Convergence of inexact Newton methods in the number of linear iterations

R. Idema, D.J.P Lahaye, and C. Vuik [*]

1 Introduction

Inexact Newton methods are great tools for the solution of large systems of non-linear equations $F(\boldsymbol{x}) = \mathbf{0}$. In each Newton iteration a Newton step $\boldsymbol{s}_i$ is obtained by solving the Jacobian system $J(\boldsymbol{x}_i)\,\boldsymbol{s}_i = -F(\boldsymbol{x}_i)$ with an iterative linear solver, up to an accuracy

$$\frac{\|\boldsymbol{r}_i\|}{\|F(\boldsymbol{x}_i)\|} \leq \eta_i, \tag{1}$$

where $\boldsymbol{r}_i = F(\boldsymbol{x}_i) + J(\boldsymbol{x}_i)\,\boldsymbol{s}_i$ is the residual error. The values η_i are called the forcing terms.

Dembo et al. [1] proved that—when using proper forcing terms—inexact Newton methods converge, and under more strict conditions on the forcing terms converge superlinearly.

Over the years a great deal of research has gone into finding good values for η_i, such that convergence is reached with the least amount of computational work. If η_i is too big many Newton iterations are needed, or convergence may even be lost, but if η_i is too small there will be a certain amount of oversolving in the linear solver. One of the most frequently used methods to calculate the forcing terms is that of Eisenstat and Walker [2].

Researching the use of inexact Newton-GMRES methods for large power flow problems, we found that a very good preconditioner—though costly to make—yielded the best solution times. Generally preconditioned GMRES will converge superlinearly; however, with a very good preconditioner GMRES was converging very fast but linearly, due to the strong clustering of the eigenvalues.

In this paper we will show that when GMRES converges linearly, the Newton–GMRES process—while converging superlinearly in the Newton iterations—convergences approximately linearly when its convergence is measured in the total number of GMRES iterations used throughout all Newton iterations. As such, the Newton convergence is independent of the forcing terms, provided that they are small enough to allow convergence, and large enough to counter oversolving.

2 Inexact Newton Convergence

We start with a variation on the inexact Newton convergence proof presented in [1, Thm. 2.3]. Consider the system of non-linear equations $F(\boldsymbol{x}) = \mathbf{0}$, where

- There is a solution $\boldsymbol{x}^*$ with $F(\boldsymbol{x}^*) = \mathbf{0}$.
- The Jacobian J of F exists in a neighborhood of $\boldsymbol{x}^*$, and J is continuous in $\boldsymbol{x}^*$.
- $J(\boldsymbol{x}^*)$ is non-singular.

[*] R. Idema (r.idema@tudelft.nl), D.J.P. Lahaye, and C. Vuik are with the Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands.

When referring to a Newton method we will silently assume that the above requirements are met for the problem that is being solved.

Theorem 2.1. *Let $\eta_i \in (0,1)$ and choose $\alpha > 0$ such that $(1+\alpha)\eta_i < 1$. Then there exists an $\varepsilon > 0$ such that, if $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| < \varepsilon$, the sequence of inexact Newton iterates $\boldsymbol{x}_i$ converges to $\boldsymbol{x}^*$, with*

$$\|J(\boldsymbol{x}^*)(\boldsymbol{x}_{i+1} - \boldsymbol{x}^*)\| < (1+\alpha)\eta_i \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|. \tag{2}$$

Proof. Define

$$\mu = \max\left[\|J(\boldsymbol{x}^*)\|, \|J(\boldsymbol{x}^*)^{-1}\|\right]. \tag{3}$$

Since $J(\boldsymbol{x}^*)$ is non-singular we then have

$$\frac{1}{\mu}\|\boldsymbol{y}\| \le \|J(\boldsymbol{x}^*)\boldsymbol{y}\| \le \mu\|\boldsymbol{y}\|. \tag{4}$$

Let

$$\gamma \in \left(0, \frac{\alpha\eta_i}{7\mu}\right) \tag{5}$$

and choose $\varepsilon > 0$ sufficiently small such that if $\|\boldsymbol{y} - \boldsymbol{x}^*\| \le \mu^2\varepsilon$ then

$$\|J(\boldsymbol{y}) - J(\boldsymbol{x}^*)\| \le \gamma, \tag{6}$$

$$\|J(\boldsymbol{y})^{-1} - J(\boldsymbol{x}^*)^{-1}\| \le \gamma, \tag{7}$$

$$\|F(\boldsymbol{y}) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{y} - \boldsymbol{x}^*)\| \le \gamma\|\boldsymbol{y} - \boldsymbol{x}^*\|. \tag{8}$$

That such an ε exists follows from [3, Thm. 2.3.3 & 3.1.5].

