dt.$$

COROLLARY 2.4. *$J(y,h)^{-1}$ is a nonnegative subinverse of $F'(y)$. If $F'(y)$ is an M-matrix, then the same holds for $J(y,h)$. Moreover, if $F'(y^0)$ is also irreducible, then there exists $h_0' > 0$ such that if $0 < h \le h_0'$, then $J(y,h)$ is irreducible for all $y \in \langle x^0, y^0 \rangle$.*

PROOF. For the last part, it is only necessary to note that the set of irreducible matrices is an open set.

REMARK 2.5. Following [1], instead of (2.7), it could have been considered the more general

$$[J(y,h)]_{i,j} := \frac{1}{h_{ij}}\left[f_i\left(y + \beta \sum_{k=1}^{j-1} h_{ik}\, e^k + h_{ij}\, e^j\right) - f_i\left(y + \beta \sum_{k=1}^{j-1} h_{ik}\, e^k\right)\right], \qquad 1 \le i \le n, \tag{2.8}$$

with $\beta \in [0,1]$ and $h_{ij} > 0$.

Lemma 2.3 and its corollary also hold for (2.8), and the proofs are again straightforward. The subsequent discussion is valid for (2.8) as well, but it will be circumscribed to (2.7) for the sake of notational simplicity. We assume that $h_0 \le h_0'$.

In (2.2) and (2.3), the choice of

$$Q_k\left(y^k\right) := P_k\left(y^k\right) := J\left(y^k, h_k\right)^{-1}, \qquad \text{with } 0 < h_k \le h_0 \tag{2.9}$$

yields convergence for these discrete Newton-Fourier (D-N-F) iterates, and it is analized in the following lemma. The proof applies some well-established arguments [1].

LEMMA 2.6.

(i) $\left\|y^{k+1} - x^{k+1}\right\| \le c_1\gamma\left[K\left\|x^k - y^k\right\| h_k + \frac{1}{2}\left\|x^k - y^k\right\|^2\right], \qquad k = 0, 1, \ldots.$
(ii) $x^k \le x_N^k \le y^* \le y_N^k \le y^k, \qquad k = 0, 1, \ldots.$

PROOF.

(i)

$$\begin{aligned}\left\|y^{k+1} - x^{k+1}\right\| &\le \left\|J(y^k,h_k)^{-1}\right\| \left\|J(y^k,h_k)(y^k - x^k) - (Fy^k - Fx^k)\right\| \\ &\le c_1 \left\|J(y^k,h_k) - F'(y^k)\right\| \left\|y^k - x^k\right\| + c_1 \left\|F'(y^k)(y^k - x^k) - (Fy^k - Fx^k)\right\| \\ &\le c_1 K\gamma h_k \left\|y^k - x^k\right\| + c_1\gamma\frac{1}{2}\left\|y^k - x^k\right\|^2.\end{aligned}$$

Note that, if for instance $\|\quad\| := \|\quad\|_1$, then $K = 1$.

(ii) It follows inductively.

REMARK 2.7. The main fact to be retained from Lemma 2.6 throughout this note is that, by choosing $h_k \leq c\left\|y^k - x^k\right\|$, quadratic convergence is attained for the D-N-F iterates, as it is the case for the nondiscrete ones. Samanskii [1] originally pointed out that the choice $|h_k| \leq c\left\|Fy^k\right\|$, with some constant c, yields at least quadratic convergence for general discrete Newton iterates, without any convexity assumption on F [1]. Note also that, if $F'(y^*)$ is nonsingular, then the inverse function theorem implies the existence of a neighborhood U of y^* and positive constants m and M, such that

$$m\,\|Fy\| \leq \|y - y^*\| \leq M\,\|Fy\|, \qquad \forall\, y \in U. \tag{2.10}$$

This means that Samanskii's choice yields a value of h_k that is of the same order as that of the error. Note also that, since on compact neighborhoods of y^*, $J(y,h)$ (resp. $J(y,h)^{-1}$) converges uniformly to $F'(y)$ (resp. $F'(y)^{-1}$)), then there exist positive constants m' and M', such that for y^0 in an appropriate neighborhood of y^*,

$$m'\left\|Fy_N^k\right\| \leq \left\|y_N^{k+1} - y_N^k\right\| \leq M'\left\|Fy_N^k\right\|. \tag{2.11}$$

Clearly, (2.10) and (2.11) imply in general that $\left\|y_N^{k+1} - y_N^k\right\|, \left\|Fy_N^k\right\|$ and $\left\|y_N^k - y^*\right\|$ converge to 0 with the same order. Also, the same holds for the D-N-F iterations. But in order to ensure at least quadratic convergence in the present context, an alternative choice is thus given by $0 < h_k \leq c\left\|y^k - x^k\right\|$. Finally, it is to be noticed that (ii) in Lemma 2.6 implies that no assymptotic improvement in the convergence of the D-N-F iterations can be expected when the N-F iterates converge quadratically if, for example $h_k := c_k\left\|Fy^k\right\|$ with $\lim c_k = 0$. To summarize, the N-F iterates always converge termwise better than the D-N-F ones and the latter also converge at least quadratically if $h_k := c\left\|Fy^k\right\|$ or $h_k := c\left\|y^k - x^k\right\|$. All these considerations will be necessary in the fourth section.

