On solving a special class of weakly nonlinear finite difference systems *

Emanuele Galligani

Department of Pure and Applied Mathematics “G. Vitali”
University of Modena and Reggio Emilia
Via Campi 213/b, 41100, Modena, Italy
e–mail: galligani@unimo.it

Abstract

In this paper we consider the Newton–iterative method for solving weakly nonlinear finite difference systems of the form \( F(u) = Au + G(u) = 0 \), where the jacobian matrix \( G'(u) \) satisfies an affine invariant Lipschitz condition. We also consider a modification of the method for which we can improve the likelihood of convergence from initial approximations that may be outside the attraction ball of the Newton–iterative method. We analyse the convergence of this damped method in the framework of the line search strategy. Numerical experiments on a diffusion–convection problem show the effectiveness of the method.

Key Words: Weakly nonlinear systems, Newton–iterative method, line search technique, finite difference.

AMS Classification: 65H10, 65H20, 65N06

C.R. Categories: G.1.5, G.1.8

1 Introduction

Let \( \Omega \) be an open and convex subset of \( \mathbb{R}^n \) and let \( F \) be a continuously differentiable mapping, with an invertible jacobian matrix \( F'(u) \) for \( u \in \Omega \). Consider the problem of solving large systems of nonlinear equations

\[
F(u) = 0
\]

where the matrix \( F'(u) \) is sparse.

*This research was supported by the Italian Ministry of Education, University and Research (MIUR), FIRB Project RBAP01877P
We assume that the solution of $u^*$ of the equation (1) exists in $\Omega$.

In many problems of practical interest the mapping $F(u)$ has the form

$$F(u) = Au + G(u)$$

(2)

where $A$ is a nonsingular matrix and $G(u)$ is a continuously mapping with $G'(u)$ invertible for all $u \in \Omega$.

Weakly nonlinear systems of the form (2) often arise from discretization of many classical semilinear partial differential equations by the finite difference method. Here $A$ is large and sparse.

There exists a lot of literature on the solvability and convergence of nonlinear elliptic difference equations. For example, the paper [27] gives conditions for the existence of at least one solution, that belongs to a well defined bounded set, of the quasi-linear system $A(u)u + \Phi(u) + f = 0$ and also for the uniqueness of a solution.

In weakly nonlinear finite difference systems of the form (2) which correspond to elliptic boundary problems related to the study of reaction–diffusion processes and of dynamics of populations (see, e.g. [3], [19], [25], [28], [36]), the vector $G(u)$ can be written as $Bg(u)$, where $B$ is a nonsingular and nonnegative matrix and $g(u)$ is a continuously differentiable diagonal mapping, i.e. a nonlinear mapping whose $i$–th component $g_i$ is a function of only the $i$–th variable $u_i$ for $i = 1, 2, ..., n$.

In the special case in which $A$ is a monotone matrix ([22, p. 360]), $G(u)$ is a $P$–bounded map (i.e. $\exists P \geq 0$ s.t. $|G(u) - G(v)| \leq P|u - v|, \forall u, v \in \Omega$) and $A^{-1}P$ has spectral radius $\rho(A^{-1}P) < 1$, then the system (1)–(2) has a unique solution $u^* \in \Omega$ (e.g. see [32, p. 433]).

This solution can be obtained efficiently on parallel computers with two–stage iterative methods (see [20] and the references therein).

A very important method for the solution of (1) is the Newton method

choose the initial guess $u^{(0)}$

for $k = 0, 1, ..., $ until the convergence do

let $\Delta u^{(k)}$ be the solution of

$$F'(u^{(k)})\Delta u = -F(u^{(k)})$$

$$u^{(k+1)} = u^{(k)} + \Delta u^{(k)}$$

The method of Newton is attractive for its local convergence properties: if $u^{(0)}$ is a sufficiently good initial guess of $u^*$, then the the sequence $\{u^{(k)}\}$ converges $q$–superlinearly to $u^*$. Usually, $F'$ is Lipschitz continuous at $u^*$ with Lipschitz constant $L$; in this case the convergence is $q$–quadratic [32, Theor. 10.2.2]. This condition on $F'(u)$ is necessary also for obtaining an estimation of the radius of the ball that is contained in the attraction basin of the zero $u^*$. An estimation
of such a radius has been determined in [7] showing an inverse proportionality between this quantity and the Lipschitz constant $L$ and the condition number of $F'(u^*)$.

A very recent theoretical work on the estimates of the radius of attraction basin when $F'$ satisfies Hölder conditions at the initial point $u^{(0)}$ are reported in [2] (see also the references therein).

At each stage of Newton method, the Newton equation

$$F'(u^{(k)}) \Delta u = -F(u^{(k)})$$

has to be solved. Computation of the exact solution can be too expensive if $n$ is large and, for any $n$, may not be justified when $u^{(k)}$ is relatively far from $u^*$. Therefore, one might prefer to compute some approximate solution of (3). In practice, the Newton step $\Delta u^{(k)}$ will most frequently be obtained by performing some iterations of an inner linear iterative solver for the equation (3).

The combined Newton–SOR method is well known (see [32, §10.3]); in [20] the Newton–Arithmetic Mean method has been proposed. Thus, if $\Delta \tilde{u}^{(k)}$ is the step actually computed (i.e., the approximate solution of (3)), the residual

$$r^{(k)} = F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} + F(\tilde{u}^{(k)})$$

is not expected to be zero. However, as long as the norm $\|F(\tilde{u}^{(k)})\|$ is still relatively large, there is a little reason for forcing $\|r^{(k)}\|$ to be very small. This suggests that an adaptive control of the inner iterations will be based on the quotient $\|r^{(k)}\|/\|F(\tilde{u}^{(k)})\|$.

Therefore, the Newton iterative method with an adaptive stopping rule, can be stated as follows.

choose the initial guess $\tilde{u}^{(0)}$

for $k = 0, 1, \ldots$, until the convergence do

$$\begin{align*}
\text{find some } & \eta_k \in [0, 1) \text{ and } \Delta \tilde{u}^{(k)} \text{ that satisfy} &
\|F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} + F(\tilde{u}^{(k)})\| &\leq \eta_k \|F(\tilde{u}^{(k)})\| \\
\tilde{u}^{(k+1)} &= \tilde{u}^{(k)} + \Delta \tilde{u}^{(k)}
\end{align*}$$

Taking $\eta_k = 0$, it gives Newton method.

This method, introduced as in (5) in [6] and there referred as inexact Newton method ($\eta_k$ is called forcing term), under standard assumptions will produce iterates which, for a sufficiently good initial guess, converges to $u^*$, provided the forcing sequence $\{\eta_k\}$ is uniformly bounded away from one; besides, the convergence is $q$–linear with respect to the weighted norm $\|\cdot\|^* = \|F'(u^*) \cdot \|$, which is not computable. If $F'(u)$ is Lipschitz continuous at $u^*$ (with Lipschitz constant $L$), the convergence of the method is $q$–quadratic if $\eta_k = O(\|F(\tilde{u}^{(k)})\|)$ as $k \to \infty$. 

3
Some authors (see e.g. [5], [9], [10], [12], [41], [42], [43]) have drawn the attention to variants of the hypothesis that $F'(u)$ is Lipschitz continuous at $u^*$; for instance, it is considered an affine invariant Lipschitz condition for $F'(u)$ with a constant $\Gamma > 0$. There exist problems of practical interest in which the constant $\Gamma$ is of moderate size independently of $L$ (see e.g. [37]). When we make this affine invariant Lipschitz condition on $F'(u)$ the system (1) is said of class $\mathcal{F}$ for short.

An interesting result on the convergence of the Newton method for systems of class $\mathcal{F}$ is reported in [11, p. 88] (see also [9]). The convergence of the Newton iterative method with the adaptive stopping rule, especially of the Newton Arithmetic Mean method for the same systems has been proved in [20]. In this paper we will analyze the Newton iterative method for solving weakly nonlinear finite difference systems (1)–(2) of a special class, called class $\mathcal{M}$. This method requires at each step an inner linear iterative solver for determining an approximate solution of (3). Under the assumption that the jacobian matrix $F'(u)$ is a $T(q, r)$ matrix (see [21]), we will consider the Arithmetic Mean method as inner solver for (3).

This paper is organized as follows. In section 2, we analyze the convergence of Newton method and Newton iterative method. Usually, the attraction ball of the Newton iterative method is not much extended. In order to increase the extent of this ball the Newton iterative method is modified introducing a damping technique. In section 3, we analyze the convergence of the damped Newton iterative method in the framework of the line search strategy. In section 4, we report the results of a numerical study performed on a weakly nonlinear finite difference system arising from the discretization of diffusion–convection equations.

2 Convergence analysis

In this section our interest lies in obtaining conditions which ensure the convergence of Newton method and Newton iterative method for the class $\mathcal{M}$ of mappings $F(u) = Au + G(u)$ defined by

$$\mathcal{M} = \mathcal{M}(\sigma, \lambda) = \{F | F : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n; \ \Omega \text{ open and convex set};$$

- $F$ continuously differentiable on $\Omega$;
- $F'(u)$ and $G'(u)$ invertible on $\Omega$; $A$ nonsingular;
- there exists $u^*$ such that $F(u^*) = 0$;
- $u^*$ is the only solution on $B(u^*, \sigma)$ with $B(u^*, \sigma) = \{v | v \in \Omega, \|v - u^*\| < \sigma\}$
- for all $u, v \in B(u^*, \sigma)$ there exists a $\lambda > 0$ such that $\|G'(u)^{-1}(G'(v) - G'(u))\| \leq \lambda \|v - u\|$.\]

The last condition is an affine invariant Lipschitz condition.

