INEXACT PERTURBED NEWTON METHODS AND APPLICATIONS TO A CLASS OF KRYLOV SOLVERS*

Emil Cătinaş†

Abstract

Inexact Newton methods are variant of the Newton method in which each step satisfies only approximately the linear system [1]. The local convergence theory given by the authors of [1] and most of the results based on it consider the error terms as being provided only by the fact that the linear systems are not solved exactly. The few existing results for the general case (when some perturbed linear systems are considered, which in turn are not solved exactly) do not offer explicit formulas in terms of the perturbations and residuals. We extend this local convergence theory to the general case, characterizing the rate of convergence in terms of the perturbations and residuals.

The Newton iterations are then analyzed when, at each step, an approximate solution of the linear system is determined by the following Krylov solvers based on backward error minimization properties: GMRES, GMBACK, MINPERT. We obtain results concerning the following topics: monotone properties of the errors in these Newton–Krylov iterates when the initial guess is taken 0 in the Krylov algorithms; control of the convergence orders of the Newton–Krylov iterations by the magnitude of the backward errors of the approximate steps; similarities of the asymptotical behavior of GMRES and MINPERT when used in a converging Newton method. At the end of the paper, the theoretical results are verified on some numerical examples.

Key words: Nonlinear systems, (inexact) Newton methods, (inexact) perturbed Newton methods, convergence orders, linear systems, backward errors, Krylov methods (GMRES, GMBACK, MINPERT), Newton–Krylov methods.

1 Introduction

Consider the system of nonlinear equations

$$F(x) = 0, \tag{1}$$

where $F : \mathbb{R}^N \to \mathbb{R}^N$ is a nonlinear mapping and suppose that:

(C1) there exists $x^* \in \mathbb{R}^N$ such that $F(x^*) = 0$.

A classical approach for approximating x^* is the Newton method, which results in the following algorithm:

$$\text{Choose an initial approximation } x_0 \in \mathbb{R}^N$$
$$\text{For } k = 0, 1, \dots \text{ until convergence do}$$
(N) $$\text{Solve } F'(x_k) \, s_k = -F(x_k)$$
$$\text{Set } x_{k+1} = x_k + s_k.$$

At each iteration of the Newton method, a routine for solving the resulting linear system must be called. The direct methods for linear systems cannot always offer the exact solution when used in floating point arithmetic and they may be inefficient for large general systems. Usually, some of the iterative methods cannot offer the exact solution after a finite number of steps even in exact arithmetic. Another fact is that, when x_k is far from x^*, it may be worthless to solve exactly the system. These reasons require a convergence analysis which takes into account some error terms.

Several Newton-type methods have been studied. Some sufficient conditions for different convergence orders have been given in [2]–[16] and in the references therein. A characterization of superlinear convergence and of convergence with orders $1 + p$, $p \in (0, 1]$, has been obtained by Dennis and Moré in [17] and [18] for the sequences given by the following quasi-Newton method:

$$x_{k+1} = x_k - B_k^{-1} F(x_k), \quad k = 0, 1, \dots, \quad x_0 \in \mathbb{R}^N,$$

$(B_k)_{k \geq 0} \subset \mathbb{R}^{N \times N}$ being a sequence of invertible matrices.

A characterization of local superlinear convergence and local convergence with orders $1 + p$, $p \in (0, 1]$, of the Newton methods which take into account other error terms has been given in 1982 by Dembo, Eisenstat and Steihaug [1]. They considered in their paper the following *inexact Newton method*

$$\text{Choose an initial approximation } x_0 \in \mathbb{R}^N$$
$$\text{For } k = 0, 1, \dots \text{ until convergence do}$$
(IN) $$\text{Find } s_k \text{ such that } F'(x_k) \, s_k = -F(x_k) + r_k$$
$$\text{Set } x_{k+1} = x_k + s_k.$$

The error terms (*residuals*) r_k represent the amounts by which the solutions s_k (determined in an unspecified manner) fail to satisfy the exact systems (N). Their magnitudes are determined by the *relative residuals* $\|r_k\| / \|F(x_k)\|$, supposed to be bounded by the *forcing sequence* $(\eta_k)_{k \geq 0}$:

(2) $$\frac{\|r_k\|}{\|F(x_k)\|} \leq \eta_k, \quad k = 0, 1, \dots$$

The local convergence analysis given in [1] characterizes the convergence orders of $(x_k)_{k \geq 0}$ given by the IN method in terms of the magnitudes of r_k (see also [19] and [20] for other convergence results). However, in this paper as well as in most others using these results, the error terms r_k are considered to appear only because the exact Newton systems (N) are

solved approximately. But in many situations, it is hard to find the exact value of $F'(x)$, and in some cases even of $F(x)$. On the other hand, $F'(x)$ (or its approximation) and $F(x)$ are both altered when represented in floating point arithmetic.

The question that naturally arises is: what magnitudes can we allow in perturbing the matrices $F'(x_k)$ and the vectors $-F(x_k)$ so that the convergence order of the resulting method does not decrease?

The Newton methods with perturbed linear systems have been considered by several authors (see for instance [9], [11], [21] and the references therein), but these methods were not analyzed with respect to their convergence orders.

Martinez, Parada and Tapia [22] have analyzed the superlinear convergence of the sequences given by *the damped and perturbed quasi-Newton method*

$$x_{k+1} = x_k - \alpha_k B_k^{-1} \left(F(x_k) + r_k \right),$$

where $0 < \alpha_k \leq 1$, $r_k \in \mathbb{R}^N$, $k = 0, 1, \ldots$ and $x_0 \in \mathbb{R}^N$, but their superlinear convergence was not characterized in terms of α_k, B_k, r_k simultaneously.

In [23], Ypma studied the case when the matrices $F'(x_k)$ and the vectors $-F(x_k)$ are calculated approximately, the resulted linear systems being solved inexactly. However, the residuals were considered to be incorporated in those perturbed linear systems. Consequently, the obtained convergence results do not offer explicit formulas in terms of the magnitude of the perturbations and residuals. Recently, Cores and Tapia [24] have considered the exact solving of perturbed linear systems, but they have obtained only sufficient conditions on the perturbations for different convergence orders to be attained.

We consider here the perturbations $(\Delta_k)_{k \geq 0} \subset \mathbb{R}^{N \times N}$ and $(\delta_k)_{k \geq 0} \subset \mathbb{R}^N$, being led to the study of the *inexact perturbed Newton method*

(IPN) $\qquad (F'(x_k) + \Delta_k) s_k = (-F(x_k) + \delta_k) + \hat{r}_k$

$\qquad\qquad x_{k+1} = x_k + s_k, \quad k = 0, 1, \ldots, \quad x_0 \in \mathbb{R}^N.$

The terms $\hat{r}_k$ denote the residuals of the approximate solutions s_k of the linear systems $(F'(x_k) + \Delta_k) s = -F(x_k) + \delta_k$. When these systems are assumed to be solved exactly (i.e. $\hat{r}_k = 0$, $k = 0, 1, \ldots$ in the IPN method), we call the resulting method a *perturbed Newton* (PN) method:

(PN) $\qquad (F'(x_k) + \Delta_k) s_k = -F(x_k) + \delta_k.$

This frame will be useful for the study of some Newton–Krylov methods in connection with backward errors.

The paper is structured as follows. In Section 2 we give two convergence results for the IPN method. In Section 3 we characterize the superlinear convergence and the convergence with orders $1 + p$, $p \in (0, 1]$, of the IPN method, and we also provide some sufficient conditions. In Section 4 we analyze the Newton methods in which the linear systems from each step are solved by GMRES, GMBACK or MINPERT. We verify the obtained theoretical results on numerical examples performed on two test problems (Section 5) and we end the paper with some concluding remarks.

Throughout the paper we consider the Euclidean and an arbitrary given norm on $\mathbb{R}^N$ (denoted by $\|\cdot\|_2$ resp. $\|\cdot\|$) together with their induced operator norms. We also use the

Frobenius norm of a matrix $Z \in \mathbb{R}^{M \times N}$, defined as $\|Z\|_F = \sqrt{\operatorname{tr}(ZZ^t)}$. We use the column notation for vectors; when we write vectors (or matrices) inside square brackets, we consider the matrix containing on columns those vectors (or the columns of those matrices). The same convention is used when joining rows. As usual, the vectors e_i, $i = 1, \ldots, n$, form the standard basis in $\mathbb{R}^n$, n being clear from the context.

2 Convergence Results for Inexact Perturbed Newton Methods

The common conditions for the local convergence analysis of the Newton method are condition **(C1)** and the following additional conditions:

(C2) the mapping F is differentiable on a neighborhood of x^* and F' is continuous at x^*;

(C3) the Jacobian $F'(x^*)$ is nonsingular.

These conditions ensure that x^* is a point of attraction for the Newton method, i.e. there exists $\varepsilon > 0$ such that $(x_k)_{k \geq 0}$ given by the Newton method converges to x^* for any initial approximation $x_0 \in \mathbb{R}^N$ with $\|x_0 - x^*\| < \varepsilon$. Moreover, the convergence is q-superlinear[3]; see [2, Th. 10.2.2], and also [10, Th. 4.4].