We prove Theorem 2.1 by induction. Note that equations (6)–(8) hold for $\boldsymbol{y} = \boldsymbol{x}_i$ because

$$\|\boldsymbol{x}_i - \boldsymbol{x}^*\| \le \mu\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| < \mu\|J(\boldsymbol{x}^*)(\boldsymbol{x}_0 - \boldsymbol{x}^*)\| \le \mu^2\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| < \mu^2\varepsilon, \tag{9}$$

where the second inequality is due to the induction hypothesis.

We can write

$$\begin{aligned} J(\boldsymbol{x}^*)(\boldsymbol{x}_{i+1} - \boldsymbol{x}^*) = &\left[I + J(\boldsymbol{x}^*)\left(J(\boldsymbol{x}_i)^{-1} - J(\boldsymbol{x}^*)^{-1}\right)\right] \\ &\cdot \left[\boldsymbol{r}_i + (J(\boldsymbol{x}_i) - J(\boldsymbol{x}^*))(\boldsymbol{x}_i - \boldsymbol{x}^*) - (F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*))\right], \end{aligned} \tag{10}$$

and taking norms we find

$$\begin{aligned} \|J(\boldsymbol{x}^*)(\boldsymbol{x}_{i+1} - \boldsymbol{x}^*)\| \le &\left[1 + \|J(\boldsymbol{x}^*)\|\|J(\boldsymbol{x}_i)^{-1} - J(\boldsymbol{x}^*)^{-1}\|\right] \\ &\cdot \left[\|\boldsymbol{r}_i\| + \|J(\boldsymbol{x}_i) - J(\boldsymbol{x}^*)\|\|\boldsymbol{x}_i - \boldsymbol{x}^*\| + \|F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|\right], \\ \le &\left[1 + \mu\gamma\right] \cdot \left[\eta_i\|F(\boldsymbol{x}_i)\| + \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\| + \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\|\right], \end{aligned} \tag{11}$$

where we used the definitions of μ and η_i, as well as equations (6)–(8).

We can also write

$$F(\boldsymbol{x}_i) = [J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)] + [F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)], \tag{12}$$

and again taking norms find

$$\begin{aligned} \|F(\boldsymbol{x}_i)\| &\le \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| + \|F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| \\ &\le \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| + \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\|. \end{aligned} \tag{13}$$

Combining equations (11) and (13), we then find that

$$
\begin{aligned}
&\|J\left(\boldsymbol{x}^*\right)\left(\boldsymbol{x}_{i+1}-\boldsymbol{x}^*\right)\| && (14)\\
&\leq \left(1+\mu\gamma\right)\left[\eta_i\left(\|J\left(\boldsymbol{x}^*\right)\left(\boldsymbol{x}_i-\boldsymbol{x}^*\right)\|+\gamma\|\boldsymbol{x}_i-\boldsymbol{x}^*\|\right)+2\gamma\|\boldsymbol{x}_i-\boldsymbol{x}^*\|\right] && (15)\\
&\leq \left(1+\mu\gamma\right)\left[\eta_i\left(\|J\left(\boldsymbol{x}^*\right)\left(\boldsymbol{x}_i-\boldsymbol{x}^*\right)\|+\mu\gamma\|J\left(\boldsymbol{x}^*\right)\left(\boldsymbol{x}_i-\boldsymbol{x}^*\right)\|\right)+2\mu\gamma\|J\left(\boldsymbol{x}^*\right)\left(\boldsymbol{x}_i-\boldsymbol{x}^*\right)\|\right] && (16)\\
&= \left(1+\mu\gamma\right)\left[\eta_i\left(1+\mu\gamma\right)+2\mu\gamma\right]\|J\left(\boldsymbol{x}^*\right)\left(\boldsymbol{x}_i-\boldsymbol{x}^*\right)\|. && (17)
\end{aligned}
$$