A convergence result on Newton method is given by the following theorem.
Theorem 1. Suppose that the mapping \( F(u) = Au + G(u) \) is of class \( M(\sigma, \lambda) \) for some \( \sigma > 0 \). Assume that
\[
\theta_{\text{max}} = \|A^{-1}\|\|G'(u)\| < 1 \quad \forall u \in B(u^*, \sigma)
\]
and the starting point \( u^{(0)} \) satisfies the condition
\[
\rho \equiv \|u^{(0)} - u^*\| < \frac{2}{\lambda} \frac{1 - \theta_{\text{max}}}{\theta_{\text{max}}} \quad (6)
\]
Then, the sequence \( \{u^{(k)}\}, k > 0 \), generated by Newton method is well defined and converges to \( u^* \), i.e.
\[
\|u^{(k)} - u^*\| < \rho \quad \text{for} \quad k \geq 1 \quad \text{and} \quad \lim_{k \to \infty} u^{(k)} = u^*
\]

Proof. By the affine invariant Lipschitz condition on \( G'(u) \)
\[
\|G'(u)^{-1}(G'(v) - G'(u))\| \leq \lambda\|v - u\| \quad \lambda > 0 \quad (7)
\]
for all \( u, v \in B(u^*, \sigma) \), we derive the following result
\[
\int_0^1 \|e(u^* + t(u - u^*), u)\| dt \leq \frac{\lambda}{2} \|u - u^*\| \quad (8)
\]
where, for all \( u, v \in B(u^*, \sigma) \)
\[
e(u, u) = G'(u)^{-1}(G'(v) - G'(u)) \quad (9)
\]
Indeed, the condition (7) implies
\[
\int_0^1 \|e(u^* + t(u - u^*), u)\| dt \leq \lambda \int_0^1 \|u - u^*\|(1 - t)dt = \frac{\lambda}{2} \|u - u^*\|
\]
Now, we consider the Newton equation (3)
\[
(A + G'(u^{(k)}))(u^{(k+1)} - u^{(k)}) = -(A u^{(k)} + G(u^{(k)}))
\]
that can be written as
\[
(A + G'(u^{(k)}))u^{(k+1)} = G'(u^{(k)})u^{(k)} - G(u^{(k)})
\]
Since \( F(u^*) = Au^* + G(u^*) = 0 \) we can determine the error
\[
(A + G'(u^{(k)}))(u^{(k+1)} - u^*) = -(A + G'(u^{(k)}))u^* + G'(u^{(k)})u^{(k)} -
\]
\[-G(u^{(k)}) + Au^* + G(u^*)
\]
\[
= G'(u^{(k)})(u^{(k)} - u^*) - (G(u^{(k)}) - G(u^*))
\]
5
Using the integral mean value theorem [31, p. 142] we obtain
\[
(A + G'(u^{(k)}))(u^{(k+1)} - u^*) = G'(u^{(k)})(u^{(k)} - u^*) - \\
- \int_0^1 G'(u^* + t(u^{(k)} - u^*)) \cdot (u^{(k)} - u^*) \cdot dt
\]
\[
= G'(u^{(k)})(u^{(k)} - u^*) - \int_0^1 G'(u^{(k)}) \left[ G'(u^{(k)}) \right]^{-1} \cdot \\
\cdot \left[ G'(u^*) + t(u^{(k)} - u^*) - G'(u^{(k)}) \right] + I \right] dt \cdot (u^{(k)} - u^*)
\]
\[
= G'(u^{(k)})(u^{(k)} - u^*) - G'(u^{(k)}) \int_0^1 e(u^* + t(u^{(k)} - u^*), u^{(k)})dt \cdot \\
\cdot (u^{(k)} - u^*) \left[ G'(u^{(k)}) \right]^{-1} \cdot (u^{(k)} - u^*)
\]
\[
= -G'(u^{(k)}) \int_0^1 e(u^* + t(u^{(k)} - u^*), u^{(k)})dt \cdot (u^{(k)} - u^*)
\]
Since \(F'(u), G'(u)\) are invertible and \(A\) is nonsingular, we have
\[
u^{(k+1)} - u^* = -(I + A^{-1}G'(u^{(k)}))^{-1}A^{-1}G'(u^{(k)}) \cdot \\
\cdot \int_0^1 e(u^* + t(u^{(k)} - u^*), u^{(k)})dt \cdot (u^{(k)} - u^*)
\]
Hence
\[
\|u^{(k+1)} - u^*\| \leq \|I + A^{-1}G'(u^{(k)})\|^{-1} \|A^{-1}G'(u^{(k)})\| \cdot \\
\cdot \int_0^1 \|e(u^* + t(u^{(k)} - u^*), u^{(k)})\|dt \cdot \|u^{(k)} - u^*\|
\]
Using the perturbation lemma [31, p. 32] and the inequality (8), we obtain
\[
\|u^{(k+1)} - u^*\| \leq \frac{\|A^{-1}G'(u^{(k)})\|}{1 - \|A^{-1}G'(u^{(k)})\|} \frac{\lambda}{2}\|u^{(k)} - u^*\|^2 \quad (10)
\]
If \(0 < \|u^{(k)} - u^*\| \leq \rho\), then
\[
\|u^{(k+1)} - u^*\| \leq \frac{\|A^{-1}G'(u^{(k)})\|}{1 - \|A^{-1}G'(u^{(k)})\|} \frac{\lambda}{2}\|u^{(k)} - u^*\| \leq \\
\leq \frac{\theta_{\max}}{1 - \theta_{\max}} \frac{\lambda}{2}\|u^{(k)} - u^*\|
\]
Since \(\rho < \frac{1}{\lambda \theta_{\max}}\) we have
\[
\|u^{(k+1)} - u^*\| \leq \theta\|u^{(k)} - u^*\|
\]
with \(\theta = \frac{\theta_{\max} \lambda}{1 - \theta_{\max}} \frac{\lambda}{2}\rho < 1\) and
\[
\|u^{(k)} - u^*\| \leq \theta^k\|u^{(0)} - u^*\|
\]
The condition (6) on the initial point $\|u^{(0)} - u^*\| = \rho$ implies, by induction, that $\|u^{(k)} - u^*\| < \rho$ for all $k \geq 1$ and the sequence $\{u^{(k)}\}$ converges to $u^*$. 

Inequality (10) leads to the following estimate of the speed of convergence

$$\|u^{(k+1)} - u^*\| \leq \frac{\theta_{\max}}{1 - \theta_{\max}} \frac{\lambda}{2} \|u^{(k)} - u^*\|^2$$

(11)

In short, Newton method has local quadratic convergence rate.

An analogous convergence result on the Newton iterative method is given by the following theorem.

**Theorem 2.** Suppose that the mapping $F(u) = Au + G(u)$ is of class $M(\sigma, \lambda)$ for some $\sigma > 0$. Assume that

$$\theta_{\max} = \|A^{-1}\|\|G'(u)\| < 1 \quad \forall u \in B(u^*, \sigma)$$

Let $\{\tilde{u}^{(k)}\}$, $k = 0, 1, \ldots$, be the sequence generated by the Newton iterative method with the condition in (5) on the residual (4) with $\eta_k$ uniformly bounded by $\eta_{\max}$. Assume that

$$\varepsilon_{\max} \equiv \eta_{\max} K(A)(1 + \theta_{\max}) + \theta_{\max} < 1$$

where $K(A) = \|A\|\|A^{-1}\|$ is the condition number of the matrix $A$. Furthermore, let us assume that the starting point $\tilde{u}^{(0)}$ satisfies the condition

$$\bar{\rho} \equiv \|\tilde{u}^{(0)} - u^*\| < \frac{2}{\lambda (1 + \eta_{\max}) \theta_{\max}}$$

(12)

$$\mathcal{B}(u^*, \bar{\rho}) \subseteq \mathcal{B}(u^*, \sigma)$$

Then, the sequence $\{\tilde{u}^{(k)}\}$, $k > 0$, generated by the Newton iterative method is well defined and converges to $u^*$.

**Proof.** We consider the inexact Newton equation (4)

$$(A + G'(\tilde{u}^{(k)}))(\tilde{u}^{(k+1)} - \tilde{u}^{(k)}) = -(A\tilde{u}^{(k)} + G(\tilde{u}^{(k)})) + r^{(k)}$$

where

$$\|r^{(k)}\| \leq \eta_k \|F(\tilde{u}^{(k)})\| \quad \eta_k \leq \eta_{\max}$$

(13)

This equation can be written as

$$(A + G'(\tilde{u}^{(k)}))\tilde{u}^{(k+1)} = G'(\tilde{u}^{(k)})\tilde{u}^{(k)} - G(\tilde{u}^{(k)}) + r^{(k)}$$

Calculations similar to those developed in the proof of Theorem 1 give

$$\tilde{u}^{(k+1)} - u^* = -(I + A^{-1}G'(\tilde{u}^{(k)}))^{-1}A^{-1}.$$
Hence, using inequality (8), we have
\[
\| \tilde{u}^{(k+1)} - u^* \| \leq \| (I + A^{-1}G'(\tilde{u}^{(k)}))^{-1} \| \left( \| A^{-1}G'(\tilde{u}^{(k)}) \| \frac{\lambda}{2} \| \tilde{u}^{(k)} - u^* \|^2 + \| A^{-1}P^{(k)} \| \right)
\]

Using the perturbation lemma [31, p. 32] and the inequality (13), we obtain
\[
\| \tilde{u}^{(k+1)} - u^* \| \leq \frac{1}{1 - \| A^{-1}G'(\tilde{u}^{(k)}) \|} \left( \| A^{-1} \| \| G'(\tilde{u}^{(k)}) \| \frac{\lambda}{2} \| \tilde{u}^{(k)} - u^* \|^2 + \eta \| A^{-1} \| \| F(\tilde{u}^{(k)}) \| \right)
\]

Now, we have ([31, p. 142])
\[
F(\tilde{u}^{(k)}) = F(\tilde{u}^{(k)}) - F(u^*) = \int_0^1 F'(u^* + t(\tilde{u}^{(k)} - u^*)) dt \cdot (\tilde{u}^{(k)} - u^*)
\]
\[
= \int_0^1 \left( A + G'(u^* + t(\tilde{u}^{(k)} - u^*)) \right) dt \cdot (\tilde{u}^{(k)} - u^*)
\]
\[
= A(\tilde{u}^{(k)} - u^*) + \int_0^1 G'(u^* + t(\tilde{u}^{(k)} - u^*)) dt \cdot (\tilde{u}^{(k)} - u^*)
\]