In the convergence analysis of the IPN iterations, we assume that

(C4) the perturbations Δ_k are such that the matrices $F'(x_k) + \Delta_k$ are nonsingular for $k = 0, 1, \ldots$

Though different in notations, this condition is in fact similar to the one considered for the quasi-Newton iterates, which requires a sequence of invertible matrices $(B_k)_{k \geq 0}$.

The following sufficient condition for the convergence of the IN method was proved by Dembo, Eisenstat and Steihaug:

Theorem 2.1. [1] *Assume Conditions* **(C1)**–**(C3)** *and* $\eta_k \leq \eta_{\max} < t < 1$, $k = 0, 1, \ldots$. *There exists* $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\| \leq \varepsilon$, *then the sequence of the IN iterates* $(x_k)_{k \geq 0}$ *satisfying* (2) *converges to* x^*. *Moreover, the convergence is linear, in the sense that*

$$\|x_{k+1} - x^*\|_* \leq t \|x_k - x^*\|_*, \quad k = 0, 1, \ldots,$$

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

Using this theorem, we obtain the following convergence results for the IPN method:

Theorem 2.2. *Assume Conditions* **(C1)**–**(C4)** *and* $\eta_k \leq \eta_{\max} < t < 1$, $k = 0, 1, \ldots$. *There exists* $\varepsilon > 0$ *such that, if* $\|x_0 - x^*\| \leq \varepsilon$ *and*

$$\left\| \Delta_k \left(F'(x_k) + \Delta_k\right)^{-1} F(x_k) + \left(I - \Delta_k \left(F'(x_k) + \Delta_k\right)^{-1}\right)(\delta_k + \hat{r}_k) \right\| \leq \eta_k \|F(x_k)\|,$$

for $k = 0, 1, \ldots$, *then the sequence of the IPN iterates* $(x_k)_{k \geq 0}$ *converges to* x^*, *the convergence being linear:*

$$\|x_{k+1} - x^*\|_* \leq t \|x_k - x^*\|_*, \quad k = 0, 1, \ldots.$$

---

[3]For definitions and results concerning convergence orders see [2, ch. 9], and also [10], [25].

Proof. The IPN can be viewed as an IN method:

$$
\begin{aligned}
s_k &= -\left(F'\left(x_k\right)+\Delta_k\right)^{-1} F\left(x_k\right)+\left(F'\left(x_k\right)+\Delta_k\right)^{-1}\left(\delta_k+\hat{r}_k\right); \\
F'\left(x_k\right) s_k &= -\Delta_k s_k - F\left(x_k\right)+\delta_k+\hat{r}_k \\
&= -F\left(x_k\right)+\Delta_k\left(F'\left(x_k\right)+\Delta_k\right)^{-1} F\left(x_k\right) - \\
&\quad\; \Delta_k\left(F'\left(x_k\right)+\Delta_k\right)^{-1}\left(\delta_k+\hat{r}_k\right)+\delta_k+\hat{r}_k \\
&= -F\left(x_k\right)+\Delta_k\left(F'\left(x_k\right)+\Delta_k\right)^{-1} F\left(x_k\right) + \\
&\quad\; \left(I-\Delta_k\left(F'\left(x_k\right)+\Delta_k\right)^{-1}\right)\left(\delta_k+\hat{r}_k\right).
\end{aligned}
$$

Denoting

$$
(3) \qquad r_k = \Delta_k\left(F'\left(x_k\right)+\Delta_k\right)^{-1} F\left(x_k\right)+\left(I-\Delta_k\left(F'\left(x_k\right)+\Delta_k\right)^{-1}\right)\left(\delta_k+\hat{r}_k\right),
$$

the conclusion follows from Theorem 2.1. $\qquad\square$

Corollary 2.1. *Assume Conditions* **(C1)**–**(C4)**. *There exists $\varepsilon > 0$ such that if $\|x_0-x^*\| \leq \varepsilon$ and*

$$
\left\|\Delta_k\left(F'\left(x_k\right)+\Delta_k\right)^{-1}\right\| \leq q_1 < 1, \quad k = 0, 1, \ldots,
$$

$$
\|\delta_k\| + \|\hat{r}_k\| \leq \tfrac{\eta_k}{1+q_1}\|F\left(x_k\right)\|, \quad \text{where} \quad \eta_k \leq q_2 < 1-q_1, \quad k = 0, 1, \ldots,
$$

then the sequence of the IPN iterates $(x_k)_{k \geq 0}$ converges to x^*. Moreover, the convergence is linear:

$$
\|x_{k+1} - x^*\|_* \leq t\,\|x_k - x^*\|_*, \quad k = 0, 1, \ldots,
$$

where $t = q_1 + q_2$.

Proof. The proof is easily obtained from the previous result making use of the hypotheses. $\qquad\square$

Remark 2.1. The idea of reducing certain perturbed Newton methods to IN iterations, which is used in the proof of Theorem 2.2, can be found in the work of several authors (see for example [9], [10] and [26]); the inexact secant methods considered by us in [27] are in fact instances of the IPN model but still it was [28] that inspired us to consider the IPN iterations. $\qquad\square$

3 Convergence Orders of Inexact Perturbed Newton Methods

Stronger conditions imposed to the continuity of F' at x^* offer higher convergence orders for the Newton method. Namely, if F' is Hölder continuous at x^* with exponent p, $p \in (0, 1]$, i.e., if there exist $\varepsilon > 0$ and $L \geq 0$ such that

$$
\|F'\left(x\right) - F'\left(x^*\right)\| \leq L\,\|x - x^*\|^p, \quad \text{for} \ \|x - x^*\| \leq \varepsilon,
$$

then the Newton method converges locally with q-order at least $1 + p$; see [2, Th. 10.2.2], and also [10, Th. 4.4].

For the inexact Newton methods, Dembo, Eisenstat and Steihaug proved the following result:

Theorem 3.1. [1] *Assume that Conditions* **(C1)**–**(C3)** *hold and that the inexact Newton iterates* $(x_k)_{k \geq 0}$ *converge to* x^*. *Then* $x_k \to x^*$ q-*superlinearly if and only if*

$$\|r_k\| = o\left(\|F(x_k)\|\right), \quad as\ k \to \infty.$$

Moreover, if F' *is Hölder continuous at* x^* *with exponent* p, $p \in (0, 1]$, *then* $x_k \to x^*$ *with* q-*order at least* $1 + p$ *if and only if*

$$\|r_k\| = \mathcal{O}\left(\|F(x_k)\|^{1+p}\right), \quad as\ k \to \infty.$$

We obtain the following result for the inexact perturbed Newton methods.

Theorem 3.2. *Assume that Conditions* **(C1)**–**(C4)** *hold and that the iterates* $(x_k)_{k \geq 0}$ *given by the IPN method converge to* x^*. *Then* $x_k \to x^*$ q-*superlinearly if and only if*

$$\left\| \Delta_k \left(F'(x_k) + \Delta_k\right)^{-1} F(x_k) + \left(I - \Delta_k \left(F'(x_k) + \Delta_k\right)^{-1}\right) (\delta_k + \hat{r}_k) \right\| = o\left(\|F(x_k)\|\right),$$

as $k \to \infty$. *Moreover, if* F' *is Hölder continuous at* x^* *with exponent* p, *then* $x_k \to x^*$ *with* q-*order at least* $1 + p$ *if and only if*

$$\left\| \Delta_k \left(F'(x_k) + \Delta_k\right)^{-1} F(x_k) + \left(I - \Delta_k \left(F'(x_k) + \Delta_k\right)^{-1}\right) (\delta_k + \hat{r}_k) \right\| = \mathcal{O}\left(\|F(x_k)\|^{1+p}\right),$$

as $k \to \infty$.

Proof. The proof is obtained from the previous theorem by using (3). $\square$

In the following result, we characterize the convergence orders of the IPN iterates in terms of the rate of convergence to zero of residuals and perturbations.

Corollary 3.1. *Assume that*

(a) *Conditions* **(C1)**–**(C4)** *hold;*

(b) $\Delta_k \to 0$, $\delta_k \to 0$ *and* $\hat{r}_k \to 0$ *as* $k \to \infty$;

(c) *the sequence* $(x_k)_{k \geq 0}$ *given by the IPN method converges to* x^*.

In addition, if

$$
\begin{aligned}
\|\delta_k\| &= o\left(\|F(x_k)\|\right) \quad and \\
\|\hat{r}_k\| &= o\left(\|F(x_k)\|\right) \quad as\ k \to \infty,
\end{aligned}
$$

then $x_k \to x^*$ q-*superlinearly.*

6

Under the same assumptions (a)–(c), *if additionally* F' *is Hölder continuous at* x^* *with exponent* p, *and if*

$$
\begin{aligned}
\|\Delta_k\| &= \mathcal{O}\left(\|F\left(x_k\right)\|^p\right), \\
\|\delta_k\| &= \mathcal{O}\left(\|F\left(x_k\right)\|^{1+p}\right), \\
\|\hat{r}_k\| &= \mathcal{O}\left(\|F\left(x_k\right)\|^{1+p}\right) \quad as\ k \to \infty,
\end{aligned}
$$

then $x_k \to x^*$ *with q-order at least* $1 + p$.