Now, using that $\gamma < \frac{\alpha\eta_i}{7\mu}$ and that both $\eta_i < 1$ and $\alpha\eta_i < 1$—the latter being a result from the requirement that $\left(1+\alpha\right)\eta_i < 1$—we can write

$$
\begin{aligned}
\left(1+\mu\gamma\right)\left[\eta_i\left(1+\mu\gamma\right)+2\mu\gamma\right] &\leq \left(1+\frac{\alpha\eta_i}{7}\right)\left[\eta_i\left(1+\frac{\alpha\eta_i}{7}\right)+\frac{2\alpha\eta_i}{7}\right] && (18)\\
&= \left[\left(1+\frac{\alpha\eta_i}{7}\right)^2+\left(1+\frac{\alpha\eta_i}{7}\right)\frac{2\alpha}{7}\right]\eta_i && (19)\\
&= \left[1+\frac{2\alpha\eta_i}{7}+\frac{\alpha^2\eta_i^2}{49}+\frac{2\alpha}{7}+\frac{2\alpha^2\eta_i}{49}\right]\eta_i && (20)\\
&< \left[1+\frac{2\alpha}{7}+\frac{\alpha}{49}+\frac{2\alpha}{7}+\frac{2\alpha}{49}\right]\eta_i && (21)\\
&< \left(1+\alpha\right)\eta_i. && (22)
\end{aligned}
$$

□

Intuitively there seems something wrong with Theorem 2.1. If we can choose η_i and α arbitrarily small, due to equation (2) we can reach any desired Newton convergence in a single step. However, this overlooks the fact that ε depends on the choices of η_i and α. The theorem states that for every choice of η_i and α, the relative convergence $\left(1+\alpha\right)\eta_i$ is attained when $\boldsymbol{x}_i$ is close enough to the solution.

Another way of looking at this is that a given iterate $\boldsymbol{x}_i$—close enough to the solution to guarantee convergence—imposes the restriction that for Theorem 2.1 to hold the forcing terms η_i cannot be chosen too small. This is nothing new, as we already noted that too small η_i lead to oversolving. Thus we can consider $\eta_i > \frac{7\mu\gamma}{\alpha}$ an upper bound for the forcing terms that wards against oversolving.

If we define oversolving as using forcing terms η_i that are too small for a certain iterate $\boldsymbol{x}_i$ in the context of Theorem 2.1, then we can characterize the theorem by saying that a convergence factor $\left(1+\alpha\right)\eta_i$ is attained if η_i is chosen such that there is convergence, but no oversolving.

Corollary 2.1. *Let $\eta_i \in (0,1)$ and choose $\alpha > 0$ such that $\left(1+\alpha\right)\eta_i < 1$. Then there exists an $\varepsilon > 0$ such that, if $\|\boldsymbol{x}_0-\boldsymbol{x}^*\| < \varepsilon$, the sequence of inexact Newton iterates $\boldsymbol{x}_i$ converges to $\boldsymbol{x}^*$, with*

$$
\|J\left(\boldsymbol{x}^*\right)\left(\boldsymbol{x}_i-\boldsymbol{x}^*\right)\| < \left(1+\alpha\right)^i\eta_{i-1}\cdots\eta_0\|J\left(\boldsymbol{x}^*\right)\left(\boldsymbol{x}_0-\boldsymbol{x}^*\right)\|. \tag{23}
$$

Proof. The stated follows readily from the repeated application of Theorem 2.1. □

3 Inexact Newton-GMRES convergence

In this section we use the results of the previous section to analyze theoretical inexact Newton convergence. We use GMRES as a practical example for the linear solver, but the results generally hold for any iterative linear solver used within an inexact Newton process.

We assume that η_i and α are such that Theorem 2.1 holds. Let M be the number of Newton iterations, N_i the number of GMRES iterations used in Newton iteration i, and $\Sigma N_i = \sum_{k=0}^{i} N_i$ the total number of GMRES iterations throughout all Newton iterations.

Obviously the convergence of the Newton-GMRES method will depend on the convergence of the GMRES processes. Therefore we introduce the following definitions.

Defintion 3.1. *Let $A\boldsymbol{x} = \boldsymbol{b}$ be a system of linear equations to be solved by an iterative linear solver, and let $\boldsymbol{r}_j$ be the residual error in iteration j. We say that the iterative method converges linearly with rate ρ, if for all j*

$$\|\boldsymbol{r}_{j+1}\| \leq 10^{-\rho}\|\boldsymbol{r}_j\|. \tag{24}$$

We say that the iterative method converges superlinearly if for all j

$$\frac{\|\boldsymbol{r}_{j+1}\|}{\|\boldsymbol{r}_j\|} < \frac{\|\boldsymbol{r}_j\|}{\|\boldsymbol{r}_{j-1}\|}. \tag{25}$$

We will explore the Newton-GMRES convergence under Theorem 2.1, for both superlinear and linear GMRES convergence.