Since, by (9),
\[
G'(v) = G'(u)(\epsilon(u, u) + I)
\]
we can write
\[
\int_0^1 G'(u^* + t(\tilde{u}^{(k)} - u^*)) dt \cdot (\tilde{u}^{(k)} - u^*) = G'(\tilde{u}^{(k)}).
\]
\[
\cdot \int_0^1 \epsilon(u^* + t(\tilde{u}^{(k)} - u^*), \tilde{u}^{(k)}) dt \cdot (\tilde{u}^{(k)} - u^*) + G'(\tilde{u}^{(k)})(\tilde{u}^{(k)} - u^*)
\]

Thus
\[
F(\tilde{u}^{(k)}) = (A + G'(\tilde{u}^{(k)}))(\tilde{u}^{(k)} - u^*) +
\]
\[
+ G'(\tilde{u}^{(k)}) \int_0^1 \epsilon(u^* + t(\tilde{u}^{(k)} - u^*), \tilde{u}^{(k)}) dt \cdot (\tilde{u}^{(k)} - u^*)
\]

Using inequality (8), we obtain
\[
\| F(\tilde{u}^{(k)}) \| \leq \| A \| \| I + A^{-1}G'(\tilde{u}^{(k)}) \| \| \tilde{u}^{(k)} - u^* \| +
\]
\[
+ \| G'(\tilde{u}^{(k)}) \| \frac{\lambda}{2} \| \tilde{u}^{(k)} - u^* \|^2
\]

Therefore, inequality (14) becomes
\[
\| \tilde{u}^{(k+1)} - u^* \| \leq \frac{1}{1 - \| A^{-1}G'(\tilde{u}^{(k)}) \|} \left( \| A^{-1} \| \| G'(\tilde{u}^{(k)}) \| \frac{\lambda}{2} \| \tilde{u}^{(k)} - u^* \|^2 + \right)
\]
\[
\eta_k K(A) \|J + A^{-1} G'(\tilde{u}^{(k)})\| \|\tilde{u}^{(k)} - u^*\| + \\
\eta_k \|A^{-1}\| \|G'(\tilde{u}^{(k)})\| \frac{\lambda}{2} \|\tilde{u}^{(k)} - u^*\|^2
\]

\[
\leq \frac{\|\tilde{u}^{(k)} - u^*\|}{1 - \|A^{-1}\| \|G'(\tilde{u}^{(k)})\|} \left( (1 + \eta_k \|A^{-1}\| \|G'(\tilde{u}^{(k)})\|) \frac{\lambda}{2} \right.

\cdot \|\tilde{u}^{(k)} - u^*\| \eta_k K(A) (1 + \|A^{-1}\| \|G'(\tilde{u}^{(k)})\|) \left. \right)
\]

If \(0 < \|\tilde{u}^{(k)} - u^*\| \leq \bar{\rho}\), then

\[
\|\tilde{u}^{(k+1)} - u^*\| \leq \frac{1}{1 - \theta_{\max}} \left( (1 + \eta_{\max}) \theta_{\max} \frac{\lambda}{2} \bar{\rho} + \\
\eta_{\max} K(A) (1 + \theta_{\max}) \right)
\]

Since \(\bar{\rho} < \frac{2}{\lambda (1 + \eta_{\max}) \theta_{\max}}\), we have

\[
\|\tilde{u}^{(k+1)} - u^*\| \leq \bar{\theta} \|\tilde{u}^{(k)} - u^*\|
\]

with

\[
\bar{\theta} = \frac{1}{1 - \theta_{\max}} \left( (1 + \eta_{\max}) \theta_{\max} \frac{\lambda}{2} \bar{\rho} + \eta_{\max} K(A) (1 + \theta_{\max}) \right)
\]

\[
< \frac{1}{1 - \theta_{\max}} \left( 1 - \varepsilon_{\max} + \eta_{\max} K(A) (1 + \theta_{\max}) \right)
\]

\[
= \frac{1}{1 - \theta_{\max}} = 1
\]

Thus

\[
\|\tilde{u}^{(k)} - u^*\| \leq \bar{\theta}^k \|\tilde{u}^{(0)} - u^*\| \quad \bar{\theta} < 1
\]

The condition (12) on the initial point \(\|\tilde{u}^{(0)} - u^*\| \leq \bar{\rho}\) implies, by induction, that \(\|\tilde{u}^{(k)} - u^*\| < \bar{\rho}\) for all \(k \geq 1\) and the sequence \(\{\tilde{u}^{(k)}\}\) converges to \(u^*\). 

3. Line-search strategy

A comparison between formulas (6) and (12) quantifies the reduction of the radius of the attraction ball of the Newton iterative method with respect to that of Newton method.

Usually, the attraction ball of the Newton iterative method is not much extended. In order to increase the extent of this ball, in the implementation of the method it is necessary to use a damping technique on the inexact Newton step \(\Delta \tilde{u}^{(k)}\). Here, we choose for the next iterate

\[
\tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \alpha_k \Delta \tilde{u}^{(k)}
\]

(15)

where \(0 < \alpha_k < 1\) is the step-length.
We describe a modification of the Newton iterative method for which we can improve the likelihood of convergence from initial approximations that may be outside the attraction ball of the method. The modification involves the introduction of a merit function $\Phi(u) \geq 0$ related to the system $F(u) = 0$, for example

$$\Phi(u) = \frac{1}{2} F(u)^T F(u) = \frac{1}{2} \|F(u)\|^2$$  \hspace{1cm} (16)$$

and a "globalization strategy" for selecting both the search direction $\Delta \tilde{u}^{(k)}$ and the step length $\alpha_k$ to guarantee that the function $\Phi(u)$ can be reduced at each iteration $k$.

Any root $u^*$ of $F(u) = 0$ has $\Phi(u^*) = 0$ and, since $\Phi(u) \geq 0$ for all $u \in \Omega$, each root is at least a local minimizer of $\Phi(u)$; but there may be local minimizers of $\Phi(u)$ that are not solution of the equation $F(u) = 0$. Besides, the choice

$$\Phi(u) = \frac{1}{2} \|F'(u)^{-1} F(u)\|^2$$

for the merit function would be more consistent with the affine invariance theory [9]. However, the merit function (16) is used successfully in many applications (e.g. see [35, p. 257]).

In the framework of the line-search strategy, a “good” iteration method

$$z^{(k+1)} = z^{(k)} + \alpha_k p^{(k)} \hspace{1cm} \alpha_k > 0 \hspace{1cm} k = 0, 1, ...$$  \hspace{1cm} (17)$$

for minimizing a function $\Phi(z)$ must satisfy, at each $k$, the following requirements (see, i.e. [33, Sect. 3.3]):

C1) $\Phi(z^{(k+1)}) < \Phi(z^{(k)})$ whenever $\nabla \Phi(z^{(k)}) \neq 0$.

A good method must be a descent method;

C2) $p^{(k)}^T \nabla \Phi(z^{(k)}) < 0$

A good method must move in a promising direction for minimization at each iteration;

C3) there exists a positive number $\beta < 1$ such that

$$\Phi(z^{(k+1)}) \leq \Phi(z^{(k)}) + \beta \alpha_k p^{(k)}^T \nabla \Phi(z^{(k)})$$

This inequality is called Armijo condition [4];

C4) there exists a positive number $\gamma$ with $0 < \beta < \gamma < 1$ such that

$$p^{(k)}^T \nabla \Phi(z^{(k+1)}) > \gamma p^{(k)}^T \nabla \Phi(z^{(k)})$$

The conditions C3 and C4 are called Wolfe conditions.

If condition C2 is satisfied, then, also condition C1 is satisfied if $\alpha_k$ is positive and sufficiently small.
Indeed, if we set 
\[ \varphi_k(\alpha) = \Phi(z^{(k)} + \alpha p^{(k)}) \]
by the Chain Rule (see e.g. [32, p. 62]) we have
\[ \varphi'_k(0) = \nabla \Phi(z^{(k)})^T p^{(k)} = p^{(k)^T} \nabla \Phi(z^{(k)}) < 0 \]
Hence, for positive value of \( \alpha \) that are sufficiently small, Taylor’s expansion ensures that
\[ \Phi(z^{(k)} + \alpha p^{(k)}) = \varphi_k(\alpha) < \varphi_k(0) = \Phi(z^{(k)}) \]
Therefore,
\[ \Phi(z^{(k+1)}) = \Phi(z^{(k)} + \alpha_k p^{(k)}) < \Phi(z^{(k)}) \]
Conditions C2 and C3 together ensure that, if the step from \( z^{(k)} \) to \( z^{(k+1)} \) is large (that is, \( \|z^{(k+1)} - z^{(k)}\| \) is large), the decrease in merit function values \( \Phi(z^{(k)}) - \Phi(z^{(k+1)}) \) between \( z^{(k)} \) and \( z^{(k+1)} \) must be also large.
Indeed, if we set
\[ \mu = -\frac{\beta p^{(k)^T} \nabla \Phi(z^{(k)})}{\|p^{(k)}\|} > 0 \quad \text{by condition C2} \]
condition C3 yields the inequality
\[ \frac{\Phi(z^{(k)}) - \Phi(z^{(k+1)})}{\|z^{(k)} - z^{(k+1)}\|} = \frac{\Phi(z^{(k)}) - \Phi(z^{(k+1)})}{\alpha_k p^{(k)}} \geq \mu \]
From this we have
\[ \Phi(z^{(k)}) - \Phi(z^{(k+1)}) \geq \mu \|z^{(k)} - z^{(k+1)}\| \]
Thus, if \( \|z^{(k)} - z^{(k+1)}\| \) is large, then \( \Phi(z^{(k)}) - \Phi(z^{(k+1)}) \) must be also large.
Besides condition C1 is automatically satisfied when conditions C2 and C3 are satisfied.
Conditions C2 and C4 together prevent arbitrarily small choice for \( \alpha_k \).
Indeed, condition C4 yields
\[ p^{(k)^T} \nabla \Phi(z^{(k)} + \alpha_k p^{(k)}) > \gamma p^{(k)^T} \nabla \Phi(z^{(k)}) > p^{(k)^T} \nabla \Phi(z^{(k)}) \]
since \( \gamma < 1 \) and \( p^{(k)^T} \nabla \Phi(z^{(k)}) < 0 \) by condition C2.
Hence,
\[ p^{(k)^T} \nabla \Phi(z^{(k)} + \alpha_k p^{(k)}) - p^{(k)^T} \nabla \Phi(z^{(k)}) > (\gamma - 1)p^{(k)^T} \nabla \Phi(z^{(k)}) > 0 \quad (18) \]
This means that \( \alpha_k \) cannot be arbitrarily small. For if we let \( \alpha_k \) approach zero in (18), the left hand side of the inequality approaches zero while the right hand side remains constant at \( (\gamma - 1)p^{(k)^T} \nabla \Phi(z^{(k)}) > 0 \), which is impossible.
Thus criteria C2 and C3 together prevent arbitrarily small choices of \( \alpha_k \).
Given $\beta$, $\gamma$ with $0 < \beta < \gamma < 1$ and $\Phi(z)$ bounded below, it is always possible to select $p^{(k)}$ and $\alpha_k > 0$ so that conditions C1–C4 are satisfied ([39],[40],[8, Theor. 6.3.2], [33, Theor. 3.3.1]).