The corresponding results for the r-convergence orders of the IPN iterates may be stated in a similar manner, taking into account the existing results for the IN methods; see [1, Th. 3.4 and Cor. 3.5].

4   Applications to Krylov Solvers Based on Backward Error Minimization Properties

Consider the nonsymmetric linear system

(4)
$$Ax = b,$$

with $A \in \mathbb{R}^{N \times N}$ nonsingular and $b \in \mathbb{R}^N$. The Krylov methods for solving such a system when the dimension N is large are methods based on the Krylov subspaces, defined for any initial approximation $x_0 \in \mathbb{R}^N$ as

$$\mathcal{K}_m = \mathcal{K}_m\left(A, r_0\right) = \operatorname{span}\left\{r_0, Ar_0, \ldots, A^{m-1}r_0\right\},$$

where $r_0 = b - Ax_0$ is the initial residual and $m \in \{1, \ldots, N\}$.

The *normwise backward error* of an approximate solution $\tilde{x}$ of (4) was introduced by Rigal and Gaches [29] and is defined by:

$$\Pi\left(\tilde{x}\right) = \min\left\{\varepsilon : \left(A + \Delta_A\right)\tilde{x} = b + \Delta_b, \ \|\Delta_A\|_F \le \varepsilon\|E\|_F, \ \|\Delta_b\|_2 \le \varepsilon\|f\|_2\right\},$$

where the parameters $E \in \mathbb{R}^{N \times N}$ and $f \in \mathbb{R}^N$ are arbitrary. The value of $\Pi\left(\tilde{x}\right)$ is

$$\Pi\left(\tilde{x}\right) = \frac{\|b - A\tilde{x}\|_2}{\|E\|_F \cdot \|\tilde{x}\|_2 + \|f\|_2},$$

and the minimum is attained by the *backward errors*

$$
\begin{aligned}
\Delta_A &= \frac{\|E\|_F \cdot \|\tilde{x}\|_2}{\|E\|_F \cdot \|\tilde{x}\|_2 + \|f\|_2}\left(b - A\tilde{x}\right)z^t, \quad \text{with } z = \frac{1}{\|\tilde{x}\|_2^2}\tilde{x}, \\
\Delta_b &= -\frac{\|f\|_2}{\|E\|_F \cdot \|\tilde{x}\|_2 + \|f\|_2}\left(b - A\tilde{x}\right).
\end{aligned}
$$

We shall analyze the behavior of the following three Krylov solvers when used for the linear systems arising in the Newton methods: GMRES, GMBACK and MINPERT. Each of these solvers is based on backward error minimization properties, as we shall see.

In order to use some existing notations, we prefer to denote hereafter by y^* the solution of the nonlinear system (1) and by y_k the iterates from the Newton methods, using x in the involved linear systems.

4.1 Newton-GMRES Method

Given an initial approximation $x_0 \in \mathbb{R}^N$, the GMRES method of Saad and Schultz [30] uses the Arnoldi process (see [31]) to construct an orthonormal basis $\{v_1, \ldots, v_m\}$ in the Krylov subspace $\mathcal{K}_m$. The approximation $x_m^{GM} \in x_0 + \mathcal{K}_m$ is then determined such that

$$(5) \qquad \left\| b - Ax_m^{GM} \right\|_2 = \min_{x_m \in x_0 + \mathcal{K}_m} \left\| b - Ax_m \right\|_2.$$

Kasenally has noted in [32] that the minimizing property of x_m^{GM} may be expressed in terms of the backward errors, namely

$$\min_{x_m \in x_0 + \mathcal{K}_m} \| b - Ax_m \|_2 = \min_{x_m \in x_0 + \mathcal{K}_m} \left\{ \| \Delta_b \|_2 : Ax_m = b - \Delta_b \right\},$$

i.e., x_m^{GM} minimizes over $x_0 + \mathcal{K}_m$ the backward error Δ_b, assuming $\Delta_A = 0$.

The solution x_m^{GM} is determined roughly in the following way (see also [33]):

Arnoldi

- Determine $V_m = [v_1 \ \ldots \ v_m] \in \mathbb{R}^{N \times m}$ and the upper Hessenberg matrix $\bar{H}_m \in \mathbb{R}^{(m+1) \times m}$;

GMRES

- Find the exact solution y_m^{GM} of the following least squares problem in $\mathbb{R}^m$

$$\min_{y_m \in \mathbb{R}^m} \left\| \beta e_1 - \bar{H}_m y_m \right\|_2,$$

  where $\beta = \| r_0 \|_2$;

- Set $x_m^{GM} = x_0 + V_m y_m^{GM}$.

The GMRES method is used in an iterative fashion. Saad and Schultz proved that, at each step, the solution x_m^{GM} is uniquely determined and that the algorithm breaks down in the Arnoldi method only if the exact solution has been reached. Moreover, the process terminates in at most N steps. In the results presented below, we shall assume, as usually, that the Krylov methods do not continue after the exact solution has been determined.

First, we introduce some notations for the restarted GMRES iterations in order to express some relations they satisfy.

Since only the small values of m are attractive in practice, an upper bound $\bar{m} \in \{1, \ldots, N-1\}$ is usually fixed for the subspace dimensions. If after $\bar{m}$ steps the computed solution does not have a sufficiently small residual, the GMRES method is restarted, taking for the initial approximation the last computed solution. We denote by $x_m^{GM(0)}$, $m \in \{1, \ldots, \bar{m}\}$, the first $\bar{m}$ solutions, by $x_m^{GM(1)}$, $m \in \{1, \ldots, \bar{m}\}$, the $\bar{m}$ solutions from the first restart, and so on. The value of the initial approximation x_0 will be clear from the context. For the nonrestarted version we shall use the common notations x_m^{GM}, while for a

generic GMRES solution[4], we shall simply write x^{GM}; the corresponding notations for the k-th correction from a Newton-GMRES method will be $s_{k,m}^{GM}$ resp. s_k^{GM}.

The choice $x_0 = 0$ in the Krylov solvers is a popular one when no better guess is known (see, e.g., [26], [34] and [35]). With the above notations, the following result is immediately obtained from the properties of the GMRES solutions. Certain affirmations from this result are more or less explicitly stated in some papers dealing with GMRES (see for example [26], [36] and [37]).

Proposition 4.1. *Consider the linear system (4) and the initial approximation $x_0 = 0$. Then the following statements are true:*

- *For any $m \in \{1, \ldots, N\}$, the residual r_m^{GM} of the GMRES solution satisfies*

$$\left\| r_m^{GM} \right\|_2 \leq \|b\|_2 .$$

  Moreover, the inequality is strict if and only if the solution x_m^{GM} is nonzero.

- *The residuals associated to the N successive solutions satisfy*

$$0 = \left\| r_N^{GM} \right\|_2 \leq \left\| r_{N-1}^{GM} \right\|_2 \leq \ldots \leq \left\| r_1^{GM} \right\|_2 \leq \|b\|_2 .$$

  The inequality between the norms of two consecutive residuals is strict if and only if the corresponding GMRES solutions are distinct.

- *For any fixed upper bound $\bar{m} \in \{1, \ldots, N-1\}$, the residuals of the restarted GMRES method satisfy*

$$\ldots \leq \left\| r_1^{GM(l+1)} \right\|_2 \leq \left\| r_{\bar{m}}^{GM(l)} \right\|_2 \leq \ldots \leq \left\| r_1^{GM(l)} \right\|_2 \leq$$
$$\leq \left\| r_{\bar{m}}^{GM(l-1)} \right\|_2 \leq \ldots \leq \left\| r_1^{GM(0)} \right\|_2 \leq \|b\|_2 .$$

  The inequality between the norms of two consecutive residuals is strict if and only if the corresponding GMRES solutions are distinct; the residuals may eventually get to zero or may indefinitely stagnate, depending on the problem.

Considering the linear systems from the Newton method we easily get the following proposition.

Proposition 4.2. *Let $y \in D$ be an element for which the derivative $F'(y)$ is nonsingular. Applying the GMRES method with $x_0 = 0$ to the linear system $F'(y)s = -F(y)$ then the residual satisfies for all $m \in \{1, \ldots, N\}$*

$$\left\| r_m^{GM} \right\|_2 \leq \|F(y)\|_2 .$$

Moreover, the above inequality is strict if and only if the correction s_m^{GM} is nonzero.

---

[4]In such a case we assume that the initial approximation $x_0 \in \mathbb{R}^N$, the upper bound $\bar{m} \in \{1, .., N-1\}$, the number of (eventual) restarts $l \geq 0$ and the number of (final, if $l \geq 1$) iterations may be arbitrary.

We have chosen to state the result corresponding only to the first part of Proposition 4.1. The other results are similarly enounced.

Brown [26] and Brown and Saad [38] considered the solving of (1) by global minimization

$$\min_{y \in \mathbb{R}^N} f(y) = \min_{y \in \mathbb{R}^N} F(y)^t F(y),$$

when $F : \mathbb{R}^N \to \mathbb{R}^N$. They obtained that any nonzero correction from a certain step of the Newton-GMRES method with $x_0 = 0$ in GMRES is a descent direction for the above minimization problem. Now we are able to describe this positive behavior in terms of the distances to the solution.