3. CONVERGENCE AFTER ACCURATE FUNCTIONAL ELIMINATION

As mentioned in the introduction, it has been proven that accurate functional elimination in (2.1) leads to convergence improvement of the N-F iterates for the reduced system [3]. Analogue results for D-N-F iterates will be discussed in this section. Some additional notation and results will be needed [3].

Since $F'(y^*)$ is a nonsingular M-matrix, the implicit function theorem yields the existence of neighborhoods U of y^*, V of $\overline{y^*} := (y_2^*, \ldots, y_n^*)$, a function $g : V \to \Re$, for which $f_1(g(\overline{y}), \overline{y}) = 0$, $(\overline{y} := (y_2, \ldots, y_n))$, such that if $y \in U$ satisfies $f_1(y) = 0$, then $y_1 = g(\overline{y})$. It will be assumed that

$$\left\langle x^0, y^0 \right\rangle \subset U \qquad \text{and} \qquad \left\langle \overline{x^0}, \overline{y^0} \right\rangle \subset V.$$

Note that distinction between row and column vectors is avoided unless strictly necessary. We may now consider the reduced system

$$\overline{F}, \overline{y} = 0, \qquad \text{with } \overline{F} := (\overline{f}_i), \quad i = 2, \ldots, n, \quad \overline{y} \in V, \qquad \text{and } \overline{f}_i(\overline{y}) := f_i(g(\overline{y}), \overline{y}). \tag{3.1}$$

The reduced system (3.1) has been shown to inherit the properties of (2.1) [3]. Consequently, the corresponding N-F iterates may be considered. They will be denoted $(\overline{y}_N^k)$ and $(\overline{x}_N^k)$, with starting points $\overline{y^0}$ and $\overline{x^0}$, respectively. For $k = 0$, we set $\overline{x}_N^0 := \overline{x^0}$ and $\overline{y}_N^0 := \overline{y^0}$. The next comparison result has been established [3].

THEOREM 3.1. *The following inequalities hold:*

$$\text{(i)}\quad \underline{x_N^k} \le \overline{x}_N^k \le \overline{y^*} \le \overline{y}_N^k \le \overline{y_N^k}, \qquad k = 0, 1, \ldots. \tag{3.2}$$

$$\text{(ii)}\quad (x_N^k)_1 \le g(\overline{x}_N^k) \le y_1^* \le g(\overline{y}_N^k) \le (y_N^k)_1, \qquad k = 0, 1, \ldots. \tag{3.3}$$

REMARK 3.2. Clearly, the meaning of (i) is that the reduced N-F iterates converge termwise faster than the corresponding nonreduced ones, whereas (ii) implies that the evaluation of g on the reduced iterates also produces better values for the eliminated coordinate. These two facts can also be exploited in order to improve the stopping criterion based on $\|y_N^k - x_N^k\|_\infty \le \text{Tol}$ by checking first for $\|\overline{y}_N^k - \overline{x}_N^k\|_\infty \le \text{Tol}$ and, once it is satisfied, the reduced iterations proceed until $|g(\overline{y}_N^k) - g(\overline{x}_N^k)| \le \text{Tol}$. In accordance with (2.7), we define the reduced approximation of $\overline{F}'$ by

$$\left[\overline{J}(\overline{y}, h)\right]_{i,j} := \frac{1}{h}\left[\overline{f}_i\left(\overline{y} + h\,\overline{e^j}\right) - \overline{f}_i(\overline{y})\right], \qquad 2 \le i, j \le n. \tag{3.4}$$

LEMMA 3.3. *If*

(a) $\partial_1 f_i$ *is constant for all* $1 \le i \le n$, *or*

(b) $\partial_j f_1$ *is constant for all* $1 \le j \le n$, *then* $\overline{J}(\overline{y}, h)$ *is nonsingular and*

$$\left[\overline{J}(\overline{y}, h)\ \right]_{i,j}^{-1} = \left[J((g(\overline{y}), \overline{y}), h)\right]_{i,j}^{-1}, \qquad 2 \le i, j \le n.$$

PROOF. Recall that

$$\begin{aligned}\left[\overline{J}(\overline{y}, h)\right]_{i,j} &= \int_0^1 \partial_j \overline{f}_i\left(\overline{y} + th\,\overline{e^j}\right) dt \\ &= \int_0^1 \left[\partial_j f_i(g(.), .) - \partial_1 f_i(g(.), .) * \frac{\partial_j f_1(g(.), .)}{\partial_1 f_1(g(.), .)}\right] dt.\end{aligned} \tag{3.5}$$