Figure 1 shows two plots of the Newton convergence as a function of the total number of GMRES iterations ΣN_i according to Corollary 2.1, when GMRES converges superlinearly. The GMRES convergence is assumed to be the same in each Newton iteration, with

$$\|\boldsymbol{r}_1\| \leq 10^{-0.5}\|\boldsymbol{r}_0\|, \ \|\boldsymbol{r}_2\| \leq 10^{-1.5}\|\boldsymbol{r}_1\|, \ \ \|\boldsymbol{r}_3\| \leq 10^{-4.0}\|\boldsymbol{r}_2\|.$$

In both plots there are $\Sigma N_i = 6$ total GMRES iterations. In the left plot each Newton iteration only has a single GMRES iteration, i.e., there are $M = 6$ Newton iterations and $N = \{1, 1, 1, 1, 1, 1\}$. In the right plot there are $M = 3$ Newton iterations with $N = \{1, 2, 3\}$. The red line shows the Newton convergence for $\alpha = 0.1$, while the blue line has $\alpha = 0$ to serve as a reference.

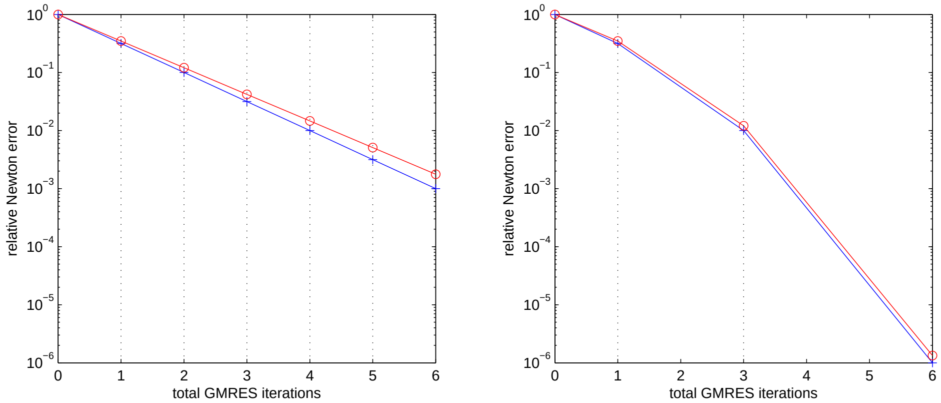

Figure 1: Newton convergence with superlinear GMRES convergence.

The difference is clear to see. In the left plot the Newton convergence is only linear in both the number of Newton iterations, and the number of GMRES iterations. This is due to the fact that only the first GMRES iteration is ever used.

The right plot yields a much better result after 6 iterations. The convergence is superlinear in both the Newton iterations and the GMRES iterations. The reason is that by doing more than one GMRES iteration per Newton iteration, we tap into the actual superlinear GMRES convergence, which was not the case in the left plot.

It is clear that, when GMRES is converging superlinearly, much is gained by doing as many GMRES iterations per Newton iteration as possible without oversolving. It will often even be worth some oversolving to get the maximum number of GMRES iterations in each Newton iteration. It is no wonder that over the years a lot of research has gone into finding good estimations for forcing terms that achieve this goal.

Now assume that in our Newton solver the GMRES method converges linearly with the same rate ρ in each iteration. Using Corollary 2.1, we then find

$$\| J\left(\boldsymbol{x}^{*}\right)\left(\boldsymbol{x}_{i}-\boldsymbol{x}^{*}\right) \| < (1+\alpha)^{i}\, 10^{-\rho N} \| J\left(\boldsymbol{x}^{*}\right)\left(\boldsymbol{x}_{0}-\boldsymbol{x}^{*}\right) \|, \tag{26}$$

or, equivalently,

$$\frac{\| J\left(\boldsymbol{x}^{*}\right)\left(\boldsymbol{x}_{i}-\boldsymbol{x}^{*}\right) \|}{\| J\left(\boldsymbol{x}^{*}\right)\left(\boldsymbol{x}_{0}-\boldsymbol{x}^{*}\right) \|} < 10^{-\rho N+\sigma i}, \text{ with } \sigma = {}^{10}\log\left(1+\alpha\right). \tag{27}$$

Figure 2 again shows two plots of the Newton convergence as a function of the GMRES iterations ΣN_i with the same variables as in Figure 1, except that GMRES convergence is now assumed to be linear with rate $\rho = 1$.

