DELFT UNIVERSITY OF TECHNOLOGY

REPORT 11-14

On the convergence of inexact Newton methods

R. Idema, D.J.P. Lahaye, and C. Vuik

ISSN 1389-6520

Reports of the Department of Applied Mathematical Analysis

Delft 2011

On the convergence of inexact Newton methods

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

Abstract

The inexact Newton method is widely used to solve systems of non-linear equations. It is well-known that forcing terms should be chosen relatively large at the start of the process, and be made smaller during the iteration process. This paper explores the mechanics behind this behavior theoretically using a simplified problem, and presents theory that shows a proper choice of the forcing terms leads to a reduction in the non-linear error that is approximately equal to the forcing term in each Newton iteration. Further it is shown that under certain conditions the inexact Newton method converges linearly in the number of linear iterations performed throughout all Newton iterations. Numerical experiments are presented to illustrate the theory in practice.

1 Introduction

The Newton-Raphson method is usually the method of choice when solving systems of non-linear equations. Power flow computations in power systems, which lead to systems of non-linear equations, are no different. In our research towards improving power flow computations we have used the inexact Newton method, where an iterative linear solver is used for the linear systems [5, 6].

Analyzing the convergence of our numerical power flow experiments, some interesting behavior surfaced. The method converged quadratically in the non-linear iterations, as expected from Newton convergence theory. However, at the same time the convergence was approximately linear in the total number of linear iterations performed throughout the non-linear iterations. This observation led us to investigate the theoretical convergence of inexact Newton methods.

In this paper we present the results of this investigation. In Section 2 we treat the Newton-Raphson method and the inexact Newton method, to provide a foundation for the rest of the paper. The notion of oversolving is treated thoroughly by means of a simplified method, as many of the concepts that play a role there are needed to correctly interpret the convergence theory presented in Section 3. Also in Section 3, using the presented convergence theory, we give a theoretical explanation of the observed linear convergence in the amount of linear iterations. Subsequently, in Section 4 we treat some numerical experiments on power flow problems, to show the practical merit and consequences of the theoretical convergence, and in Section 5 we reflect on some possible applications of the presented convergence theory. Finally, in Section 6 we gather the conclusions of the paper.

*Delft University of Technology, Delft Institute of Applied Mathematics, Mekelweg 4, 2628 CD Delft, The Netherlands. E-mail: r.idema@tudelft.nl, d.j.p.lahaye@tudelft.nl, c.vuik@tudelft.nl

2 Inexact Newton Method

Consider the problem of finding a solution $\boldsymbol{x}^* \in \mathbb{R}$ for the system of non-linear equations

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

with $\boldsymbol{x} \in \mathbb{R}$, and $F : \mathbb{R}^n \to \mathbb{R}^n$ a non-linear function. Note that any system of non-linear equations can be written in this form.

The Newton-Raphson method is an iterative method for systems of non-linear equations, that linearizes the problem in each step, and updates iterate $\boldsymbol{x}_i$ with the solution of this linearized problem. Problem (1) may be solved with the Newton-Raphson method (Algorithm 1), provided that the Jacobian $J(\boldsymbol{x}_i)$ exists and is invertible for each iterate $\boldsymbol{x}_i$. The algorithm has local quadratic convergence, meaning that it converges quadratically when the initial solution is close enough to the solution. Global convergence can be attained by using line search or trust regions methods [8, 3].

Algorithm 1 Newton-Raphson Method

1: $i = 0$
2: given initial solution $\boldsymbol{x}_0$
3: **while** not converged **do**
4: solve $J(\boldsymbol{x}_i)\,\boldsymbol{s}_i = -F(\boldsymbol{x}_i)$
5: update solution $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \boldsymbol{s}_i$
6: $i = i + 1$
7: **end while**

The calculation of the exact solution of the linearized problem implies the use of a direct solver. Often it is impossible, or at least undesirable, to use a direct solver, mostly when the problem is very large. The inexact Newton method (Algorithm 2) is a natural adaptation of the Newton-Raphson method for use with a iterative linear solvers.

Algorithm 2 Inexact Newton Method

1: $i = 0$
2: given initial solution $\boldsymbol{x}_0$
3: **while** not converged **do**
4: *determine forcing term* η_i
5: solve $J(\boldsymbol{x}_i)\,\boldsymbol{s}_i = -F(\boldsymbol{x}_i)$ *up to accuracy* $\frac{\|J(\boldsymbol{x}_i)\boldsymbol{s}_i + F(\boldsymbol{x}_i)\|}{\|F(\boldsymbol{x}_i)\|} \leq \eta_i$
6: update solution $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \boldsymbol{s}_i$
7: $i = i + 1$
8: **end while**

The convergence of the inexact Newton method depends on the choice of the forcing terms η_i. The inexact Newton method is locally convergent provided that $\eta_i < 1$, and has local quadratic convergence if $\eta_i = \mathcal{O}(\|F(x_i)\|)$ [2]. Global convergence can again be attained by using line search or trust regions methods [1]. For a survey of Jacobian-free Newton-Krylov methods see [7].