Once the search direction $p^{(k)}$ has been selected to guarantee that condition C2 is satisfied, it is possible to determine the step length $\alpha_k$ with a backtracking procedure.

We begin by setting $\alpha = 1$ and then we reduce $\alpha$ until we reach a value for which conditions C3 and C4 are satisfied.

A criterion for doing this is to check for failure of the Armijo condition (condition C3); that is, we examine whether

$$
\Phi(z^{(k)} + \alpha p^{(k)}) \geq \Phi(z^{(k)}) + \beta \alpha p^{(k)} T \nabla \Phi(z^{(k)})
$$

and if this inequality holds, we replace $\alpha$ by $\chi \alpha$ for some appropriate $\chi < 1$ and we check again until the failure of the condition (see [8, Sect. 6.3.2]).

Set $\xi_k$ the angle between $p^{(k)}$ and $\nabla \Phi(z^{(k)})$, thus

$$
\cos^2 \xi_k \|\nabla \Phi(z^{(k)})\|^2 < \infty
$$

The convergence scheme for line-search methods can be stated as follows (e.g. see [30, §3.1, §3.2])

- Let $p^{(k)}$ be a descent direction; i.e. $p^{(k)}$ satisfies condition C2;
- Let the damping parameter $\alpha_k$ satisfy Wolfe conditions (C3 and C4);
- Let $L_0 = \{ z \mid \Phi(z) \leq \Phi(z^{(0)}) \}$ be a compact set or let $\Phi(z)$ be bounded below;
- Let $\Phi(z)$ be continuously differentiable in a open set containing $L_0$, and in the same set let $\nabla \Phi(z)$ be Lipschitz–continuous;

Then, the following condition (Zoutendijk condition) holds

$$
\sum_{k \geq 0} \cos^2 \xi_k \|\nabla \Phi(z^{(k)})\|^2 < \infty
$$

thus,

$$
\cos^2 \xi_k \|\nabla \Phi(z^{(k)})\|^2 \rightarrow 0
$$

If $\cos \xi_k > 0$ then $\lim_{k \rightarrow \infty} \|\nabla \Phi(z^{(k)})\| = 0$.

We can summarize the convergence result for a line search method with the following scheme:
Convergence scheme for line-search methods

- Let $p(k)$ be a descent direction for the merit function $\Phi(z)$;
- Let $\alpha_k$ satisfy Wolfe conditions;
- Let $\mathcal{L}_0 = \{ z | \Phi(z) \leq \Phi(z^{(0)}) \}$ be a level set\(^1\), where $z^{(0)}$ is the initial point of the method (17) and
  - $\Phi(z)$ is continuously differentiable on an open set $\tilde{\mathcal{L}} (\mathcal{L}_0 \subset \tilde{\mathcal{L}})$;
  - $\nabla \Phi(z)$ is Lipschitz continuous on $\tilde{\mathcal{L}}$;
  - $\Phi(z)$ is bounded below (e.g. $\mathcal{L}_0$ is a compact set)

\[ \sum_{k \geq 0} \cos^2 \xi_k \| \nabla \Phi(z^{(k)}) \|^2 > \infty \quad \text{Zoutendijk condition} \]

\[ \cos^2 \xi_k \| \nabla \Phi(z^{(k)}) \|^2 \rightarrow 0 \]

$\cos \xi_k > 0 \Rightarrow$

\[ \| \nabla \Phi(z^{(k)}) \| \rightarrow 0 \]

If $\Phi(z) = \| F(z) \|^2$; $\nabla \Phi(z) = 2F'(z)^TF(z)$;

\[ \| \nabla \Phi(z^{(k)}) \| \rightarrow 0 \]

\[ \| F'(z^{(k)})^TF(z^{(k)}) \| \rightarrow 0 \]

$F'(z^{(k)})$ nonsingular $\Rightarrow$

\[ \| F(z^{(k)}) \| \rightarrow 0 \]

- If $\{ z^{(k)} \}$ has a limit point $z^*$ $\implies$ $F(z^{(k)}) \rightarrow F(z^*) = 0$.

\(^1\)\(\mathcal{L}_0\) is a closed set. Indeed, if $z^*$ is a limit point of the sequence $\{ z^{(k)} \}$, where $z^{(k)} \in \mathcal{L}_0$, the hypothesis on continuity of $\Phi(z)$ implies that

\[ \lim_{k \to \infty} \Phi(z^{(k)}) = \Phi(\lim_{k \to \infty} z^{(k)}) = \Phi(z^*) \]

and since $\Phi(z^{(k)}) \leq \Phi(z^{(0)})$ for all $k$, we have that

\[ \lim_{k \to \infty} \Phi(z^{(k)}) \leq \lim_{k \to \infty} \Phi(z^{(0)}) \]

i.e. $\Phi(z^*) \leq \Phi(z^{(0)})$. 

13
In other words, any sequence \( \{ z^{(k)} \} \) generated by (17) with \( p^{(k)} \) satisfying condition C2 and the Wolfe conditions C3 and C4, unless the angle between \( \nabla \Phi(z^{(k)}) \) and \( p^{(k)} \) converges to 90° as \( k \) diverges, either the gradient \( \nabla \Phi(z^{(k)}) \) converges to 0, or \( \Phi(z) \) is unbounded below.

We now investigate the convergence properties of the iteration (15).

First, we identify in (17) the search direction \( p^{(k)} \) with the inexact Newton step \( \Delta \tilde{u}^{(k)} \) that satisfies the inequality (5).

One can show that \( \Delta \tilde{u}^{(k)} \) is a descent direction for the merit function (16) with

\[
\cos \xi_k = -\frac{\nabla \Phi(\tilde{u}^{(k)})^T \Delta \tilde{u}^{(k)}}{\|\nabla \Phi(\tilde{u}^{(k)})\| \|\Delta \tilde{u}^{(k)}\|} \geq \frac{1 - \eta_k}{2K(F'(\tilde{u}^{(k)}))}
\]

(19)

and \( \Phi(\tilde{u}^{(k+1)}) < \Phi(\tilde{u}^{(k)}) \), at least for small positive values of \( \alpha_k \).

Here, \( K(F'(\tilde{u}^{(k)})) \) is the condition number of the Jacobian matrix of \( F(u) \) computed at \( \tilde{u}^{(k)} \), that is

\[
K(F'(\tilde{u}^{(k)})) = \|F'(\tilde{u}^{(k)})\| \|F'(\tilde{u}^{(k)})^{-1}\|
\]

(see e.g. [20]).

If \( K(F'(\tilde{u}^{(k)})) \) is uniformly bounded for all \( k \), then \( \cos \xi_k \) is bounded below.

When \( K(F'(\tilde{u}^{(k)})) \) is large this lower bound is close to zero, and the use of the direction \( \Delta \tilde{u}^{(k)} \) may cause poor performance of the method.

The vector \( \Delta \tilde{u}^{(k)} \), which satisfies the inequality (5) with \( \eta_k \in [0, 1) \), is a descent direction for \( \Phi(u) \) if satisfies condition C2, i.e., if

\[
\nabla \Phi(\tilde{u}^{(k)})^T \Delta \tilde{u}^{(k)} < 0
\]

that in our case \( \nabla \Phi(u) = F'(u)^T F(u) \), takes the form

\[
F(\tilde{u}^{(k)})^T F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} < 0
\]

(20)

which may be written

\[
F(\tilde{u}^{(k)})^T \left( F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} + F(\tilde{u}^{(k)}) - F(\tilde{u}^{(k)}) \right) < 0
\]

and hence as

\[
F(\tilde{u}^{(k)})^T \left( F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} + F(\tilde{u}^{(k)}) \right) < F(\tilde{u}^{(k)})^T F(\tilde{u}^{(k)})
\]

Applying the Cauchy–Schwarz inequality, the condition (26) is satisfied if \( \Delta \tilde{u}^{(k)} \) satisfies

\[
\|F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} + F(\tilde{u}^{(k)})\| < \|F(\tilde{u}^{(k)})\|
\]

(21)