When (a) holds, consider

$$M_L := \begin{pmatrix} 1 & 0 & 0 & \ldots & \ldots & 0 \\ -m_{2,1} & 1 & 0 & \ldots & \ldots & 0 \\ -m_{3,1} & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 1 & 0 \\ -m_{n,1} & 0 & \ldots & \ldots & 0 & 1 \end{pmatrix}, \qquad \text{with } m_{i,1} := \frac{J_{i,1}}{J_{1,1}}, \qquad 1 \le i \le n,$$

whereas if (b) holds, we define

$$M_U := \begin{pmatrix} 1 & -m_{1,2} & -m_{1,3} & \ldots & \ldots & -m_{1,n} \\ 0 & 1 & 0 & \ldots & \ldots & 0 \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 1 & 0 \\ 0 & 0 & \ldots & \ldots & 0 & 1 \end{pmatrix}, \qquad \text{with } m_{1,j} := \frac{J_{1,j}}{J_{1,1}}, \qquad 1 \le j \le n.$$

Now (3.5) and (a) imply that

$$M_L * J = \begin{pmatrix} J_{1,1} & \ldots & \ldots & J_{1,n} \\ 0 & & & \\ \vdots & & \overline{J} & \\ 0 & & & \end{pmatrix}, \tag{3.6}$$

while (3.5) and (b) imply that

$$J * M_U = \begin{pmatrix} J_{1,1} & 0 & \dots & 0 \\ \vdots & & & \\ \vdots & & \overline{J} & \\ J_{n,1} & & & \end{pmatrix}. \tag{3.7}$$

The conclusion follows in either case by considering a formal inversion in (3.6) and (3.7) and by taking into account the block structure of the involved matrices.

The analogue of Theorem 3.1 may now be stated for discrete and reduced discrete N-F iterates. Recall that the reduced D-N-F iterates are defined by

$$\begin{aligned} \overline{y}^{k+1} &:= \overline{y}^k - \overline{J}(\overline{y}^k, h_k)^{-1}\overline{F}\overline{y}^k, \\ \overline{x}^{k+1} &:= \overline{x}^k - \overline{J}(\overline{y}^k, h_k)^{-1}\overline{F}\overline{x}^k, \quad \text{where} \\ \overline{y}^0 &:= \overline{y^0} \quad \text{and} \quad \overline{x}^0 := \overline{x^0}. \end{aligned}$$

THEOREM 3.4. *If* (a) *or* (b) *in Lemma 3.3 hold, then*

$$\text{(i)} \quad \overline{x^k} \le \overline{x}^k \le \overline{y^*} \le \overline{y}^k \le \overline{y^k}, \qquad k = 0, 1, \dots,$$

and

$$\text{(ii)} \quad x_1^k \le g(\overline{x}^k) \le y_1^* \le g(\overline{y}^k) \le y_1^k, \qquad k = 0, 1, \dots.$$

PROOF. The proof follows as that of Theorem 3.1, and by taking account of Lemma 3.3.

REMARK 3.5. When (a) holds, the question arises whether to use some i^{th} equation, $i \neq 1$, for which $\partial_1 f_i \neq 0$, in order to eliminate the first unknown in (2.1). Such procedure may lead to a reduced system that doesn't satisfy the hypotheses in Theorem 2.1, but even if it does satisfy them, then the convergence for the reduced N-F iterates turns out to be slower [3]. This is the reason for calling *accurate partial elimination* the procedure based on eliminating an unknown by means of its corresponding equation. However, if a subset of unknowns is to be eliminated before proceeding with the iterations, the corresponding subset of equations may be handled at will, because the final reduced system does not depend on the order of the elimination.

The hypotheses in Theorem 3.4 are restrictive enough, when compared with those in Theorem 3.1. This is the reason for the approach that will now be examined. Notice that explicit mention of intermediate points in the application of mean value theorems will be omitted.

LEMMA 3.6. *If $F'(y^0)$ is irreducible, then g is strictly isotone on $\left\langle \overline{x^0}, \overline{y^0} \right\rangle$; i.e.,*

$$\overline{x} < \overline{y} \quad \text{implies} \quad g(\overline{x}) < g(\overline{y}).$$

Moreover, if $x^0 \neq y^*$ or $y^0 \neq y^*$, then necessarily $x^0 < y^0$. Finally, $f_1(y^0) > 0$ if and only if $g(\overline{y^0}) < y_1^0$, and $f_1(x^0) < 0$ if and only if $x_1^0 < g(\overline{x^0})$.