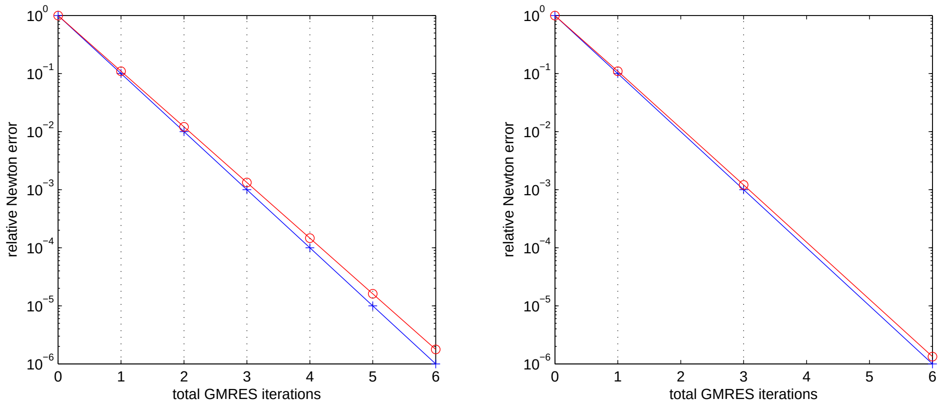

Figure 2: Newton convergence with linear GMRES convergence.

In the left plot the Newton convergence is linear in both the Newton and the GMRES iterations. Since $\Sigma N_i = M$, the relative error in this plot is $10^{(-\rho+\sigma)\Sigma N_i} \approx 10^{-0.996 \Sigma N_i}$.

In the right plot the Newton convergence is superlinear in the number of Newton iterations; however, it is approximately linear in the number of GMRES iterations. The convergence is very slightly better than in the left plot, because the factor $(1+\alpha)$ has only been applied 3 times, as opposed to 6 times in the left plot. However, for small α the difference will be negligible, as both plots approach the blue reference line.

We can conclude that when GMRES converges linearly and no oversolving is done, the convergence of the Newton-GMRES method is approximately linear in the GMRES iterations, independent of how many GMRES iterations are performed in each Newton iteration.

The only real downside of the method used for the left plot here is the overhead of the extra Newton iterations. Moreover, since GMRES becomes more expensive in each iteration, there is actually a benefit in doing only a single GMRES iteration per Newton iteration. In practice this will usually not be enough to outweigh the cost of the extra Newton iterations; however, it is clear that doing a maximum number of GMRES iterations per Newton iteration—as was clearly of great importance when GMRES converges superlinearly—is a lot less important when GMRES converges linearly. Instead, making sure that there is no oversolving may be much more important.

4 Conclusions

With Theorem 2.1 we were able to show some theoretical properties of the convergence of inexact Newton methods, that are especially interesting when the linear iterative solver is converging linearly. The key idea is to measure the convergence not in the number of Newton iterations, but in the total number of linear iterations done throughout all Newton iterations.

Using GMRES as an example we showed that when the linear solver converges superlinearly, Newton convergence is as generally expected and it is paramount to choose forcing terms that are as small as possible without doing too much oversolving.

However, when GMRES convergence is linear, we showed that Newton convergence—while still converging superlinearly in the Newton iterations, as expected—actually converges linearly in the GMRES iterations. As a result it is less clear how the forcing terms should be chosen, as it depends on the cost of doing extra Newton iterations, versus the extra cost of later GMRES iterations compared to the first GMRES iteration.

Where with superlinear GMRES convergence it seems appropriate to maximize the number of GMRES iterations per Newton iteration, and taking a little oversolving for granted, with linear GMRES convergence it seems more appealing to choose slightly larger forcing terms to ensure that no oversolving is done.

References

[1] R. S. Dembo, S. C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Numer. Anal.*, 19(2):400–408, 1982.

[2] S. C. Eisenstat and H. F. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM J. Sci. Comput.*, 17(1):16–32, 1996.

[3] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.