The vector \( \Delta \tilde{u}^{(k)} \) of (5) satisfies the inequality and therefore, it is a descent direction for \( \Phi(u) \) at \( \tilde{u}^{(k)} \).
Now, the condition
\[ \Phi(\tilde{u}^{(k)}) + \alpha \Delta \tilde{u}^{(k)} < \Phi(\tilde{u}^{(k)}) \quad \alpha > 0 \]
is equivalent to
\[ \|F(\tilde{u}^{(k)}) + \alpha \Delta \tilde{u}^{(k)}\| < \|F(\tilde{u}^{(k)})\| \quad (22) \]
Writing (by using mean value theorem (see e.g. [31, p.142]))
\[ F(\tilde{u}^{(k)} + \alpha \Delta \tilde{u}^{(k)}) = F(\tilde{u}^{(k)}) + \alpha F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} + \]
\[ + F(\tilde{u}^{(k)}) + \alpha \Delta \tilde{u}^{(k)} - \alpha F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} - F(\tilde{u}^{(k)}) \]
\[ = F(\tilde{u}^{(k)}) + \alpha F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} + \]
\[ + \int_0^\alpha \left( F'(\tilde{u}^{(k)} + t \Delta \tilde{u}^{(k)}) - F'(\tilde{u}^{(k)}) \right) dt \cdot \Delta \tilde{u}^{(k)} \]
and using (2), we obtain
\[ F(\tilde{u}^{(k)} + \alpha \Delta \tilde{u}^{(k)}) = (1 - \alpha) F(\tilde{u}^{(k)}) + \alpha \left( F(\tilde{u}^{(k)}) + F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} \right) + \]
\[ + G'(\tilde{u}^{(k)}) \int_0^\alpha G'(\tilde{u}^{(k)})^{-1}(G'(\tilde{u}^{(k)} + t \Delta \tilde{u}^{(k)}) - G'(\tilde{u}^{(k)})) dt \cdot \Delta \tilde{u}^{(k)} \]
Using, (7), the triangle inequality gives
\[ \|F(\tilde{u}^{(k)} + \alpha \Delta \tilde{u}^{(k)})\| \leq \left( 1 - \alpha + \alpha \frac{\|F(\tilde{u}^{(k)}) + F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)}\|}{\|F(\tilde{u}^{(k)})\|} \right) \|F(\tilde{u}^{(k)})\| + \]
\[ + \|G'(\tilde{u}^{(k)})\| \int_0^\alpha \lambda ||t \Delta \tilde{u}^{(k)}|| dt \cdot ||\Delta \tilde{u}^{(k)}|| \]
\[ = \left( 1 - \alpha + \alpha \frac{\|F(\tilde{u}^{(k)}) + F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)}\|}{\|F(\tilde{u}^{(k)})\|} \right) \|F(\tilde{u}^{(k)})\| + \]
\[ + \frac{\lambda \alpha^2}{2} ||G'(\tilde{u}^{(k)})|| ||\Delta \tilde{u}^{(k)}||^2 \]
For \( \alpha \) sufficiently small, in the last formula, we may ignore the higher order term and hence satisfy (28) if (27) holds.
By squaring inequality (5) we obtain
\[ 2 \Delta \tilde{u}^{(k)} T F'(\tilde{u}^{(k)}) T F(\tilde{u}^{(k)}) + \|F(\tilde{u}^{(k)})\|^2 + \|F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)}\|^2 \leq \eta_k^2 \|F(\tilde{u}^{(k)})\|^2 \]
Hence
\[ \Delta \tilde{u}^{(k)} T \nabla \Phi(\tilde{u}^{(k)}) = \Delta \tilde{u}^{(k)} T F'(\tilde{u}^{(k)}) T F(\tilde{u}^{(k)}) \leq \frac{\eta_k^2}{2} \|F(\tilde{u}^{(k)})\|^2 \]
But
\[ \|\Delta \tilde{u}^{(k)}\| = \|F'(\tilde{u}^{(k)})^{-1} \left( F(\tilde{u}^{(k)}) + F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} - F(\tilde{u}^{(k)}) \right) \| \]
\[ \leq \|F'(\tilde{u}^{(k)})^{-1} \| \left( \|F(\tilde{u}^{(k)}) + F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)}\| + \|F(\tilde{u}^{(k)})\| \right) \]
\[ \leq \|F'(\tilde{u}^{(k)})^{-1} \| (1 + \eta_k) \|F(\tilde{u}^{(k)})\| \]
Thus, \[
\cos \xi_k = \frac{-\nabla \Phi(\tilde{u}(k))^T \Delta \tilde{u}(k)}{\|\nabla \Phi(\tilde{u}(k))\|\|\Delta \tilde{u}(k)\|}
\geq \frac{(1 - \eta_k^2)\|F'(\tilde{u}(k))\|^2}{2\|F'(\tilde{u}(k))\|\|F'(\tilde{u}(k))^{-1}\|(1 + \eta_k)\|F'(\tilde{u}(k))\|}
= \frac{1 - \eta_k^2}{2K(F'(\tilde{u}(k)))}
\]

Secondly, we determine in (15) the step length \(\alpha_k\) with a backtracking strategy, guaranteeing that the Wolfe conditions are satisfied. With these appropriate choices of \(\Delta \tilde{u}(k)\) and \(\alpha_k\) we have the following result on the convergence of the damped Newton iterative method.

Theorem 3. Suppose that the mapping \(F(u) = Au + G(u)\) is of class \(M(\sigma, \lambda)\) for some \(\sigma > 0\). Suppose that the open set \(B(u^*, \sigma)\) is containing the level set \(L_0 = \{u \mid \Phi(u) \leq \Phi(u^{(0)})\}\), where \(u^{(0)}\) is the initial guess of the Newton iterative method. Suppose that the iteration method (15) is applied for minimizing \(\Phi(u)\), where the search direction \(\Delta \tilde{u}(k)\) is the inexact Newton step that satisfies the inequality (5) with \(\eta_k \leq \eta_{\text{max}} < 1\) and the step length \(\alpha_k\) satisfies the Wolfe conditions.

Furthermore, let us assume that the condition number \(K(F'(u))\) of the Jacobian matrix of \(F(u)\) is bounded above for all \(u \in B(u^*, \sigma)\).

Then, the sequence \(\{\tilde{u}(k)\}, k > 0\), generated by the damped Newton iterative method is well defined and converges to \(u^*\), unique zero of \(F(u)\) in \(B(u^*, \sigma)\).

Proof. Using (9) and (7), we have that for all \(u \in B(u^*, \sigma)\):

\[G'(u) = G'(u^*) + G'(u^*)e(u, u^*)\]

and

\[
\|G'(u)\| \leq \|G'(u^*)\| + \|G'(u^*)\||e(u, u^*)| \tag{23}
\]

By the mean value theorem, we have for \(u \in B(u^*, \sigma)\), \((B(u^*, \sigma)\) is a convex set):

\[F(u) = F(u^*) + \int_0^1 F'(u^* + t(u - u^*))((u - u^*))dt\]
Since $F(u^*) = 0$, we can write

$$
\|F(u)\| \leq \int_0^1 \|F'(u^* + t(u - u^*))\| dt \cdot \|u - u^*\|
$$

$$
\leq \int_0^1 (\|A\| + \|G'(u^* + t(u - u^*))\|) dt \cdot \|u - u^*\|
$$

We use (23). Thus, for all $u \in \mathcal{B}(u^*, \sigma)$:

$$
\|F(u)\| \leq (\|A\| + \|G'(u^*)\|)(1 + \lambda \sigma) \sigma < \infty \quad (24)
$$

By the hypothesis that $K(F'(u)) = \|F'(u)||F'(u)^{-1}\|$ is bounded above for all $u \in \mathcal{B}(u^*, \sigma)$, it follows from (23) that $\|F'(u)^{-1}\|$ is bounded on the set $\mathcal{B}(u^*, \sigma)$.

For any $u, v$ in $\mathcal{B}(u^*, \sigma)$, we have

$$
\|\nabla \Phi(u) - \nabla \Phi(v)\| = \|F'(u)^T F(u) - F'(v)^T F(v)\|
$$

$$
= \|F'(u)^T (F(u) - F(v)) + F'(v)^T (F(u) - F(v))\|
$$

$$
\leq \|F'(u) - F'(v)\| \|F(u)\| + \|F'(v)\| \|F(u) - F(v)\|
$$

$$
= \|G'(u) - G'(v)\| \|F(u)\| + \|G'(v)\| + \|A\| \|u - v\| + \|G(u) - G(v)\|
$$

Using (9) and (7), we have

$$
\|G'(u) - G'(v)\| = \|G'(v)\| \epsilon(u, v) \leq \|G'(v)\| \lambda \|u - v\|
$$

By the result in [31, p. 143], we have

$$
\|G(u) - G(v)\| \leq \sup_{0 \leq t \leq 1} \|G'(u + t(v - u))\| \|u - v\|
$$

For (24) and (23), $\|F(u)\|$ and $\|G'(u)\|$ are bounded above for all $u \in \mathcal{B}(u^*, \sigma)$; thus, there exist the supremum of $\|F(u)\|$ and of $\|G'(u)\|$. If we set

$$
M = \max \{ \sup_{u \in \mathcal{B}(u^*, \sigma)} \Phi(u), \sup_{u \in \mathcal{B}(u^*, \sigma)} \|G'(u)\| \}
$$

we obtain

$$
\|\nabla \Phi(u) - \nabla \Phi(v)\| \leq \lambda \|G'(v)\| \|F(u)\| \|u - v\| + \|G'(v)\| \|A\| \|u - v\| + \|G(u) - G(v)\|
$$

$$
\leq \lambda M (2M)^{1/2} \|u - v\| + (\|A\| + M) \|u - v\|
$$

$$
= \left( \lambda M (2M)^{1/2} + (\|A\| + M)^2 \right) \|u - v\|
$$

Thus, $\nabla \Phi(u)$ is Lipschitz continuous on $\mathcal{B}(u^*, \sigma)$. Besides, $\Phi(u)$ is bounded below by zero on $\mathcal{B}(u^*, \sigma)$.
The iteration method (15), where the search direction $\Delta \tilde{u}^{(k)}$ is the inexact Newton step that satisfies the inequality (5) and the step length $\alpha_k$ satisfies the Wolfe conditions, generates a sequence of iterates $\tilde{u}^{(k)}$ which reduces $\Phi(u)$ at each iteration $k$. Thus, these iterates belong to the level set $L_0 \subset \mathcal{B}(u^*, \sigma)$. Therefore, the hypotheses of Theorem 3.2 in [30, p. 43] are satisfied, obtaining that, for $\cos \xi_k$ defined as in (19), the series
\[
\sum_{k=0}^{\infty} \cos \xi_k \|\nabla \Phi(\tilde{u}^{(k)})\|^2
\]
converges.

By (19) we have
\[
\cos \xi_k \geq \frac{1 - \eta_k}{K(F'(\tilde{u}^{(k)}))} \geq \frac{1 - \eta_{\text{max}}}{K(F'(\tilde{u}^{(k)}))}
\]
Since $K(F'(u))$ is bounded above for all $u \in \mathcal{B}(u^*, \sigma)$, $K(F'(\tilde{u}^{(k)}))$ is uniformly bounded for all $k$. Thus, there exists a positive constant $\delta \in (0, 1)$ such that $\cos \theta_k \geq \delta > 0$ for all $k$ sufficiently large.