Theorem 4.1. *Assume that the mapping F satisfies Conditions* **(C1)**–**(C3)** *and consider a current approximation $y_c \neq y^*$ for y^*. Let s^{GM} be a nonzero approximate solution of the linear system $F'(y_c) s = -F(y_c)$ provided by GMRES, assuming the initial guess $x_0 = 0$. Let $y_+ = y_c + s^{GM}$, denote $\eta = \left\| r^{GM} \right\|_2 / \left\| F(y_c) \right\|_2$ and take $t \in (\eta, 1)$. If y_c is sufficiently close to y^* then*

$$\|y_+ - y^*\|_* \leq t \|y_c - y^*\|_*$$

where $\|y\|_* = \|F'(y^*) y\|_2$.

Proof. The thesis can be easily proved with slight modifications of the proof of Theorem 2.1, given in [1]. □

The above result says that, when using GMRES (with $x_0 = 0$) in the Newton method, any nonzero correction improves the current outer approximation, provided that approximation is sufficiently good. Though this theorem guarantees a nonincreasing curve for the errors of the Newton-GMRES iterations starting from a sufficiently good approximation, the local convergence is not granted. It is sufficient to notice that, considering a sequence obeying at each step the hypotheses of the above result, there may appear the situation when the set of the forcing terms has 1 as accumulation point, in which case Theorem 2.1 cannot be applied.

The following theoretical example shows that, when the dimension of the Krylov subspaces are smaller than N, the outer iterations may stagnate at the initial approximation $y_0 \neq y^*$, no matter how close to y^* we choose y_0.

Example 4.1. Consider $F : \mathbb{R}^N \to \mathbb{R}^N$, $F(y) = \left(y^{(N)}, y^{(1)}, y^{(2)}, \ldots, y^{(N-1)}\right)^t$, $y = \left(y^{(1)}, \ldots, y^{(N)}\right)^t \in \mathbb{R}^N$, for which

$$F'(y) \equiv \begin{pmatrix} 0 & & & 1 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix} = [e_2 \ e_3 \ \ldots \ e_N \ e_1] =: A.$$

The Newton method for solving $F(y) = 0$ should yield the unique solution $y^* = 0$ in one iteration for any $y_0 \in \mathbb{R}^N$, whenever the corresponding linear system is solved exactly. However, when the correction is approximately determined, the situation changes, and we may use the IN setting.

Taking $y_0 = -he_N$ for some arbitrarily small $h \neq 0$, we must solve $F'(y_0) s = -F(y_0)$, i.e. $Ax = b$ with $b = he_1$.

Applying $m \leq N - 1$ steps of the Arnoldi process with $x_0 = 0$, we obtain $V_m = [e_1 \ \ldots \ e_m] \in \mathbb{R}^{N \times m}$ and $\bar{H}_m = [e_2 \ \ldots \ e_{m+1}] \in \mathbb{R}^{(m+1) \times m}$.

The GMRES solution is given by (see [36] and [39])

$$x_m^{GM} = x_0 + V_m \left(\bar{H}_m^t \bar{H}_m \right)^{-1} \bar{H}_m^t V_{m+1}^t b,$$

such that $V_{m+1}^t b = he_1$, $\bar{H}_m^t he_1 = 0$, and so $x_m^{GM} = s_{0,m}^{GM} = 0$, for all $m = 1, \ldots, N - 1$.

We also note that the restarting version of the GMRES yields the same result whenever $x_0 = 0$ and $\bar{m}, m \leq N - 1$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

This theoretical example shows that the dimension of the Krylov subspaces may sometimes be crucial for the efficiency of the Newton-GMRES method, and that the use of the restarted version of GMRES cannot overcome the difficulties. The stagnation of the GMRES algorithm was first reported on in [36] (where A was as above and for b stood e_1), but we are not aware of any extension to a (non)linear mapping in order to show such a stagnation of the Newton-GMRES method.

4.2 Newton-GMBACK Method

The GMBACK algorithm for solving the linear system (4) was introduced by Kasenally in [32]. Given $x_0 \in \mathbb{R}^N$ it computes a vector $x_m^{GB} \in x_0 + \mathcal{K}_m$ which minimizes the backward error in the matrix A, assuming $\Delta_b = 0$:

$$\min_{x_m \in x_0 + \mathcal{K}_m} \|\Delta_A\|_F \quad \text{subject to} \quad (A - \Delta_A) x_m = b.$$

The following steps are performed for determining x_m^{GB}:

Arnoldi

- Compute V_m and $\bar{H}_m$;

GMBACK

- Let $\beta = \|r_0\|_2$,

  $\hat{H}_m = \begin{bmatrix} -\beta e_1 & \bar{H}_m \end{bmatrix} \in \mathbb{R}^{(m+1) \times (m+1)}, \quad \hat{G}_m = \begin{bmatrix} x_0 & V_m \end{bmatrix} \in \mathbb{R}^{N \times (m+1)},$

  $P = \hat{H}_m^t \hat{H}_m \in \mathbb{R}^{(m+1) \times (m+1)} \quad \text{and} \quad Q = \hat{G}_m^t \hat{G}_m \in \mathbb{R}^{(m+1) \times (m+1)};$

- Determine an eigenvector u_{m+1} corresponding to the smallest eigenvalue λ_{m+1}^{GB} of the generalized eigenproblem $Pu = \lambda Q u$;

- If the first component $u_{m+1}^{(1)}$ is nonzero, compute the vector $y_m^{GB} \in \mathbb{R}^m$ by scaling u_{m+1} such that

  $$\begin{bmatrix} 1 \\ y_m^{GB} \end{bmatrix} = \frac{1}{u_{m+1}^{(1)}} u_{m+1};$$

- Set $x_m^{GB} = x_0 + V_m y_m^{GB}$.

This algorithm may lead to two possible breakdowns, either in the Arnoldi method or in the scaling of u_{m+1}. The first is, as in the case of GMRES, a happy breakdown, because the solution may be determined exactly using $\bar{H}_m$ and V_m. The second appears when all the eigenvectors associated to λ_{m+1}^{GB} have the first component zero, the inevitable divisions by zero leading to uncircumventible breakdowns. In such a case either m is increased or the algorithm is restarted with a different initial approximation x_0. We shall assume in the following analysis that x_m^{GB} exists.

Kasenally proved that for any $x_0 \in \mathbb{R}^N$ and $m \in \{1, \ldots, N\}$, the backward error $\Delta_{A,m}^{GB}$ corresponding to the GMBACK solution satisfies

$$(6) \qquad \left\| \Delta_{A,m}^{GB} \right\|_F = \sqrt{\lambda_{m+1}^{GB}}.$$

The Newton-GMBACK iterates may be written in two equivalent ways, taking into account the properties of the Krylov solutions:

$$F'(y_k) s_k^{GB} = -F(y_k) + r_k^{GB},$$
$$\left(F'(y_k) - \Delta_{A_k}^{GB} \right) s_k^{GB} = -F(y_k),$$

$k = 0, 1, \ldots, y_0 \in D$, where we considered $A_k = F'(y_k)$ and $b_k = -F(y_k)$. There are three results which may be applied to characterize the high convergence orders of these sequences (the corresponding enounces are left to the reader): Theorem 2.1 of Dembo, Eisenstat and Steihaug, the results of Dennis and Moré for the quasi-Newton methods (see [17] and [18]) and Theorem 2.2 for IPN iterates (in which we must take $\delta_k = \hat{r}_k = 0$, $k = 0, 1, \ldots$). Naturally, these results must be equivalent, but we do not analyze here this aspect.

Concerning the sufficient conditions, the convergence orders of the Newton-GMBACK method may be controlled by the computed eigenvalues λ_k^{GB} from GMBACK.

Theorem 4.2. *Consider the sequence of the Newton-GMBACK iterates $y_{k+1} = y_k + s_k^{GB}$, where s_k^{GB} satisfies*

$$\left(F'(y_k) - \Delta_{A_k}^{GB} \right) s_k^{GB} = -F(y_k), \quad k = 0, 1, \ldots$$

and assume the following:

a) *Conditions* **(C1)**–**(C4)** *hold;*

b) *the derivative F' is Hölder continuous with exponent p at y^*;*

c) *the sequence $(y_k)_{k\geq0}$ converges to y^*.*

If, moreover,

$$\sqrt{\lambda_k^{GB}} = \mathcal{O}\left(\|F(y_k)\|^p \right) \quad \text{as } k \to \infty,$$

then the Newton-GMBACK iterates converge with q-order at least $1 + p$.

Proof. The proof follows from Corollary 3.1, taking into account formula (6), the inequality $\|Z\|_2 \leq \|Z\|_F$, true for all $Z \in \mathbb{R}^{N \times N}$, and the fact that all norms are equivalent on a finite dimensional normed space. $\square$

We are interested now if the curve of the Newton-GMBACK errors is nonincreasing when starting with a sufficiently good approximation. In the following we shall construct an example which shows that, unlike the Newton-GMRES case, this property is not generally shared by the Newton-GMBACK method when GMBACK is used in the nonrestarted version.