PROOF. Since $F'(y)$ is irreducible for $y \le y^0$, for each such y there is some nonvanishing partial derivative $\partial_j f_1(y)$ with $2 \le j \le n$. Thus,

$$g(\overline{y}) - g(\overline{x}) = \sum_{j=2}^{n} \partial_j g * (y_j - x_j) = -\sum_{j=2}^{n} \frac{\partial_j f_1}{\partial_1 f_1} * (y_j - x_j) > 0.$$

As for the second statement, assume for instance that $x^0 \neq y^*$; i.e., $Fx^0 \neq 0$ and $Fx^0 \leq 0$. Then, since $F'(y^0)^{-1} > 0$, it follows that

$$x^0 < x^0 - F'(y^0)^{-1}Fx^0 = x^1. \tag{3.8}$$

Theorem 2.1 also yields

$$x^1 \leq y^* \leq y^0,$$

which combined with (3.8) gives $x^0 < y^0$. The third part can be dealt with by noting that

$$f_1\left(y^0\right) = f_1\left(y^0\right) - f_1\left(g\left(\overline{y^0}\right), \overline{y^0}\right) = \partial_1 f_1 * \left(y_1^0 - g\left(\overline{y^0}\right)\right).$$

THEOREM 3.7. *Suppose that for some* $k \geq 0$, $F'(g(\overline{y}_N^k), \overline{y}_N^k) \neq F'(y_N^k)$ *and* $F'(y^k)$ *is irreducible. Then the following propositions hold:*

(i) *If* $\overline{y}_N^k \neq \overline{y^*}$, *then* $\overline{y}_N^{k+1} < \overline{y_N^{k+1}}$ *and* $g(\overline{y}_N^{k+1}) < (y_N^{k+1})_1$.

(ii) *If* $\overline{x}_N^k \neq \overline{y^*}$, *then* $\overline{x_N^{k+1}} < \overline{x}_N^{k+1}$ *and* $(x_N^{k+1})_1 < g(\overline{x}_N^{k+1})$.

PROOF. (i) Consider the first inequality; $F'(y_N^k)$ being irreducible yields $F'(y_N^k)^{-1} > 0$, while irreducibility and $F'(g(\overline{y}_N^k), \overline{y}_N^k) \neq F'(y_N^k)$ imply

$$F'(y_N^k)^{-1} < F'(g(\overline{y}_N^k), \overline{y}_N^k)^{-1}.$$

Clearly, Theorem 2.1 applied to (3.1) yields $\overline{F}\,\overline{y}_N^k \geq 0$; thus,

$$\begin{aligned}\left(g\left(\overline{y}_N^k\right), \overline{y}_N^k\right) - F'\left(g\left(\overline{y}_N^k\right), \overline{y}_N^k\right)^{-1} F\left(g\left(\overline{y}_N^k\right), \overline{y}_N^k\right) < \left(g\left(\overline{y}_N^k\right), \overline{y}_N^k\right) \\ - F'\left(y_N^k\right)^{-1} F\left(g\left(\overline{y}_N^k\right), \overline{y}_N^k\right) \leq y_N^k - F'\left(y_N^k\right)^{-1} F y_N^k = y_N^{k+1}.\end{aligned} \tag{3.9}$$

The last inequality in (3.9) is a consequence of the order convexity of F. The proof for the first inequality can now be completed by applying the arguments needed in the proof of Theorem 3.1, i.e., Theorem 3.9 in [3]. As for the second inequality, Lemma 3.6 yields

$$g\left(\overline{y}_N^{k+1}\right) < g\left(\overline{y_N^{k+1}}\right),$$

while

$$0 \leq f_1\left(y_N^{k+1}\right) = f_1\left(y_N^{k+1}\right) - f_1\left(g\left(\overline{y_N^{k+1}}\right), \overline{y_N^{k+1}}\right) = \partial_1 f_1 * \left(\left(y_N^{k+1}\right)_1 - g\left(\overline{y_N^{k+1}}\right)\right).$$

Since $\partial_1 f_1 > 0$, the conclusion follows from the preceding two inequalities.

(ii) The proof is very similar to the corresponding one for (i).

COROLLARY 3.8. *Suppose that* $F'(y^0)$ *is irreducible, that* F' *is not constant on any open subset of* $\langle x^0, y^0 \rangle$, *and that* $F'(g(\overline{y^0}), \overline{y^0}) \neq F'(y^0)$.

(i) *If* $\overline{y}_N^k \neq \overline{y^*}$, *then* $\overline{y}_N^m < \overline{y_N^m}$ *and* $g(\overline{y}_N^m) < (y_N^m)_1$ *for* $1 \leq m \leq k+1$.

(ii) *If* $\overline{x}_N^k \neq \overline{y^*}$, *then* $\overline{x_N^m} < \overline{x}_N^m$ *and* $(x_N^m)_1 < g(\overline{x}_N^m)$ *for* $1 \leq m \leq k+1$.