Since the series converges, we have ([1, p. 186])
\[
\lim_{k \to \infty} \|\nabla \Phi(\tilde{u}^{(k)})\| = 0
\]
that is
\[
\lim_{k \to \infty} \|F'(\tilde{u}^{(k)})^T F'(\tilde{u}^{(k)})\| = 0
\]
This guarantees that
\[
\lim_{k \to \infty} F'(\tilde{u}^{(k)})^T F'(\tilde{u}^{(k)}) = 0
\]
Since $\|F'(\tilde{u}^{(k)})^{-1}\|$ is uniformly bounded for all $k$, condition (25) is equivalent to
\[
\lim_{k \to \infty} F'(\tilde{u}^{(k)}) = 0
\]
Therefore, the iterates $\tilde{u}^{(k)}$ must converge to a limit point that solves the equation $F(u) = 0$. But in $\mathcal{B}(u^*, \sigma)$ this equation has the unique solution $u^*$. Thus
\[
\lim_{k \to \infty} \tilde{u}^{(k)} = u^* \quad \sharp
\]
The theorem assumes that the level set $L_0$ is bounded; this assures that the function $\Phi(u)$ takes its minimum value at a finite point. It rules out functions such as $\Phi(u) = \sum_{i=1}^{n} e^{u_i}$ that are bounded below (in this case by zero) but only approach this bound in the limit (see [29, p. 316]).
4 A Numerical example

We study a nonlinear finite difference system which corresponds to a diffusion–convection partial differential equation expressed as

\[-d(u_{xx} + u_{yy}) + pu_x + qu_y + \frac{au}{1+bu} = f(x, y) \quad (x, y) \in R\]

\[u(x, y) = 0 \quad (x, y) \in \partial R\]

(26)

where \(R = (0, 1) \times (0, 1)\) and \(f(x, y)\) is a given source term. The parameters \(d, p, q, a\) and \(b\) are given and \(d, a\) and \(b\) are positive.

We suppose that the source term is also a positive function.

Equation (26) is discretized by using the finite difference method. We cover the domain \(R\) with a uniform square mesh \(R_h\) of mesh width \(\Delta x = \Delta y = h = 1/(N + 1)\), where \(N\) is the number of internal nodes in each direction.

We then replace the differential operators with the central difference approximations in \(R_h\). Thus the truncation error at each node \((i, j)\) is \(O(h^2))\), \(1 \leq i, j \leq N\).

This leads to a system of weakly nonlinear equations of the form

\[Au + G(u) = 0\]

(27)

where \(A\) is a \(N^2 \times N^2\) block tridiagonal matrix where the diagonal blocks are tridiagonal and the off-diagonal blocks are diagonal (e.g. [38]); \(G(u)\) is a vector of \(N^2\) components. \(G'(u)\) is an \(N^2 \times N^2\) diagonal matrix having diagonal elements

\[\frac{h^2}{(1 + bu_{ij})^2} a \quad 1 \leq i, j \leq N\]

(28)

\(u_{ij}\) is an approximation at the node \((x_i, y_j)\) of the solution \(u(x_i, y_j)\) of (26).

If \(|p| h < 2\) and \(|q| h < 2\), then the matrix \(A\) is an irreducibly diagonally dominant matrix with positive diagonal entries and nonpositive off-diagonal entries, and, hence, an M–matrix ([31, p. 110]); furthermore, when \(p = q = 0\), it is symmetric and positive definite matrix ([31, p. 107]).

In both the cases, \(A\) can be splitted in the form described in [20, Sect. 2]. This splitting is a variant in the block diagonal part of the Alternating Group Explicit (AGE) decomposition introduced by Evans (see i.e. [17], [18]). Thus, we can use the Arithmetic Mean method as the inner linear solver for the Newton equation (3). The convergence of this method is proved by results in ([26]), ([34]) (see, propositions 1 and 2 in [20]).

We evaluate the efficiency of the Newton–Arithmetic Mean method.

In the following, the tables show the accuracy and the efficiency of the method when we solve problem (26) with the source \(f(x, y)\) chosen to satisfy in \(R\) the given exact solution

\[u^*(x, y) = \sin(\pi x)\sin(\pi y)\]

(29)

and the parameters are \(d = 1; a = 10; b = 0.5\) while \(p\) and \(q\) assume different values.

The termination criterion for the outer iteration is to stop the iteration if

\[\|F(u^{(k)})\| \leq \tau\]

(30)
where \( \tau = \tau_r \| F(\hat{u}^{(0)}) \| + \tau_a \) and the relative error tolerance \( \tau_r \) and the absolute error tolerance \( \tau_a \) are input to the method ([24, p. 73]). Here, \( \tau_r = \tau_a = 10^{-5} \).

If we denote with \( u_{\min} \) and \( u_{\max} \) the smallest and the largest values of \( u_i \geq 0, i = 1, \ldots, N^2 \), respectively, thus for problem (27), the function \( F \) belongs to class \( \mathcal{M} \); indeed

\[
\frac{g'(u)^{-1}(g'(v) - g'(u))}{g'(u)} = \frac{(1 + bu)^2}{ah^2} \left( \frac{ah^2}{(1 + bv)^2} - \frac{ah^2}{(1 + bu)^2} \right)
\]

\[
= (1 + bu)^2 \left( \frac{1 + 2bu + b^2u^2 - 1 - 2bv - b^2v^2}{(1 + bu)^2(1 + bv)^2} \right)
\]

\[
= \frac{1}{(1 + bv)^2} \left( 2b(u - v) + b^2(u^2 - v^2) \right)
\]

\[
= \frac{1}{(1 + bv)^2} \left( 2b + b^2(u + v) \right) (u - v)
\]

\[
\leq 2b(1 + bu_{\max})(u - v)
\]

for all \( u, v \geq 0 \) bounded by \( u_{\max} \), the constant

\[
\lambda = 2b(1 + bu_{\max})
\]

(31)

gives an upper bound for \( \lambda \) in the subset \( \Omega \) of \( \mathbb{R}_n^+ \)

\[
\Omega = \{ u \in \mathbb{R}_n^+ \mid \| u \|_\infty \leq u_{\max} \}
\]

When we consider the case of diffusion \( p = q = 0 \), we have that the smallest and the largest eigenvalues of the symmetric positive definite matrix \( A \) are, respectively ([23, p. 455])

\[
\lambda_{\min}(A) = 8d \sin^2\left(\frac{\pi}{2}h\right)
\]

\[
\lambda_{\max}(A) = 8d \cos^2\left(\frac{\pi}{2}h\right)
\]

Thus,

\[
K(A) \approx \frac{1}{8d \sin^2\left(\frac{\pi}{2}h\right)} \sim \frac{1}{2d \pi h^2}
\]

(32)

Then, if \( u_i \geq 0 \), Theorem 2 requires

\[
\theta_{\max} = \| A^{-1} \| \| G'(u) \| \leq \frac{1}{8d \sin^2\left(\frac{\pi}{2}h\right)} \frac{h^2 a}{(1 + bu_{\min})^2} < 1
\]

(33)

and

\[
\varepsilon_{\max} = \eta_{\max} K(A)(1 + \theta_{\max}) + \theta_{\max} < 1
\]

that is

\[
\eta_{\max} < \frac{1}{K(A)} \frac{1 - \theta_{\max}}{1 + \theta_{\max}}
\]

(34)

This inequality provides at each outer iteration \( k \) of the Newton–iterative method (5) an upper bound for the forcing term \( \eta_k \). This term depends on the condition number \( K(A) \) of the matrix \( A \) and, for ill–conditioned matrices, it may be
excessively small. However, the performance of the Newton–iterative method can be greatly improved by the addition of preconditioning strategies (see Table 5 in [20]).

It is important to note that a suitable set of physical parameters \(d, a, b\) in diffusion problems of a practical interest (see, i.e., [25]) has values such that, under constraints (33) and (34), \(1/\bar{\lambda}\) in (31) defines an attraction ball whose radius (12) is sufficiently large.

We choose \(N = 100\); then, denoting \(u^*_{\text{min}}\) and \(u^*_{\text{max}}\) the minimum and the maximum values of the nonnegative solution (29) at the mesh points respectively, we have the following values:

\[
\begin{align*}
  u^*_{\text{min}} &= 9.67 \cdot 10^{-4}; \quad u^*_{\text{max}} = 0.999; \\
  K(A) &= 4134.31; \quad \theta_{\text{max}} = 0.5; \quad \lambda = 1.5
\end{align*}
\]

From (12), we have the upper bound of \(\bar{\rho}\) equal to

\[
\frac{2}{\lambda 1 + \eta_{\text{max}}} R
\]

where \(R = (1 - \varepsilon_{\text{max}})/\theta_{\text{max}}\); by the expression of \(\varepsilon_{\text{max}}\), we obtain

\[
\gamma_1 - R\gamma_2 = \eta_{\text{max}} K(A)
\]

with \(\gamma_1 = (1 - \theta_{\text{max}})/(1 + \theta_{\text{max}})\) and \(\gamma_2 = \theta_{\text{max}}/(1 + \theta_{\text{max}})\); by hypothesis in Theorem 2, we have \(\gamma_1 > 0, \gamma_2 > 0\) and \(\gamma_1 < 1, \gamma_2 < 1\).

Since we must have \(0 < \eta_{\text{max}} K(A) < 1\), it yields \(0 < \gamma_1 - R\gamma_2 < 1\). That is \(R < \gamma_1/\gamma_2\).