Example 4.2. Consider the system $F(y) = 0$ with

$F : \mathbb{R}^N \to \mathbb{R}^N$, $F(y) = \left(y^{(1)} + y^{(N)}, y^{(1)}, y^{(2)}, \ldots, y^{(N-1)}\right)^t$, $y = \left(y^{(1)}, \ldots, y^{(N)}\right)^t \in \mathbb{R}^N$, having the unique solution $y^* = 0$. The derivative of F is given for all $y \in \mathbb{R}^N$ by

$$F'(y) \equiv [e_1 + e_2 \quad e_3 \; \ldots \; e_N \; e_1] =: A.$$

Taking $y_0 = -he_N$ with arbitrarily small $h \neq 0$, we must solve $F'(y_0)s = -F(y_0)$, or, equivalently, $Ax = b$ with $b = he_1$.

Applying $m \leq N - 1$ steps of the Arnoldi process with $x_0 = 0$, we successively obtain: $V_m = [e_1 \; \ldots \; e_m] \in \mathbb{R}^{N \times m}$, $\bar{H}_m = [e_1 + e_2 \quad e_3 \ldots e_{m+1}] \in \mathbb{R}^{(m+1) \times m}$, $P = \hat{\bar{H}}_m^t \hat{\bar{H}}_m = \left[h^2 e_1 - he_2 \quad 2e_2 - he_1 \quad e_3 \; \ldots \; e_{m+1}\right]$ and $Q = \hat{G}_m^t \hat{G}_m = [0 \; e_2 \; \ldots \; e_{m+1}] \in \mathbb{R}^{(m+1) \times (m+1)}$.

The eigenpair $\left(u_{m+1}, \lambda_{m+1}^{GB}\right)$ of $Pu = \lambda Qu$ is uniquely determined by $u_{m+1} = (1/h, 1, 0, \ldots, 0)^t$ and $\lambda_{m+1}^{GB} = 1$. It follows that $y_m^{GB} = he_1 \in \mathbb{R}^m$ and the GMBACK solution is $x_m^{GB} = he_1$, such that when $N \geq 3$

$$\|y_1 - y^*\|_2 = \sqrt{2}\,|h| > |h| = \|y_0 - y^*\|_2\,.$$

$\square$

4.3 Newton-MINPERT Method

The MINPERT method for solving (4) was introduced by Kasenally and Simoncini in [40]. Given an initial approximation $x_0 \in \mathbb{R}^N$, it computes an element $x_m^{MP} \in x_0 + \mathcal{K}_m$ which minimizes the joint backward error:

$$(7) \qquad \min_{x_m \in x_0 + \mathcal{K}_m} \|[\Delta_A \; \Delta_b]\|_F \quad \text{subject to} \quad (A - \Delta_A)x_m = b + \Delta_b.$$

In other words, x_m^{MP} minimizes the distance from the original system to the nearest one that an approximation $x_m \in x_0 + \mathcal{K}_m$ actually satisfies. The above minimization problem may also be viewed as a total least squares problem in the Krylov subspace (see [40]).

The algorithm is similar to GMBACK, the only difference being given by the computation of the matrices from the eigenproblem $Pu = \lambda Qu$. The following is a sketch of the steps performed by MINPERT:

Arnoldi

- Compute V_m and $\bar{H}_m$;

MINPERT

- Let $\beta = \|r_0\|_2$,

$$\hat{H}_m = \left[-\beta e_1 \quad \bar{H}_m\right] \in \mathbb{R}^{(m+1)\times(m+1)}, \quad \hat{G}_m = \begin{bmatrix} x_0 & V_m \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(m+1)},$$

$$P = \hat{H}_m^t \hat{H}_m \in \mathbb{R}^{(m+1)\times(m+1)} \quad \text{and} \quad Q = \hat{G}_m^t \hat{G}_m \in \mathbb{R}^{(m+1)\times(m+1)};$$

- Determine an eigenvector u_{m+1} corresponding to the smallest eigenvalue λ_{m+1}^{MP} of the generalized eigenproblem $Pu = \lambda Qu$;

- If the first component $u_{m+1}^{(1)}$ is nonzero, compute the vector $y_m^{MP} \in \mathbb{R}^m$ by scaling u_{m+1} such that

$$\begin{bmatrix} 1 \\ y_m^{MP} \end{bmatrix} = \frac{1}{u_{m+1}^{(1)}} u_{m+1};$$

- Set $x_m^{MP} = x_0 + V_m y_m^{MP}$.

The remarks concerning the breakdowns of the GMBACK method hold also for MIN-PERT.

Kasenally and Simoncini proved that for any $x_0 \in \mathbb{R}^N$ and $m \in \{1, \ldots, N\}$ the following relations hold:

$$\left\| \begin{bmatrix} \Delta_{A,m}^{MP} & \Delta_{b,m}^{MP} \end{bmatrix} \right\|_F = \sqrt{\lambda_{m+1}^{MP}}, \tag{8}$$

$$\left\| r_m^{MP} \right\|_2 = \sqrt{\lambda_{m+1}^{MP}} \left\| \begin{bmatrix} (x_m^{MP})^t & 1 \end{bmatrix}^t \right\|_2, \tag{9}$$

where $\Delta_{A,m}^{MP}, \Delta_{b,m}^{MP}$ resp. r_m^{MP} represent the backward errors and the residual corresponding to the MINPERT solution x_m^{MP}.

We shall prove for MINPERT a result somehow similar to Proposition 4.1:

Proposition 4.3. *Consider the linear system* (4), *the initial approximation* $x_0 = 0$ *and an arbitrary value* $m \in \{1, \ldots, N\}$. *If there exists a MINPERT solution* x_m^{MP}, *then its joint backward error satisfies*

$$\left\| \begin{bmatrix} \Delta_{A,m}^{MP} & \Delta_{b,m}^{MP} \end{bmatrix} \right\|_F \le \|b\|_2 \,.$$

Proof. When $x_0 = 0$, as noticed in [40], we are led to a regular eigenproblem in the MIN-PERT algorithm, since $Q = I_{m+1}$. The boundness of the Rayleigh quotient then implies

$$\lambda_{m+1}^{MP} = \min_{z \in \mathbb{R}^{m+1}} \frac{z^t P z}{z^t z} \le \frac{e_1^t P e_1}{e_1^t e_1} = e_1^t \hat{H}_m^t \hat{H}_m e_1 = \beta^2 = \|b\|_2^2 \,,$$

so, by (8), the conclusion is immediate. $\qquad \square$

The other two affirmations similar to those from Proposition 4.1 may be correspondingly stated with the single remark that, since the solution of the minimization problem (7) may not be uniquely determined, the inequality between two consecutive joint backward errors may not be strict even if the consecutive solutions x_m^{MP} and x_{m+1}^{MP} are distinct.

We also note that when the initial approximation in GMBACK is taken $x_0 = 0$, one cannot similarly use the vector e_1 in the Fisher result (see, e.g., [41, Cor. VI.1.16]—which says that $\lambda_{m+1}^{GB} = \min_{z \in \mathbb{R}^{m+1}} \frac{z^t P z}{z^t Q z}$) in order to bound λ_{m+1}^{GB} by β^2, since in this case $e_1^t Q e_1 = 0$, i.e. e_1 is an eigenvector corresponding to the infinite eigenvalue $\lambda_1 = +\infty$. Such a result could not be expected to hold in general, because by Corollary 3.1 it would imply that for an arbitrary mapping F satisfying assumptions **(C1)**–**(C3)** and with Lipschitz derivative, the

Newton-GMBACK iterations with $x_0 = 0$ in GMBACK and obeying **(C4)** converge locally with q-order 2 even for one-dimensional Krylov subspaces.

The convergence orders of the Newton-MINPERT iterates may be characterized by (9) in terms of the computed eigenvalues λ_k^{MP}.

Theorem 4.3. *Assume that* **(C1)**–**(C3)** *hold, that the derivative F' is Hölder continuous with exponent p at y^* and that the sequence $(y_k)_{k \geq 0}$ given by the Newton-MINPERT iterations $y_{k+1} = y_k + s_k^{MP}$ with*

$$F'(y_k) s_k^{MP} = -F(y_k) + r_k^{MP}, \quad k = 0, 1, \ldots, \ y_0 \in D,$$

converges to y^*. Then $(y_k)_{k \geq 0}$ converges with q-order at least $1 + p$ if and only if

$$\sqrt{\lambda_k^{MP}} = \mathcal{O}\left(\|F(y_k)\|^{1+p} \right) \quad \text{as } k \to \infty.$$

The direct application of Proposition 4.3 to the study of the monotonicity of the Newton-MINPERT errors leads to a result similar to Theorem 4.1, but which requires some additional conditions.

Theorem 4.4. *Assume that the mapping F satisfies conditions* **(C1)**–**(C3)** *and consider a current approximation $y_c \neq y^*$ for y^*. Let s^{MP} be a nonzero approximate solution of the linear system $F'(y_c) s = -F(y_c)$ provided by MINPERT with the initial guess $x_0 = 0$. Let $y_+ = y_c + s^{MP}$. If*

$$\eta = \sqrt{\lambda^{MP}} \left\| \left[\left(s^{MP} \right)^t \ 1 \right]^t \right\|_2 / \|F(y_c)\|_2 < 1,$$

and y_c is sufficiently close to y^*, then

$$\|y_+ - y^*\|_* \leq t \|y_c - y^*\|_*$$

for all $t \in (\eta, 1)$, where $\|y\|_* = \|F'(y^*) y\|_2$.