PROOF. (i) Theorem 2.1 applied to $\overline{F}$ with respect to $\left\langle \overline{x^0}, \overline{y^0} \right\rangle$ implies that

$$\overline{y}_N^m \neq \overline{y^*}, \qquad 0 \leq m \leq k.$$

Thus, inductively,

$$(g(\overline{y}_N^m), \overline{y}_N^m) < y_N^m$$

and

$$F'(g(\overline{y}_N^m), \overline{y}_N^m) \neq F'(y_N^m), \qquad \text{for } 1 \leq m \leq k+1.$$

THEOREM 3.9. *With the hypotheses in Corollary* 3.8, *if* $0 < h_m$ *in* (2.9), $0 \leq m \leq k$, *are chosen conveniently small, then the following hold:*

(i) *If* $\overline{y}_N^k \neq \overline{y^*}$, *then* $\overline{y}^m < \overline{y^m}$ *and* $g(\overline{y}^m) < y_1^m$ *for* $1 \leq m \leq k+1$.

(ii) *If* $\overline{x}_N^k \neq \overline{y^*}$, *then* $\overline{x^m} < \overline{x}^m$ *and* $y_1^m < g(\overline{y}^m)$ *for* $1 \leq m \leq k+1$.

PROOF. It is only necessary to recall that $J(y,h)^{-1}$ (resp. $\overline{J}(\overline{y},h)^{-1}$) converges uniformly on compact sets to $F'(y)^{-1}$ (resp. $\overline{F}'(\overline{y})^{-1}$).

REMARK 3.10. Unlike Theorem 3.4, Theorem 3.9 only has a qualitative meaning. It is our purpose now to show that the hypotheses in (i) and (ii) are automatically satisfied if $\overline{F}'$ is not constant on any open set.

LEMMA 3.11. *Suppose that $F'(y^0)$ is irreducible and that F' is not constant on any open subset of $\langle x^0, y^0\rangle$.*

(i) *If $y^0 \neq y^*$, then $y^* < y_N^{k+1} < y_N^k$ for $k = 0, 1, \ldots$.*

(ii) *If $x^0 \neq y^*$, then $x_N^k < x_N^{k+1} < y^*$ for $k = 0, 1, \ldots$.*

PROOF. The proof is simple, so that only (i) will be considered. Since $F'(y^0)^{-1} > 0$ and $Fy^0 \geq 0$, then

$$y_N^1 = y^0 - F'(y^0)^{-1} F y^0 < y^0.$$

Recall that $Fy_N^1 \geq 0$ and suppose that $Fy_N^1 = 0$. Thus,

$$\begin{aligned} Fy^0 = Fy^0 - Fy_N^1 &= \int_0^1 F'\left(y_N^1 + t\left(y^0 - y_N^1\right)\right)\left(y^0 - y_N^1\right) dt \\ &\leq F'\left(y^0\right)\left(y^0 - y_N^1\right) = Fy^0, \end{aligned}$$

and

$$\int_0^1 F'\left(y_N^1 + t\left(y^0 - y_N^1\right)\right)\left(y^0 - y_N^1\right) dt = F'\left(y^0\right)\left(y^0 - y_N^1\right). \tag{3.10}$$

If

$$\partial_j f_i\left(y_N^1 + t\left(y^0 - y_N^1\right)\right) = \partial_j f_i\left(y^0\right), \qquad \forall\, t \in (0,1), \quad 1 \leq i, j \leq n,$$

isotonicity would imply that

$$\partial_j f_i(z) = \partial_j f_i\left(y^0\right), \qquad \forall\, z \in \left\langle y_N^1, y^0\right\rangle,$$

which contradicts that F' is not constant on any open set. Thus, there exist i_0, j_0 and $t_0 \in (0,1)$, such that

$$\partial_{j_0} f_{i_0}\left(y_N^1 + t\left(y^0 - y_N^1\right)\right) < \partial_{j_0} f_{i_0}\left(y^0\right), \qquad \forall t \leq t_0. \tag{3.11}$$

Consider

$$a_{ij} := \int_0^1 \partial_j f_i\left(y_N^1 + t\left(y^0 - y_N^1\right)\right) dt, \quad \text{for } 1 \leq i, j \leq n.$$

Equation (3.11) implies that

$$A \leq F'\left(y^0\right),$$

with equality excluded, whence

$$A\left(y^0 - y_N^1\right) \leq F'\left(y^0\right)\left(y^0 - y_N^1\right),$$

with equality excluded as well. But this contradicts (3.10), which yields

$$Fy_N^1 \neq 0.$$

An induction argument completes the proof.

THEOREM 3.12. *Suppose that $F'(y^0)$ is irreducible, $F'(g(\overline{y^0}), \overline{y^0}) \neq F'(y^0)$, $\overline{x^0} \neq \overline{y^*}$ and that $\overline{F}'$ is not constant on any open subset of $\left\langle \overline{x^0}, \overline{y^0}\right\rangle$. Then, for $k = 1, 2, \ldots$,*

$$\overline{x_N^k} < \overline{x}_N^k < \overline{y^*} < \overline{y}_N^k < \overline{y_N^k} \quad \text{and} \quad (x_N^k)_1 < g(\overline{x}_N^k) < y_1^* < g(\overline{y}_N^k) < (y_N^k)_1.$$