It is clear that the same speed of convergence as the Newton-Raphson method can be attained with the inexact Newton method, simply by setting the forcing terms η_i very small. However, the power of the inexact Newton method lies in the fact that the same order of convergence can be reached with much less strict forcing terms, thus saving a lot of computing time. In each Newton iteration,

at some number of iterations of the linear solver performing extra linear iterations—although significantly reducing η_i—no longer significantly improve the non-linear iterate $\boldsymbol{x}_i$. Performing linear iterations beyond this point is known as oversolving. The concept op oversolving is treated in detail in Section 2.1.

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 altogether, 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 [4].

2.1 Oversolving

In this section we elaborate on when and why oversolving occurs. To this end we look at a single Newton step, simplifying the problem by assuming that we know true errors, instead of having to deal with residuals. The simplified situation is sketched in Figure 1, where the current iterate is $\boldsymbol{x}_i$, and an exact Newton step would make the next iterate $\tilde{\boldsymbol{x}}_{i+1}$.

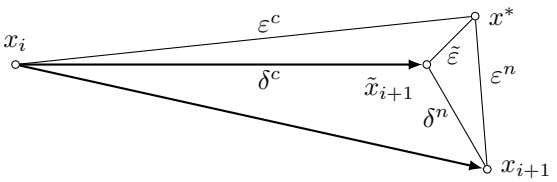

Figure 1: Inexact Newton step

Further we define:

$$\delta^c = \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_{i+1}\| > 0, \tag{2}$$
$$\delta^n = \|\boldsymbol{x}_{i+1} - \tilde{\boldsymbol{x}}_{i+1}\| \geq 0, \tag{3}$$
$$\varepsilon^c = \|\boldsymbol{x}_i - \boldsymbol{x}^*\| > 0 \tag{4}$$
$$\varepsilon^n = \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\|, \tag{5}$$
$$\tilde{\varepsilon} = \|\tilde{\boldsymbol{x}}_{i+1} - \boldsymbol{x}^*\| \geq 0, \tag{6}$$
$$\gamma = \frac{\tilde{\varepsilon}}{\delta^c} > 0. \tag{7}$$

Thus the relative error to the solution is given by $\frac{\varepsilon^n}{\varepsilon^c}$. We assume that this error is not known, but that we do know, and have direct control over, the relative error to the exact Newton step $\frac{\delta^n}{\delta^c}$. In the actual Newton method $\frac{\delta^n}{\delta^c}$ would translate to the forcing term η_i.

We would like that an improvement in the controllable error $\frac{\delta^n}{\delta^c}$, translates into a similar improvement in the error to the solution, i.e., we want that

$$\frac{\varepsilon^n}{\varepsilon^c} \leq (1+\alpha)\,\frac{\delta^n}{\delta^c} \tag{8}$$

for some small $\alpha > 0$. If α is too large, a reduction of $\frac{\delta^n}{\delta^c}$ only leads to a far smaller reduction of the relative distance to the solution $\frac{\varepsilon^n}{\varepsilon^c}$.

The worst case scenario can be identified as

$$\max \frac{\varepsilon^n}{\varepsilon^c} = \frac{\delta^n + \tilde{\varepsilon}}{|\delta^c - \tilde{\varepsilon}|} = \frac{\delta^n + \gamma\delta^c}{|1-\gamma|\,\delta^c} = \frac{1}{|1-\gamma|}\,\frac{\delta^n}{\delta^c} + \frac{\gamma}{|1-\gamma|}. \tag{9}$$

To guarantee that the inexact Newton step $\boldsymbol{x}_{i+1}$ is an improvement over $\boldsymbol{x}_i$, using equation (9), we require that

$$\frac{1}{|1-\gamma|}\frac{\delta^n}{\delta^c} + \frac{\gamma}{|1-\gamma|} < 1 \Leftrightarrow \frac{\delta^n}{\delta^c} + \gamma < |1-\gamma| \Leftrightarrow \frac{\delta^n}{\delta^c} < |1-\gamma| - \gamma. \tag{10}$$

If $\gamma \geq 1$ this would mean that $\frac{\delta^n}{\delta^c} < -1$, which is impossible. Thus to guarantee reduction of the error to the solution we need that

$$\frac{\delta^n}{\delta^c} < 1 - 2\gamma \Leftrightarrow 2\gamma < 1 - \frac{\delta^n}{\delta^c} \Leftrightarrow \gamma < \frac{1}{2} - \frac{1}{2}\frac{\delta^n}{\delta^c}. \tag{11}$$

As a result we can drop the absolute operators in equation (9).

This behavior of the simplified problem is also evident in the Newton method in practice. If the starting point is too far from the solution, then even the exact Newton step does not guarantee convergence. In terms of the simplified problem this means that γ is too large, i.e., that $\tilde{\varepsilon}$ is too large compared to δ^c. Note that for a quadratically converging method, like the Newton-Raphson method, we expect γ to converge to 0 quadratically. Therefore, if the method is converging, then equation (11) is guaranteed to hold from some point on forward in the iteration process.