The following example shows that the Newton-MINPERT iterates may stagnate no matter how close to the solution y^* we choose y_0.

Example 4.3. Consider the mapping F from Example 4.1, which leads us again to the solving of the linear system $Ax = b$. The MINPERT method with $x_0 = 0$ and $m \in \{1, \ldots, N-1\}$ yields the diagonal matrix $P \in \mathbb{R}^{(m+1) \times (m+1)}$ with $\operatorname{diag}(P) = \left(h^2, 1, \ldots, 1 \right)$ and $Q = I_{m+1}$, such that for any h with $0 < |h| < 1$ we get $\lambda_{m+1}^{MP} = h^2$ and $u_{m+1} = e_1$. It follows the unique choice $y_m^{MP} = 0$ and so $s_{0,m}^{MP} = 0$. □

Remark 4.1. This example also shows that for the linear systems $Ax = b$ above, the MINPERT method behaves identically with GMRES, i.e. for $x_0 = 0$ and $\bar{m}, m \in \{1, \ldots, N-1\}$ it cannot improve the accuracy of the approximate solution, even in the restarted version. The systems with the same matrix A but with $b = h e_1$, $|h| > 1$, constitute some examples when the MINPERT method for a linear system leads to uncircumventible breakdowns. The stagnations and the uncircumventible breakdowns of MINPERT were theoretically known to be possible, but we have not encountered concrete examples before. □

We end the theoretical considerations by applying a result which relates the GMRES and MINPERT approximations to the Newton iterations. However, we do not tackle here this problem in its general setting nor do we analyze the conditions required in this result.

Kasenally and Simoncini proved that the difference between the GMRES and MINPERT solutions may be bounded under certain circumstances.

Theorem 4.5. [40] *Consider the same arbitrary elements $m \in \{1, \ldots, N\}$ and $x_0 \in \mathbb{R}^N$ in the GMRES and MINPERT methods for solving the linear system* (4). *Denote by σ_m^2 the smallest eigenvalue of $\bar{H}_m^t \bar{H}_m$ (i.e. the smallest squared singular value of $\bar{H}_m$) and assume that $\sigma_m^2 \neq \lambda_{m+1}^{MP}$. Then*

$$\left\| x_m^{MP} - x_m^{GM} \right\|_2 \leq \frac{\lambda_{m+1}^{MP}}{\sigma_m^2 - \lambda_{m+1}^{MP}} \left\| x_m^{GM} \right\|_2 .$$

An auxiliary result proved by these authors shows that the eigenvalues determined at the step m interlace in the following way: $\lambda_i^{MP} \geq \sigma_i^2 \geq \lambda_{i+1}^{MP}$ for $i = 1, \ldots, m$. The bound from the above theorem is large whenever σ_m^2 is close to λ_{m+1}^{MP} and the result cannot be applied when $\sigma_m^2 = \lambda_{m+1}^{MP}$ (such a situation arises for example when λ_{m+1}^{MP} is a repeated eigenvalue).

Theorem 4.8 may determine theoretical possibilities when GMRES and MINPERT have the same asymptotical behavior when used in a converging Newton method.

Theorem 4.6. *Assume that Conditions* **(C1)**–**(C3)** *hold and that the sequence $\left(y_k^{GM}\right)_{k \geq 0}$ given by the Newton-GMRES method converges to y^*, where the elements from the GMRES algorithm are arbitrary at each outer step. Denote by s_k^{GM} the obtained corrections from the linear systems*

$$F'\left(y_k^{GM}\right) s = -F\left(y_k^{GM}\right),$$

and consider the approximate solutions s_k^{MP} of these linear systems obtained with MIN-PERT, which is assumed to use the same initial approximations, the same subspace dimensions and the same number of restarts as GMRES. If $\sigma_k^2 \neq \lambda_k^{MP}$ for $k = 0, 1, \ldots$ and $\lambda_k^{MP} = o\left(\sigma_k^2\right)$ as $k \to \infty$ then GMRES and MINPERT yield asimptotically the same normalized corrections:

$$\frac{1}{\left\|s_k^{GM}\right\|_2} s_k^{GM} - \frac{1}{\left\|s_k^{MP}\right\|_2} s_k^{MP} \to 0 \quad as \ k \to \infty.$$

The same result holds inverting the role of GMRES and MINPERT.

Proof. The hypotheses of the theorem imply that $\lambda_k^{MP} \to 0$ as $k \to \infty$. For the nonrestarted version of the Krylov solvers the affirmation is a straightforward application of Theorem 4.5, while for the restarted version the conclusion is obtained by an inductive argument on the number of restarts from each outer step. □

5 Numerical Examples

We shall consider two test problems from [34] and [42], that will provide some nonlinear systems in $\mathbb{R}^N$. We shall apply to them the studied Newton–Krylov methods. The Krylov solvers are considered in the nonrestarted version and the initial guess is taken 0 both in

the inner and in the outer iterations. We are interested in the behavior of the magnitude of the backward errors of these methods and in verifying theoretical results and therefore we have not aimed to implement some efficient methods for solving the problems under consideration.

5.1 Bratu problem

Consider the nonlinear partial differential equation

$$ - \triangle u + \alpha u_x + \lambda e^u = f, $$

over the unit square of $\mathbb{R}^2$, with Dirichlet boundary conditions. As mentioned in [34] and [42], this is a standard problem, a simplified form of which is known as the Bratu problem (See [43]). We have discretized by 5-point finite differences, respectively by central finite differences on a uniform mesh, obtaining a system of nonlinear equations of size $N = (n-2)^2$, where n is the number of mesh points in each direction. As in [34], we took f such that the solution of the discretized problem to be the constant unity. We have considered $\alpha = 10$, $\lambda = 1$, $N = 1024$ and $m = 10$. The runs were made on a PC, using MATLAB, version 4.0. The symbol $\|\cdot\|$ denotes either the Euclidean or the Frobenius norm, and e_k, F_k stand for $y^* - y_k$, resp. $F(y_k)$.

Table 1 contains the results obtained by using the Newton-GMRES method. We have also considered the corrections obtained with MINPERT at each step, denoting by a_k the norm of the difference $(1/\|s_k^{GM}\|)s_k^{GM} - (1/\|s_k^{MP}\|)s_k^{MP}$ (the corrections s_k^{MP} were computed only for the comparison).

k	$\|e_k\|$	$\|F_k\|^2$	$\|r_k\|$	a_k	k	$\|e_k\|$	$\|F_k\|^2$	$\|r_k\|$	a_k
0	3e+1	1e+2	9e-1	4e-2	11	3e-4	1e-10	8e-06	1e-9
1	2e+1	9e-1	5e-1	3e-2	12	2e-4	7e-11	2e-06	4e-11
2	1e+1	3e-1	4e-1	2e-2	13	3e-5	8e-12	7e-07	4e-12
3	1e+1	1e-1	2e-1	9e-3	14	1e-5	5e-13	4e-07	3e-12
4	6e+0	8e-2	3e-2	9e-5	15	1e-5	2e-13	2e-07	3e-13
5	1e-1	1e-3	6e-3	2e-4	16	1e-6	4e-14	2e-08	1e-14
6	5e-2	4e-5	2e-3	5e-5	17	5e-7	6e-16	1e-08	5e-14
7	2e-2	5e-6	6e-4	2e-6	18	4e-7	1e-16	6e-09	3e-15
8	1e-2	4e-7	3e-4	2e-6	19	8e-8	4e-17	5e-10	2e-15
9	1e-2	1e-7	1e-4	2e-7	20	9e-9	2e-19	2e-10	4e-14
10	1e-3	2e-8	1e-5	4e-10	21	6e-9	5e-20	9e-11	2e-16

Table 1: Newton-GMRES applied to the Bratu problem.

It can be seen that the normalized corrections agree the closer the iterations get to the solution.

Table 2 contains the results obtained by using the Newton-GMBACK method.

We notice the rather constant magnitudes of $\Delta_{A_k}^{GB}$, which do not approach zero even when the iterates are close to the solution.