For each k, there exist $\epsilon_m > 0, 0 \leq m \leq k$, depending on k and such that if $0 < h_m \leq \epsilon_m$, then for $1 \leq m \leq k+1$,

$$\overline{x^m} < \overline{x}^m < \overline{y^*} < \overline{y}^m < \overline{y^m} \quad \text{and} \quad x_1^m < g(\overline{x}^m) < y_1^* < g(\overline{y}^m) < y_1^m.$$

PROOF. Notice first that $F'(g(\overline{y^0}), \overline{y^0}) \neq F'(y^0)$ assumes that $g(\overline{y^0}) < y_1^0$. Since $F'(y^0)$ is irreducible, there exists $i \geq 2$ for which

$$\partial_1 f_i(y^0) > 0,$$

whence

$$f_i(y^0) - f_i\left(g\left(\overline{y^0}\right), \overline{y^0}\right) = \partial_1 f_i * \left(y_1^0 - g\left(\overline{y^0}\right)\right) < 0.$$

Thus,

$$\overline{y^0} \neq \overline{y^*}.$$

Since $(\overline{F}')_{ij}^{-1}(\overline{y^0}) = (F')_{ij}^{-1}(g(\overline{y^0}), \overline{y^0})$, for $i \neq 1 \neq j$ [3], the irreducibility of $\overline{F}'(\overline{y^0})$ follows from that of $F'(y^0)$. Lemma 3.11 can now be applied to $\overline{F}$ with respect to $\left\langle \overline{x^0}, \overline{y^0} \right\rangle$ and it yields that

$$\overline{y}_N^k \neq y^* \neq \overline{x}_N^k, \qquad k = 0, 1, \ldots.$$

Recall that

$$\partial_j \overline{f}_i(.) = \partial_j f_i(g(.), .) - \partial_1 f_i(g(.), .) * \frac{\partial_j f_1(g(.), .)}{\partial_1 f_1(g(.), .)},$$

which implies that also F' cannot be constant on any open subset of $\langle x^0, y^0 \rangle$. The inequalities for the N-F iterates now follow from Corollary 3.8, while those relating the D-N-F iterates can be obtained with the argument in Theorem 3.9.

REMARK 3.13. Although the necessary condition $f_1(y^0) > 0$ (i.e., $g(\overline{y^0}) < y_1^0$) may not be satisfied, note that the procedure leading to the reduced system (3.1) might have been applied to any other variable with its corresponding equation. Thus, unless $F(y^0) = 0$, it may be supposed that $f_1(y^0) > 0$. However, in actual problems, the elimination criterion should take into account both the simplicity of the equation to be used and the resulting complexity for the reduced iterations. Note that the condition $\overline{x^0} \neq \overline{y^*}$, which is equivalent to $\overline{F}\,\overline{x^0} \neq 0$, is implied by $x_1^0 < g(\overline{x^0})$; i.e., $f_1(x^0) < 0$. Note also that it is easy to exhibit examples in the context given by Theorem 2.1 for which it may happen that $\overline{F}'$ is constant while F' is not constant on any open subset of $\langle x^0, y^0 \rangle$. However, Theorem 3.12 also holds if, instead of asking $\overline{F}'$ not being constant on open sets, it is only supposed that F' is not constant on any open subset of $\langle x^0, y^0 \rangle$. The proof in this case takes into account strict inequalities like the one in (3.9) and can be obtained by making slight changes in the proof of Theorem 3.1. A different proof for this modification of Theorem 3.12 may be found in [7].

4. AN EXAMPLE

Let us define $F : \Re^{10} \longrightarrow \Re^{10}$ by

$$\begin{aligned} f_1 &:= \frac{2y_1 - y_2}{h^2} + y_1^3, \\ f_i &:= \frac{2y_i - y_{i-1} - y_{i+1}}{h^2} + y_i^3, \qquad 2 \leq i \leq 9, \\ f_{10} &:= \frac{2y_{10}^3 - y_9}{h^2}, \qquad \text{with } h := \frac{1}{10}. \end{aligned}$$

By eliminating y_{10} by means of f_{10}, the following reduced system is obtained:

$$\overline{f}_1 = \frac{2y_1 - y_2}{h^2} + y_1^3$$
$$\overline{f}_i = \frac{2y_i - y_{i-1} - y_{i+1}}{h^2} + y_i^3, \qquad 2 \le i \le 8$$
$$\overline{f}_9 = \frac{2y_9 - y_8 - g}{h^2} + y_9^3, \qquad \text{where } g := \left(\frac{y_9}{2}\right)^{1/3}.$$

Consider $x^0 := (0, \ldots, 0, 0.14, 0.41)$ and $y^0 := (1, \ldots, 1)$. This example is taken from [3], where the verification of the hypotheses in Theorem 2.1 is included. Clearly, also the hypotheses in Theorem 3.12 are satisfied. The calculations have been carried out on a PC, with the double precision of Fortran 5.0. The stopping criteria have been $\left\|Fy^k\right\|_\infty < \epsilon := 0.5 * 10^{-13}$ and $\left\|Fx^k\right\|_\infty < \epsilon$, and their analogues for the reduced counterparts. It is to be noted that, with the exceptions in Tables 2 and 5, the reduced iterations converge faster than the nonreduced ones.