Let \(R_1 = c_1 \gamma_1/\gamma_2\) and \(R_2 = c_2 \gamma_1/\gamma_2\), with \(c_1 < 1, c_2 < 1\); thus

\[
\eta^{(1)}_{\text{max}} = \frac{\gamma_1 - R_1 \gamma_2}{K(A)}; \quad \eta^{(2)}_{\text{max}} = \frac{\gamma_1 - R_2 \gamma_2}{K(A)}
\]

and

\[
\bar{\rho} = \bar{\rho}_1 = \frac{1}{\lambda 1 + \eta^{(1)}_{\text{max}} R_1}; \quad \bar{\rho} = \bar{\rho}_2 = \frac{1}{\lambda 1 + \eta^{(2)}_{\text{max}} R_2}
\]

For \(c_1 = 0.9, c_2 = 0.1\), we have the following values:

\[
\begin{align*}
  R_1 &= 0.8 \quad &R_2 &= 9.7 \cdot 10^{-2} \\
  \eta^{(1)}_{\text{max}} &= 7.9 \cdot 10^{-6} \quad &\eta^{(2)}_{\text{max}} &= 7.1 \cdot 10^{-5} \\
  \bar{\rho}_1 &= 1.17 \quad &\bar{\rho}_2 &= 0.13
\end{align*}
\]

Starting from the initial point \(\hat{u}^{(0)} = (0.1, \ldots, 0.1)^T\), we have \(\|F(\hat{u}^{(0)})\| = 2.02\) and \(\tau = 3.02 \cdot 10^{-5}\); if we choose the forcing parameter \(\eta_k = 10^{-6}\) the initial point \(\hat{u}^{(0)}\) is an interior point of the convergence region with radius \(\rho = \rho_1\) (see the minimum and the maximum values \(u^*_{\text{min}} \) and \(u^*_{\text{max}}\) of the solution), while from \(\eta_k = 10^{-5}\) up to larger values, the initial point \(\hat{u}^{(0)}\) is out of this convergence region.
Here and in the following we indicate the relative error \( e^* = \| \tilde{u}^* - \tilde{u}^{(k*)} \| / \| \tilde{u}^* \| \), where \( k^* \) is the iteration for which the stopping criterium (30) is satisfied.

In Table 1 are reported the number of Newton–AM iterations (outer iterations), the total amount of the Arithmetic Mean iterations (inner iterations) and the norm of the relative error \( e^* \) for the solution of the problem (26), for the diffusion case, with the values of the parameters as described above.

We observe that, in this experiment, we can choose larger values of the forcing term \( \eta_k \) that the convergence is still obtained avoiding an extremely large value of the total amount of the inner iterations, even if the initial point is out of the convergence region.

| Table 1 |
| --- | --- | --- | --- |
| \( \eta_k \) | outer it. | total inner it. | \( e^* \) |
| 10^{-6} | 2 | 9396 | 1.49 \cdot 10^{-4} |
| 10^{-5} | 2 | 7656 | 1.51 \cdot 10^{-4} |
| 10^{-2} | 3 | 4281 | 1.72 \cdot 10^{-5} |
| 10^{-1} | 5 | 3376 | 1.61 \cdot 10^{-4} |

Consider, now, the damped Newton–AM method that we have implemented with three different line–search backtracking strategies.

We consider the Armijo backtracking technique ([4]):

\[
\alpha_k = 1 \\
\text{while } (\| F(\tilde{u}^{(k)} + \alpha_k \Delta \tilde{u}^{(k)}) \| \geq \| F(\tilde{u}^{(k)}) \| + \alpha_k \beta F(\tilde{u}^{(k)})^T F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)}) \text{ do} \\
\alpha_k = \chi \alpha_k \\
\text{endwhile} \\
\tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \alpha_k \Delta \tilde{u}^{(k)}
\]

(35)

the minimum reduction algorithm ([14, p. 410]):

\[
\eta_k = 1 - \chi (1 - \eta_k) \\
\Delta \tilde{u}^{(k)} = \chi \Delta \tilde{u}^{(k)} \\
\text{while } (\| F(\tilde{u}^{(k)} + \Delta \tilde{u}^{(k)}) \| \geq (1 - \eta_k) \| F(\tilde{u}^{(k)}) \|) \text{ do} \\
\eta_k = 1 - \chi (1 - \eta_k) \\
\Delta \tilde{u}^{(k)} = \chi \Delta \tilde{u}^{(k)} \\
\text{endwhile} \\
\tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \Delta \tilde{u}^{(k)}
\]

(36)

and the line–search strategy described in [13] (see also [16]):

\[
\alpha_k = 1 \\
\text{while } (\| F(\tilde{u}^{(k)} + \alpha_k \Delta \tilde{u}^{(k)}) \| \geq (1 - \alpha_k \beta (1 - \eta_k)) \| F(\tilde{u}^{(k)}) \|) \text{ do} \\
\alpha_k = \chi \alpha_k \\
\text{endwhile} \\
\tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \alpha_k \Delta \tilde{u}^{(k)}
\]

(37)

where \( \Delta \tilde{u}^{(k)} \) is the inexact Newton step that satisfies inequality (5) for some \( \eta_k \leq \eta_{\text{max}} < 1 \).
Here, $\beta \in (0, 1)$ and $\chi \in (0, 1)$ (we set $\chi = 0.8$). We observe that the rule stated in (37) is more general than the one in (35). Indeed, if (35) is verified, we have from (4) and (13)

$$
\| F(\tilde{u}^{(k)} + \alpha_k \Delta \tilde{u}^{(k)}) \|^2 \\
\leq \| F(\tilde{u}^{(k)}) \|^2 + \alpha_k \beta F(\tilde{u}^{(k)})^T F'(\tilde{u}^{(k)}) \Delta \tilde{u}^{(k)} \\
\leq \| F(\tilde{u}^{(k)}) \|^2 + \alpha_k \beta \| F(\tilde{u}^{(k)}) \|^2 (\eta_k - 1) \\
\leq (1 - \alpha_k \beta (1 - \eta_k))^2 \| F(\tilde{u}^{(k)}) \|^2
$$

then (37) is verified.

Choices of $\eta_k$ must be motivated by a desire to prevent the forcing term from becoming too small too quickly (in this case we had to solve the Newton equation almost exactly when we are far from the solution, requiring too much inner iterations) and to avoid oversolving phenomenon (see, i.e., [15]), that is, to avoid that, the norm of the inner residual has to be too small, to satisfy the inner stopping rule, without any notable decreasing of the norm of $F$; that provokes no progress toward a solution.

In our experiments we have also considered safeguard values ([15]) for the forcing term $\eta_k$ defined as $\eta_k = \max\{\tilde{\eta}_k, t\}$ for $k = 1, 2, \ldots$ and $\tilde{\eta}_0 = 0.5$, where

$$
\tilde{\eta}_k = \left(\| F(\tilde{u}^{(k)}) \| - \| F(\tilde{u}^{(k-1)}) + F'(\tilde{u}^{(k-1)}) \Delta \tilde{u}^{(k-1)} \|\right) / \| F(\tilde{u}^{(k-1)}) \| \quad (38)
$$

and $t$ is a threshold that we set equal to $10^{-1}$.

A too small value of the forcing term $\eta_k$ can reduce $\| F(\tilde{u}^{(k)}) \|$ at the final stage $k^*$ of the method far beyond the desired level; so the computation of the solution of the linear system for the last stage is more expensive than that is really needed.

We remark that there is a relation between the oversolving and the backtracking rules (36) and (37). When the inner method for the computation of $\Delta \tilde{u}^{(k)}$ in (5) does not provoke oversolving, we have that (e.g. see figures 3.1–3.4 in [15])

$$
\| r^{(k)} \| \simeq \| F(\tilde{u}^{(k)} + \Delta \tilde{u}^{(k)}) \|
$$

then, in general

$$
\| F(\tilde{u}^{(k)} + \Delta \tilde{u}^{(k)}) \| \leq \eta_k \| F(\tilde{u}^{(k)}) \|
$$

For the backtracking techniques (36) and (37) (we note that $\alpha_k = 1$ at the first step of (37)), since $\eta_k < 1$ and $\beta < 1$, we have

$$
(1 - \beta (1 - \eta_k)) \geq \eta_k
$$

Then, generally, backtracking steps (36) or (37) have been performed when oversolving happens in (5).

For instance, consider the Newton–AM method with the backtracking technique (36), with $\tilde{u}^{(0)} = (0.1, \ldots, 0.1)^T$ as initial point, $\eta_k = 0.5$ as forcing term and $\beta = 1 - 10^{-5}$ as backtracking parameter; it gives 23 outer iterations, 5358 total inner iterations, 44 total backtracking steps and the relative error $e^* = 8.3 \cdot 10^{-5}$; we also have $\| F(\tilde{u}^{(0)}) \| = 2$ and $\tau = 3.0 \cdot 10^{-5}$. 

23
For this case, in Table 2, for each outer iteration $k$, are reported: the number of inner iteration of the Arithmetic Mean method, the norm of residual ($nr = \|r^{(k)}\|$) and the norm of $F$ ($nf = \|F(\tilde{u}^{(k)} + \Delta \tilde{u}^{(k)})\|$) after the inner solution and before of checking backtracking condition, the bound of backtracking rule at the first checking ($bb = (1 - \alpha_k \beta (1 - \eta_k))\|F(\tilde{u}^{(k)})\|$, here $\alpha_k = 1$) and the number of the performed backtracking steps.