In Table 3 we have denoted $b_k = \left\| s_{k,10}^{GM} - s_{k,10}^{MP} \right\|_2 / \left\| s_{k,10}^{GM} \right\|_2$ and $c_k = \lambda_{k,10}^{MP}/(\sigma_{k,10}^2 - \lambda_{k,10}^{MP})$.

k	$\|e_k\|$	$\|F_k\|$	$\|\Delta_{A_k}^{GB}\|$	$\|r_k\|$	k	$\|e_k\|$	$\|F_k\|$	$\|\Delta_{A_k}^{GB}\|$	$\|r_k\|$
0	3e+1	1e+1	6e-2	1e+0	11	1e-4	2e-5	4e-2	4e-6
1	2e+1	1e+0	5e-2	7e-1	12	8e-5	4e-6	5e-2	3e-6
2	1e+1	7e-1	6e-2	5e-1	13	3e-5	3e-6	1e-2	5e-7
3	9e+0	5e-1	6e-2	3e-1	14	8e-6	5e-7	6e-2	3e-7
4	4e+0	3e-1	2e-2	7e-2	15	4e-6	3e-7	2e-2	8e-8
5	1e+0	7e-2	4e-2	6e-2	16	1e-6	8e-8	5e-2	5e-8
6	5e-1	6e-2	2e-2	1e-2	17	5e-7	5e-8	2e-2	1e-8
7	2e-1	1e-2	1e-2	4e-3	18	2e-7	1e-8	3e-2	6e-9
8	1e-2	4e-3	5e-2	5e-4	19	3e-8	6e-9	2e-2	7e-10
9	4e-3	5e-4	4e-2	9e-5	20	1e-8	7e-10	2e-2	3e-10
10	2e-3	9e-5	1e-2	2e-5	21	1e-9	3e-10	1e-2	1e-11

Table 2: Newton-GMBACK applied to the Bratu problem.

k	$\|e_k\|$	$\|F_k\|$	$\sqrt{\lambda_{k,10}^{MP}}$	$\|F_k\|^2$	$\|r_k\|$	b_k	c_k
0	3e+1	1e+1	6e-2	1e+2	1e+0	8.3594e-2	1.0364e-1
1	2e+1	1e+0	5e-2	1e+0	7e-1	6.6518e-1	6.7180e-1
2	1e+1	7e-1	6e-2	5e-1	5e-1	7.3033e-1	7.3995e-1
3	9e+0	5e-1	6e-2	2e-1	3e-1	6.1259e-1	6.2053e-1
4	4e+0	3e-1	2e-2	1e-1	7e-2	4.4297e-2	4.4403e-2
5	1e+0	7e-2	3e-2	5e-3	5e-2	3.3261e-1	3.3456e-1
6	8e-1	5e-2	1e-2	2e-3	1e-2	3.0693e-2	3.0808e-2
7	3e-1	1e-2	5e-3	2e-4	5e-3	1.5905e-2	1.6004e-2
8	6e-2	5e-3	1e-3	3e-5	1e-3	2.0734e-4	2.0883e-4
9	2e-2	1e-3	7e-4	2e-6	7e-4	1.1074e-4	1.1233e-4
10	1e-2	7e-4	1e-4	5e-7	1e-4	1.0523e-5	1.0526e-5

Table 3: Newton-MINPERT applied to the Bratu problem.

The bounds $\lambda_{k,10}^{MP}/\left(\sigma_{k,10}^2 - \lambda_{k,10}^{MP}\right)$ seem to be tight here for $\left\|s_{k,10}^{GM} - s_{k,10}^{MP}\right\|_2 / \left\|s_{k,10}^{GM}\right\|_2$.
As soon as the iterations approach the solution, the inequalities from Theorem 4.5 cease
to hold numerically; we believe that this is due to the different types of errors which have
appeared in computing the eigenpairs, the Krylov approximations and the elements b_k, c_k.

5.2 Driven cavity flow problem

This is a classical problem from incompressible fluid flow. It has the following equations in
stream function-vorticity formulation:

18

k	$\|e_k\|$	$\|F_k\|$	$\sqrt{\lambda_{k,10}^{MP}}$	$\|F_k\|^2$	$\|r_k\|$	b_k	c_k
11	2e-3	1e-4	7e-05	3e-08	7e-05	5.2230e-7	5.3930e-7
12	1e-3	7e-5	6e-06	5e-09	6e-06	2.6605e-8	2.6605e-8
13	4e-5	6e-6	1e-06	4e-11	1e-06	3.1452e-11	3.2821e-11
14	2e-5	1e-6	6e-07	1e-12	6e-07	1.6474e-10	1.6433e-10
15	9e-6	6e-7	1e-07	3e-13	1e-07	4.9880e-12	5.0888e-12
16	2e-6	1e-7	8e-08	2e-14	8e-08	1.6094e-12	1.6504e-12
17	1e-6	8e-8	1e-08	6e-15	1e-08	1.6952e-13	8.3076e-14
18	1e-7	1e-8	4e-09	1e-16	4e-09	1.2156e-14	1.7097e-15
19	1e-7	4e-9	1e-09	2e-17	1e-09	4.7003e-13	9.6014e-16
20	5e-9	1e-9	2e-11	1e-18	2e-11	2.2673e-14	2.2768e-20

Table 3: (continued)

$$\nu \triangle \omega + (\psi_y \omega_x - \psi_x \omega_y) = 0 \quad \text{in } \Omega, \tag{10}$$

$$-\triangle \psi = \omega \quad \text{in } \Omega, \tag{11}$$

$$\psi = 0 \quad \text{on } \partial\Omega,$$

$$\left. \frac{\partial \psi}{\partial n}(x,y) \right|_{\partial\Omega} = \begin{cases} 1, & \text{if } y = 1 \\ 0, & \text{if } 0 \leq y < 1, \end{cases}$$

where Ω denotes again the interior of the unit square from $\mathbb{R}^2$, and the viscosity ν is the reciprocal of the Reynolds number Re. In terms of ψ alone, (10) and (11) are replaced by

$$\nu \triangle^2 \psi + \left(\psi_y \left(\triangle\psi \right)_x - \psi_x \left(\triangle\psi \right)_y \right) = 0 \quad \text{in } \Omega.$$

This equation was discretized again on a uniform mesh. We have considered $n = 25$, choosing the boundary conditions to be incorporated in the nonlinear system. We have used the symbolic facilities offered by Maple V Release 3.0 in order to determine the discretized equations. The Reynolds number was taken small ($Re = 10$) and the subspace dimensions of the Krylov methods were considered large ($m = 50$), since the nonrestarted versions of these methods were not efficient for this problem. Tables 4 and 5 contain the results of the three Newton–Krylov methods applied to this problem.

We notice again that, close to the solution, the computed elements b_k and c_k do not satisfy the inequalities $b_k \leq c_k$.

In Figure 1 we have plotted on a semilog scale the evolution of $\|F_k\|$ for the three methods. There were considered the first 150 steps. It can be seen the monotone behavior of Newton-GMRES and Newton-MINPERT starting from certain steps, and the non-monotone behavior of Newton-GMBACK.

k	$\|F_k\|$	$\|F_k\|^2$	$\|r_k\|$	a_k
0	4e+0	2e+1	9e-1	2e-2
1	8e+0	7e+1	1e+0	7e-2
2	4e+0	2e+1	8e-1	4e-2
3	1e+0	3e+0	6e-1	4e-2
4	7e-1	5e-1	5e-1	1e-2
...				
131	9e-5	8e-9	8e-5	6e-9
132	8e-5	7e-9	8e-5	7e-9
133	8e-5	6e-9	7e-5	5e-9
134	7e-5	6e-9	7e-5	5e-9
135	7e-5	5e-9	6e-5	4e-9

k	$\|F_k\|$	$\left\|\Delta^{GB}_{A_k}\right\|$	$\|F_k\|^2$	$\|r_k\|$
0	4e+0	6e-2	2e+1	9e-1
1	9e+0	3e-1	9e+1	1e+0
2	6e+0	3e-1	3e+1	1e+0
3	1e+0	3e-1	3e+0	9e-1
4	9e-1	2e-1	8e-1	8e-1
...				
131	4e-5	6e-1	1e-09	2e-5
132	2e-5	4e-1	5e-10	3e-5
133	3e-5	6e-1	1e-09	2e-5
134	2e-5	4e-1	4e-10	3e-5
135	3e-5	6e-1	1e-09	1e-5

Table 4: Newton-GMRES, resp. Newton-GMBACK applied to the driven cavity problem.

k	$\|F_k\|$	$\sqrt{\lambda^{MP}_{k,50}}$	$\|F_k\|^2$	$\|r_k\|$	b_k	c_k
0	4e+0	6e-2	2e+1	9e-1	7.0262e-2	7.8592e-2
1	9e+0	3e-1	9e+1	1e+0	3.3821e-1	3.5336e-1
2	6e+0	3e-1	3e+1	1e+0	8.0633e-1	8.2988e-1
3	1e+0	2e-1	3e+0	8e-1	1.1452e+0	1.1887e+0
4	8e-1	2e-1	7e-1	7e-1	1.6037e+0	1.6438e+0
...						
131	1e-5	1e-5	2e-10	1e-5	4.0286e-8	4.0292e-8
132	1e-5	1e-5	2e-10	1e-5	3.2414e-9	3.2489e-9
133	1e-5	1e-5	2e-10	1e-5	3.6246e-8	3.6241e-8
134	1e-5	1e-5	1e-10	1e-5	2.4452e-9	2.4576e-9
135	1e-5	1e-5	1e-10	1e-5	2.1203e-8	2.1201e-8

Table 5: Newton-MINPERT applied to the driven cavity problem.

6 Conclusions

The local superlinear convergence and the local convergence with orders $1 + p$, $p \in (0, 1]$, of the Newton methods were characterized in a more natural setting, which assumes that the Jacobians are perturbed, the function evaluations are performed approximately for F and the resulted linear systems are solved inexactly. The results of Dembo, Eisenstat and Steihaug allowed us to extend their local convergence analysis to this setting.