Table 1.

	$\|\|Fy_N^k\|\|_\infty < \epsilon$	$\|\|Fx_N^k\|\|_\infty < \epsilon$	$\|\|\overline{F}\,\overline{y}_N^k\|\|_\infty < \epsilon$	$\|\|\overline{F}\,\overline{x}_N^k\|\|_\infty < \epsilon$
k	6	8	5	5

Table 1 gives for reference the number of necessary N-F and reduced N-F iterations in order to satisfy the stopping criteria (see [8] for the actual iterates); because of (ii) in Lemma 2.6, these values are theoretical lower bounds for the corresponding ones of discrete iterations that are given in the following tables. The tables represent the tests with different ways of producing h_m, i.e., the value of h in (2.7) used to approximate the Jacobian matrix for iteration number $m+1$. In each table, the first column shows the values of c used to produce h_m; each of the following four columns gives the corresponding minimum value of k for which the stopping criterion on top of the column has been satisfied (or not).

Table 2.

$h_m :=$	$c * \|\|Fy^m\|\|_\infty$		$c * \|\|\overline{F}\,\overline{y}^m\|\|_\infty$	
c	$\|\|Fy^k\|\|_\infty < \epsilon$	$\|\|Fx^k\|\|_\infty < \epsilon$	$\|\|\overline{F}\,\overline{y}^k\|\|_\infty < \epsilon$	$\|\|\overline{F}\,\overline{x}^k\|\|_\infty < \epsilon$
10^{-1}	73	73	8	8
10^{-2}	10	11	5	8
10^{-3}	6	12	5	13
10^{-4}	6	12	6	16
10^{-5}	6	∞(52)	9	RT (335)
10^{-6}	7	∞(39)	16	RT (96)

RT stands here for breakdown in the calculations due to singularity of the approximate Jacobian matrix; what is in parentheses indicates the value of k for its occurence. Analogously, ∞ stands for machine overflow. Note the almost uniform bad behavior with repect to the Fourier iterates; as a matter of fact, no choice of c has been permitted to attain either the value $k = 6$ for the nonreduced Newton iterates or the value $k = 5$ for the reduced ones. This can be only partially explained by the fact that when the Newton iterates approach the root, the approximation of the Jacobian matrix may tend to be numerically ill-conditioned. Besides, note that for $c := 10^{-1}$, convergence for the nonreduced iterations is very slow. Another significant point is given by the comparison between the values of k for $c = 10^{-3}$ and $c = 10^{-4}$; they suggest that, with the notation in Theorem 3.12, some h_m are larger than ϵ_m. Notice also that with $c = 10^{-5}$ or $c = 10^{-6}$, the reduced Fourier iterates converge much slower than the nonreduced ones. Finally, note that only for $c := 10^{-3}$ the optimal values in Table 1 for the Newton iterations are attained. All this suggests that Samanskii's choice, although analytically sound, may be numerically bad.

Table 3.

$h_m :=$	$c * \max(\|Fy^m\|_\infty, \|Fx^m\|_\infty)$		$c * \max(\|\overline{F}\overline{y}^m\|_\infty, \|\overline{F}\overline{x}^m\|_\infty)$	
c	$\|Fy^k\|_\infty < \epsilon$	$\|Fx^k\|_\infty < \epsilon$	$\|\overline{F}\overline{y}^k\|_\infty < \epsilon$	$\|\overline{F}\overline{x}^k\|_\infty < \epsilon$
10^{-1}	80	80	8	8
10^{-2}	11	11	6	6
10^{-3}	8	8	5	5
10^{-4}	7	8	5	6
10^{-5}	7	8	5	8
10^{-6}	18	19	8	13
10^{-7}	303	303	157	157

In the present context, one wonders whether it may be numerically better to somehow take into account both residues, i.e., at Newton's and at Fourier's iterates. The corresponding results are described in Table 3.

The situation described here has improved when compared with that corresponding to Table 2; however, it is still far from good. It is to be noted that for the reduced iterations the behavior is definitely better than for the nonreduced ones. Another possibility involving the residues can be based on (2.4) and is described in Table 4.

Table 4.

$h_m :=$	$c * \|Fy^m - Fx^m\|_\infty$		$c * \|\overline{F}\overline{y}^m - \overline{F}\overline{x}^m\|_\infty$	
c	$\|Fy^k\|_\infty < \epsilon$	$\|Fx^k\|_\infty < \epsilon$	$\|\overline{F}\overline{y}^k\|_\infty < \epsilon$	$\|\overline{F}\overline{x}^k\|_\infty < \epsilon$
10^{-1}	134	134	9	9
10^{-2}	13	13	6	6
10^{-3}	8	8	5	5
10^{-4}	7	8	5	5
10^{-5}	6	8	5	9
10^{-6}	6	8	5	10
10^{-7}	6	208	5	52

Table 5.