**Table 2:** $N = 100; d = 1; p = 0; q = 0; a = 10; b = 0.5$

<table>
<thead>
<tr>
<th>$k$</th>
<th>AM it.</th>
<th>$nr$</th>
<th>$nf$</th>
<th>$bb$</th>
<th>backtr. steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.73418</td>
<td>0.73417</td>
<td>1.01229</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.31312</td>
<td>0.31310</td>
<td>0.36709</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.15066</td>
<td>0.15066</td>
<td>0.15655</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>130</td>
<td>$7.516 \cdot 10^{-2}$</td>
<td>$7.604 \cdot 10^{-2}$</td>
<td>$7.533 \cdot 10^{-2}$</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>241</td>
<td>$4.268 \cdot 10^{-2}$</td>
<td>$4.364 \cdot 10^{-2}$</td>
<td>$4.275 \cdot 10^{-2}$</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>260</td>
<td>$2.894 \cdot 10^{-2}$</td>
<td>$2.938 \cdot 10^{-2}$</td>
<td>$2.897 \cdot 10^{-2}$</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>267</td>
<td>$2.152 \cdot 10^{-2}$</td>
<td>$2.176 \cdot 10^{-2}$</td>
<td>$2.154 \cdot 10^{-2}$</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>271</td>
<td>$1.709 \cdot 10^{-2}$</td>
<td>$1.723 \cdot 10^{-2}$</td>
<td>$1.713 \cdot 10^{-2}$</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>273</td>
<td>$1.360 \cdot 10^{-2}$</td>
<td>$1.369 \cdot 10^{-2}$</td>
<td>$1.362 \cdot 10^{-2}$</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>274</td>
<td>$1.138 \cdot 10^{-2}$</td>
<td>$1.145 \cdot 10^{-2}$</td>
<td>$1.138 \cdot 10^{-2}$</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>275</td>
<td>$1.042 \cdot 10^{-2}$</td>
<td>$1.047 \cdot 10^{-2}$</td>
<td>$1.043 \cdot 10^{-2}$</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>276</td>
<td>$8.714 \cdot 10^{-3}$</td>
<td>$8.750 \cdot 10^{-3}$</td>
<td>$8.723 \cdot 10^{-3}$</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>277</td>
<td>$7.283 \cdot 10^{-3}$</td>
<td>$7.308 \cdot 10^{-3}$</td>
<td>$7.294 \cdot 10^{-3}$</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>278</td>
<td>$5.789 \cdot 10^{-3}$</td>
<td>$5.804 \cdot 10^{-3}$</td>
<td>$5.799 \cdot 10^{-3}$</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>279</td>
<td>$3.936 \cdot 10^{-3}$</td>
<td>$3.943 \cdot 10^{-3}$</td>
<td>$3.943 \cdot 10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>280</td>
<td>$2.360 \cdot 10^{-3}$</td>
<td>$2.363 \cdot 10^{-3}$</td>
<td>$2.365 \cdot 10^{-3}$</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>281</td>
<td>$1.180 \cdot 10^{-3}$</td>
<td>$1.181 \cdot 10^{-3}$</td>
<td>$1.181 \cdot 10^{-3}$</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>281</td>
<td>$5.891 \cdot 10^{-4}$</td>
<td>$5.893 \cdot 10^{-4}$</td>
<td>$5.906 \cdot 10^{-4}$</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>281</td>
<td>$2.940 \cdot 10^{-4}$</td>
<td>$2.940 \cdot 10^{-4}$</td>
<td>$2.946 \cdot 10^{-4}$</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>281</td>
<td>$1.467 \cdot 10^{-4}$</td>
<td>$1.467 \cdot 10^{-4}$</td>
<td>$1.470 \cdot 10^{-4}$</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>281</td>
<td>$7.325 \cdot 10^{-5}$</td>
<td>$7.325 \cdot 10^{-5}$</td>
<td>$7.338 \cdot 10^{-5}$</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>281</td>
<td>$3.656 \cdot 10^{-5}$</td>
<td>$3.656 \cdot 10^{-5}$</td>
<td>$3.662 \cdot 10^{-5}$</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>281</td>
<td>$1.825 \cdot 10^{-5}$</td>
<td>$1.825 \cdot 10^{-5}$</td>
<td>$1.828 \cdot 10^{-5}$</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3 shows the results for the Newton–AM method with different backtracking rules, for different values of the initial point $\tilde{u}^{(0)}$ and with $\eta_k = 0.5$. We denote: $u_{0001} = \tilde{u}^{(0)} = (0.01, ..., 0.01)^T$, $u_{001} = \tilde{u}^{(0)} = (0.1, ..., 0.1)^T$, $u_{01} = \tilde{u}^{(0)} = (1.0, ..., 1.0)^T$, $u_{010} = \tilde{u}^{(0)} = (10.0, ..., 10.0)^T$.

In the table, it represents the number of outer iterations with the total number of inner iterations (in brackets), bs the total number of backtracking steps and $e^*$ is the relative error. Here the backtracking parameter $\beta$ is equal to $1 - 10^{-5}$.

In the table are also reported the values of $F(\tilde{u}^{(0)})$ and of $\tau$ for the different choices of the initial points.
Table 3. \( N = 100; d = 1; p = 0; q = 0; a = 10; b = 0.5 \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( u_{001}^{(0)} )</th>
<th>( u_{001}^{(3)} )</th>
<th>( u_{01}^{(3)} )</th>
<th>( u_{10}^{(3)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau )</td>
<td>0.2</td>
<td>1.2 ( \cdot 10^{-5} )</td>
<td>2.0</td>
<td>3.0 ( \cdot 10^{-5} )</td>
</tr>
<tr>
<td>( it )</td>
<td>20.2</td>
<td>20.2 ( \cdot 10^{-4} )</td>
<td>20.2 ( \cdot 10^{-3} )</td>
<td></td>
</tr>
<tr>
<td>( bs )</td>
<td>15(3892)</td>
<td>16(3467)</td>
<td>16(2518)</td>
<td>16(3053)</td>
</tr>
<tr>
<td>( e^* )</td>
<td>8.9 ( \cdot 10^{-5} )</td>
<td>1.3 ( \cdot 10^{-3} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
<td></td>
</tr>
<tr>
<td>( \tau )</td>
<td>0.2</td>
<td>1.2 ( \cdot 10^{-5} )</td>
<td>2.0</td>
<td>3.0 ( \cdot 10^{-5} )</td>
</tr>
<tr>
<td>( it )</td>
<td>20.2</td>
<td>20.2 ( \cdot 10^{-4} )</td>
<td>20.2 ( \cdot 10^{-3} )</td>
<td></td>
</tr>
<tr>
<td>( bs )</td>
<td>15(3892)</td>
<td>16(3467)</td>
<td>16(2518)</td>
<td>16(3053)</td>
</tr>
<tr>
<td>( e^* )</td>
<td>8.9 ( \cdot 10^{-5} )</td>
<td>1.3 ( \cdot 10^{-3} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
<td></td>
</tr>
<tr>
<td>( \tau )</td>
<td>0.2</td>
<td>1.2 ( \cdot 10^{-5} )</td>
<td>2.0</td>
<td>3.0 ( \cdot 10^{-5} )</td>
</tr>
<tr>
<td>( it )</td>
<td>20.2</td>
<td>20.2 ( \cdot 10^{-4} )</td>
<td>20.2 ( \cdot 10^{-3} )</td>
<td></td>
</tr>
<tr>
<td>( bs )</td>
<td>15(3892)</td>
<td>16(3467)</td>
<td>16(2518)</td>
<td>16(3053)</td>
</tr>
<tr>
<td>( e^* )</td>
<td>8.9 ( \cdot 10^{-5} )</td>
<td>1.3 ( \cdot 10^{-3} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. \( N = 100; d = 1; a = 10; b = 0.5 \)

<table>
<thead>
<tr>
<th>( p )</th>
<th>( q )</th>
<th>( \text{out. it.} )</th>
<th>( \text{inn. it.} )</th>
<th>( | F(u^{(k)}) | )</th>
<th>( \tau )</th>
<th>( e^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>16</td>
<td>3467</td>
<td>2.02</td>
<td>2.1 ( \cdot 10^{-5} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>16</td>
<td>3412</td>
<td>2.02</td>
<td>2.1 ( \cdot 10^{-5} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>16</td>
<td>1580</td>
<td>2.03</td>
<td>2.1 ( \cdot 10^{-5} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>16</td>
<td>275</td>
<td>2.13</td>
<td>2.1 ( \cdot 10^{-5} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>16</td>
<td>184</td>
<td>2.56</td>
<td>2.1 ( \cdot 10^{-5} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>15</td>
<td>186</td>
<td>2.56</td>
<td>2.1 ( \cdot 10^{-5} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>15</td>
<td>186</td>
<td>2.56</td>
<td>2.1 ( \cdot 10^{-5} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>15</td>
<td>186</td>
<td>2.56</td>
<td>2.1 ( \cdot 10^{-5} )</td>
<td>1.1 ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>14</td>
<td>87</td>
<td>4.88</td>
<td>4.1 ( \cdot 10^{-5} )</td>
<td>1.5 ( \cdot 10^{-4} )</td>
</tr>
</tbody>
</table>

We observe that, in the case of \( \tilde{u}^{(0)} = (0.1, \ldots, 0.1)^T \), where many backtracking steps occur for the choices (36) and (37), the choice (38) of the forcing term yields the following results for Newton–AM–(37) method: 9 outer iterations, 5090 total inner iterations, 6 total backtracking steps with a relative error \( e^* = 1.5 \cdot 10^{-4} \).

Since the Arithmetic–Mean method for linear systems and the Newton–AM method for weakly nonlinear systems give better results when the problem is nonsymmetric (e.g. see [34], [20]), in Table 4, we report results of the Newton–AM method for different values of \( p \) and \( q \). We report the number of outer iteration (out. it), the total number of inner iterations (inn. it.), the euclidean norm of \( F \) at the initial point, the tolerance of the stopping rule (\( \tau \)) and the relative error \( e^* \).

Here \( \eta_b = 0.5 \), \( \tilde{u}^{(0)} = (0.1, \ldots, 0.1)^T \) and the backtracking rule (36) with \( \beta = 10^{-4} \) does not occur.

All the experiments have been obtained by a double precision Fortran code carried out on Workstation Alpha 21264 EV6 with 667 MHz.

Table 4. \( N = 100; d = 1; a = 10; b = 0.5 \)
5 Conclusions

In this paper the following conclusions can be drawn:

- theorems of local convergence for both Newton and Newton iterative method with adaptive stopping rule are developed for the problem \( F(u) = Au + G(u) = 0 \) when the function \( G' \) satisfies an affine invariant Lipschitz condition.

  In the case of Newton iterative method the radius of attraction ball of the method depends on the condition number of the matrix \( A \).

- In the framework of line–search strategy, a theorem for the convergence of Newton iterative method with damping parameter is stated under the affine invariant Lipschitz assumption on \( G' \).

- The Arithmetic Mean solver for block linear system has been used as iterative solver for the solution with Newton iterative method of a semilinear diffusion convection equation.

  A rational function is the nonlinear source term.

- The numerical experiments are concerning with an analysis of the forcing term, the oversolving phenomenon related with the backtracking steps and the effectiveness of the method related with the symmetry of the problem.

References