The sufficient conditions for local convergence with different convergence orders show that the perturbations in the Jacobians may be allowed to have much greater magnitudes than those in the function evaluations and than the magnitudes of the residuals (this may be explained by the fact that the right-hand sides $-F(x_k)$ of the linear systems from the Newton method tend to zero, while the matrices $F'(x_k)$ tend to $F'(x^*)$, assumed here to be nonsingular). It follows that the methods for solving the linear systems from the Newton
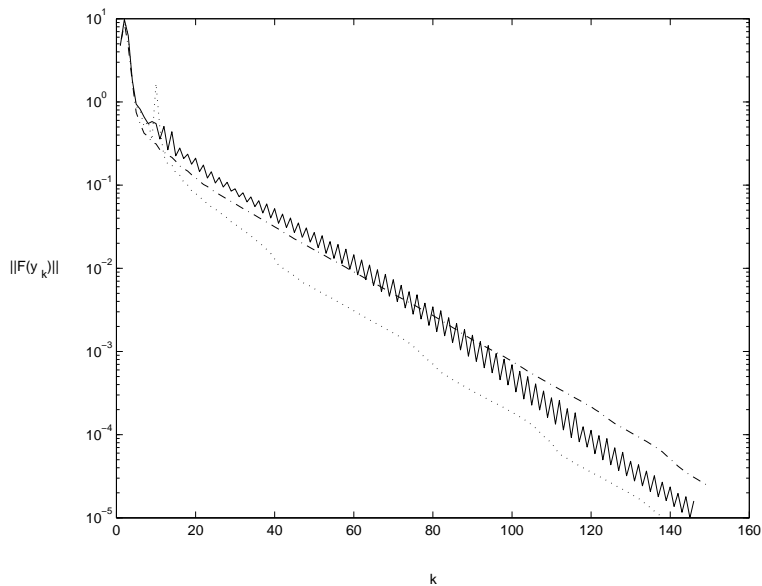
Figure 1: Newton-GMRES, Newton-GMBACK and Newton-MINPERT

methods must be analyzed in this respect and also that special care must be taken when the function evaluations may be affected by significant errors. The greater sensitivity of the right-hand sides $-F(x_k)$ is known for a longer time (see for example [11], [21] and the references therein) such that the evaluation of these vectors in double or extended precision is already a standard.

The existing results for the IN iterates allowed us to show that the Newton-GMRES iterates have a monotone property. We were able to prove that Newton-MINPERT shares this property only under some additional conditions, whereas from a concrete example we saw that Newton-GMBACK does not generally share this property. The convergence orders of Newton-GMBACK and Newton-MINPERT methods can be controlled in terms of the magnitude of the backward errors of the approximate steps. The theoretical results were confronted with the performed numerical examples.

Many of the existing results for different Newton-type methods may be reconsidered in the IPN setting. For example, the local convergence orders of some finite difference Newton–Krylov methods studied by Brown [26] now may be analyzed in the IPN setting. On the other hand, new approaches can be developed as well; for instance we should mention the local convergence and acceleration techniques for the IPN methods in the case of singular Jacobian at the solution. Other results and directions of research are included in our Ph.D. Thesis [44].

paper and helpful criticism and suggestions.

References

[1] R. S. DEMBO, S. C. EISENSTAT and T. STEIHAUG, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[2] J. M. ORTEGA and W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[3] L. V. KANTOROVICH and G. P. AKILOV, *Functional Analysis*, Editura Ştiinţifică şi Enciclopedică, Bucharest, Romania, 1986 (in Romanian).

[4] I. PĂVĂLOIU, *Introduction to the Theory of Approximating the Solutions of Equations*, Editura Dacia, Cluj–Napoca, Romania, 1976 (in Romanian).

[5] ŞT. MĂRUŞER, *Numerical Methods for Solving Nonlinear Equations*, Editura Tehnică, Bucharest, Romania, 1981 (in Romanian).

[6] J. E. DENNIS, JR., and R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall Series in Computational Mathematics, Englewood Cliffs, New Jersey, 1983.

[7] F. A. POTRA and V. PTÁK, *Nondiscrete Induction and Iterative Processes*, Pitman, London, 1984.

[8] I. ARGYROS and F. SZIDAROVSZKY, *The Theory and Applications of Iteration Methods*, CRC Press, Boca Raton, 1993.

[9] C. T. KELLEY, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.

[10] W. C. RHEINBOLDT, *Methods for Solving Systems of Nonlinear Equations*, SIAM, Philadelphia, 1998.

[11] J. E. DENNIS, JR., and H. F. WALKER, *Inaccuracy in quasi-Newton methods: local improvement theorems*, Math. Prog. Study, 22 (1984), pp. 70–85.

[12] F. A. POTRA and V. PTÁK, *Sharp error bounds for Newton's process*, Numer. Math., 34 (1980), pp. 63–72.

[13] P. DEUFLHARD and F. A. POTRA, *Asymptotic mesh independence of Newton–Galerkin methods via a refined Mysovskii theorem*, SIAM J. Numer. Anal., 29 (1992), pp. 1395–1412.

[14] I. ARGYROS, *Concerning the convergence of inexact Newton methods*, J. Comp. Appl. Math., 79 (1997), pp. 235–247.

[15] I. ARGYROS, *On a new Newton–Mysovskii–type theorem with applications to inexact Newton-like methods and their discretizations*, IMA J. Numer. Anal., 18 (1998), pp. 37–56.

[16] A. H. SHERMAN, *On Newton-iterative methods for the solution of systems of nonlinear equations*, SIAM J. Numer. Anal., 15 (1978), pp. 755–771.

[17] J. E. DENNIS, JR., and J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[18] J. E. DENNIS, JR., and J. J. MORÉ, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.

[19] S. C. EISENSTAT and H. F. WALKER, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[20] S. C. EISENSTAT and H. F. WALKER, *Globally convergent inexact Newton methods*, SIAM J. Optim., 4 (1994), pp. 393–422.

[21] YPMA, T. J., *The effect of rounding errors on Newton-like methods*, IMA J. Numer. Anal., 3 (1983), pp. 109–118.

[22] H. J. MARTINEZ, Z. PARADA and R. A. TAPIA, *On the characterization of q-superlinear convergence of quasi-Newton interior-point methods for nonlinear programming*, Bol. Soc. Mat. Mexicana (3), 1 (1995), pp. 137–148.

[23] T. J. YPMA, *Local convergence of inexact Newton methods*, SIAM J. Numer. Anal., 21 (1983), pp. 583–590.

[24] D. CORES and R. A. TAPIA, *Perturbation lemma for the Newton method with application to the SQP Newton method*, J. Optim. Theory Appl., 97 (1998), pp.271–280.

[25] F. A. POTRA, *On q-order and r-order of convergence*, J. Optim. Theory Appl., 63 (1989), pp. 415–431.

[26] P. N. BROWN, *A local convergence theory for combined inexact-Newton/finite-difference projection methods*, SIAM J. Numer. Anal., 24 (1987), pp. 407–434.

[27] E. CĂTINAŞ, *A note on inexact secant methods*, Rev. Anal. Numér. Théor. Approx., 25 (1996), pp. 33–41.

[28] KRÓLIKOWSKA, M., *Nonlinear mappings with an almost sparse Jacobian matrix*, Annal. Univ. Mariae Curie-Skłodowska, Lublin-Polonia, 49 (A) (1995), pp. 145–157.

[29] J. L. RIGAL and J. GACHES, *On the compatibility of a given solution with the data of a linear system*, J. ACM, 14 (1967), pp. 543–548.

[30] Y. SAAD and M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[31] Y. SAAD, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[32] E. M. KASENALLY, *GMBACK: a generalised minimum backward error algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comp., 16 (1995), pp. 698–719.

23

[33] Y. Saad, *Iterative methods for sparse linear systems*, PWS Publishing Company, Boston, 1996.

[34] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 450–481.

[35] K. Turner and H. F. Walker, *Efficient high accuracy solutions with GMRES(m)*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 815–825.

[36] P. N. Brown and A. C. Hindmarsh, *Reduced storage matrix methods in stiff ODE systems*, Appl. Math. Comp., 31 (1989), pp. 40–91.

[37] A. Greenbaum, V. Pták and Z. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.

[38] P. N. Brown and Y. Saad, *Convergence theory of nonlinear Newton–Krylov algorithms*, SIAM J. Optim., 4 (1994), pp. 297–330.

[39] P. N. Brown, *A Theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 58–78.

[40] E. M. Kasenally and V. Simoncini, *Analysis of a minimum perturbation algorithm for nonsymmetric linear systems*, SIAM J. Numer. Anal., 34 (1997), pp. 48–66.

[41] G. W. Stewart and J. G. Sun, *Matrix perturbation theory*, Academic Press, New York, 1990.

[42] J. J. Moré, *A collection of nonlinear model problems,* Computational Solutions of Nonlinear Systems of Equations, Lectures in Applied Mathematics, Edited by E. L. Allgower and K. Georg, American Mathematical Society, Providence, Rhode Island, vol. 26, pp. 723–762, 1990.

[43] R. Glowinsky, H. B. Keller and L. Reinhart, *Continuation-conjugate gradient methods for the least-squares solution of nonlinear boundary value problems*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 793–832.

[44] E. Cătinaş, *Newton and Newton–Krylov methods for solving nonlinear systems in $\mathbb{R}^n$*, PhD Thesis, Babeş–Bolyai University, Cluj–Napoca, Romania, 1999.