$h_m :=$	$c * \|y^m - x^m\|_\infty$		$c * \|\overline{y}^m - \overline{x}^m\|_\infty$	
c	$\|Fy^k\|_\infty < \epsilon$	$\|Fx^k\|_\infty < \epsilon$	$\|\overline{F}\overline{y}^k\|_\infty < \epsilon$	$\|\overline{F}\overline{x}^k\|_\infty < \epsilon$
10^{-1}	7	8	5	5
10^{-2}	7	8	5	5
10^{-3}	6	8	5	9
10^{-4}	6	9	5	10
10^{-5}	6	82	5	24

In Table 5, we turn to the second criterion discussed in Remark 2.7; i.e., we consider $h_m := c * \|y^m - x^m\|_\infty$. Table 5 clearly shows a more robust behavior for the discrete iterations than those described by the previous three tables. However, no best possible results are attained; i.e., no row matches the data row in Table 1. Note also that, as with Table 2, the values $c = 10^{-3}$ and $c = 10^{-4}$ yield slower converging reduced Fourier iterations. Tables 2 through 4 suggest the need of introducing a damping element that makes h_m small when the residues are large, and keeps h_m of the same size of the residues when they are small; here, small is intended with respect to working precision. The usefulness of this point of view is illustrated in Table 6. The vertical dots in Table 6 mean that k remains constant columnwise for the corresponding intermediate values of c. The described results give good evidence of the usefulness of the damping element; Table 7

Table 6.

$h_m :=$	$\min(c, \max(\|\|Fy^m\|\|_\infty, \|\|Fx^m\|\|_\infty))$		$\min(c, \max(\|\|\overline{F}\,\overline{y}^m\|\|_\infty, \|\|\overline{F}\,\overline{x}^m\|\|_\infty))$	
c	$\|\|Fy^k\|\|_\infty < \epsilon$	$\|\|Fx^k\|\|_\infty < \epsilon$	$\|\|\overline{F}\,\overline{y}^k\|\|_\infty < \epsilon$	$\|\|\overline{F}\,\overline{x}^k\|\|_\infty < \epsilon$
10^{-1}	11	12	7	7
10^{-2}	9	9	6	6
10^{-3}	8	8	5	6
10^{-4}	7	8	5	6
10^{-5}	7	8	5	6
10^{-6}	6	8	5	6
10^{-7}	6	8	5	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
10^{-10}	6	8	5	5

Table 7.

$h_m :=$	$\min(c, \|\|y^m - x^m\|\|_\infty)$		$\min(c, \|\|\overline{y}^m - \overline{x}^m\|\|_\infty)$	
c	$\|\|Fy^k\|\|_\infty < \epsilon$	$\|\|Fx^k\|\|_\infty < \epsilon$	$\|\|\overline{F}\,\overline{y}^k\|\|_\infty < \epsilon$	$\|\|\overline{F}\,\overline{x}^k\|\|_\infty < \epsilon$
10^{-1}	8	9	6	6
10^{-2}	8	8	5	6
10^{-3}	7	8	5	5
10^{-4}	7	8	5	5
10^{-5}	7	8	5	5
10^{-6}	6	8	5	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
10^{-10}	6	8	5	5

shows that it may yield better results in the context of Table 5, i.e., when related to the sizes of the enclosing intervals given by the Newton-Fourier iterates.

FINAL REMARKS. Discrete Newton-Fourier iterations may be a useful substitute for Newton-Fourier iterations, especially when the Jacobian matrix is difficult to calculate. However, some caution should be excercised as to the choice of the step size used in the difference approximations; the numerical results exhibited here suggest that a reliable step size may be $\|y^m - x^m\|_\infty$ truncated by half the working precision.

REFERENCES

1. J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, (1970).
2. J.P. Milaszewicz, Improving Jacobi and Gauss-Seidel iterations, *Linear Algebra and its Applications* **93**, 161–170 (1987).
3. J.P. Milaszewicz and S. Abdel Masih, On elimination and fixed point iterations, *Computers Math. Applic.* **25** (5), 43–53 (1993).
4. A.M. Ostrowski, *Solution of Equations in Euclidean and Banach Spaces*, Academic Press, New York, (1973).
5. A. Baluev, On the method of Chaplygin (Russian), *Doklady Akademii Nauk SSSR* **83**, 781–784 (1952).
6. J.M. Ortega and W.C. Rheinboldt, Monotone iterations for nonlinear equations with application to Gauss-Seidel methods, *Siam Journal on Numerical Analysis* **4**, 171–190 (1967).
7. J.P. Milaszewicz, Comparison theorems for monotone Newton-Fourier iterations and applications in functional elimination, *Linear Algebra and its Applications* (1995).
8. J.P. Milaszewicz and S. Abdel Masih, Errata to "On elimination and fixed point iterations", *Computers Math. Applic.* **27** (4), 113–115 (1994).