Figure 2 shows plots of equation (9) for several values of γ on a logarithmic scale. The horizontal axis shows the number of digits improvement in the distance to the exact Newton step $d_\delta = -\log\frac{\delta^n}{\delta^c}$, while the vertical axis depicts the resulting minimum number of digits improvement in the distance to the solution $d_\varepsilon = -\log\left(\max\frac{\varepsilon^n}{\varepsilon^c}\right)$.

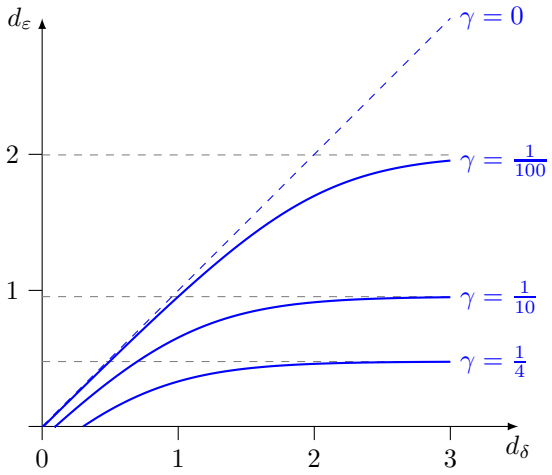

Figure 2: Number of digits improvement in the error

For fixed d_δ, the smaller the value of γ, the better the improvement d_ε is. For $\gamma = \frac{1}{10}$, we have a significant start-up cost on d_δ before d_ε becomes positive, and we can never get a full digit improvement on the distance to the solution. Making more than a 2 digit improvement in the error to the exact Newton step would result in a lot of effort with hardly any return at $\gamma = \frac{1}{10}$. When $\gamma = \frac{1}{100}$, however, there is hardly any start-up cost on d_δ, and we can improve the error to the solution up to nearly 2 digits.

The above mentioned start-up cost can be derived from equation (11) to be $d_\delta = -\log(1-2\gamma)$. The asymptote to which d_ε approaches is given by $d_\varepsilon = -\log\left(\frac{\gamma}{1-\gamma}\right) = \log\left(\frac{1}{\gamma} - 1\right)$, which is the improvement obtained when taking the exact Newton step.

The value α, as introduced in equation (8), is a measure of how far the graph of d_ε deviates from the ideal $d_\varepsilon = d_\delta$, which is attained only in the fictitious case that $\gamma = 0$. Combining equations (8)

and (9), we can investigate the minimum value of α needed for equation (8) to be guaranteed to hold:

$$\frac{1}{1-\gamma}\frac{\delta^n}{\delta^c} + \frac{\gamma}{1-\gamma} = (1+\alpha_{min})\frac{\delta^n}{\delta^c} \Leftrightarrow \tag{12}$$

$$\frac{1}{1-\gamma} + \frac{\gamma}{1-\gamma}\left(\frac{\delta^n}{\delta^c}\right)^{-1} = (1+\alpha_{min}) \Leftrightarrow \tag{13}$$

$$\alpha_{min} = \frac{\gamma}{1-\gamma}\left[\left(\frac{\delta^n}{\delta^c}\right)^{-1} + 1\right] \tag{14}$$

Figure 3 shows α_{min} as a function of $\frac{\delta^n}{\delta^c} \in [0, 1)$ for several values of γ. Left of the dotted line the equation (11) is met, i.e., improvement of the distance to the solution is guaranteed, whereas right of the dotted line this is not the case.

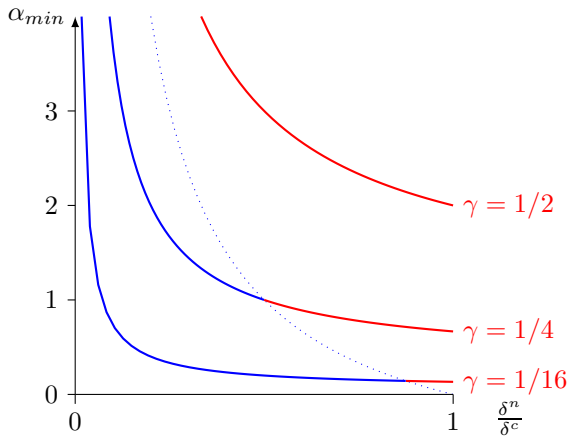

Figure 3: α_{min} as a function of the relative error to the exact Newton step

We see that for given γ, reducing $\frac{\delta^n}{\delta^c}$, makes α_{min} becomes larger. Especially for small $\frac{\delta^n}{\delta^c}$, α_{min} grows rapidly. Thus, the closer we bring $\boldsymbol{x}_{i+1}$ to the exact Newton step, the less the expected return in the error to the solution is. In the Newton method this means that if we choose η_i too small, we will be oversolving.

Further, we find that if γ is reduced, then α_{min} becomes smaller. It is clear that for very small γ a great improvement in the distance to the full Newton step can be made without compromising the return of investment on the distance to the actual solution. However, for γ nearing $\frac{1}{2}$, or more, any improvement in the distance to the full Newton step no longer guarantees a similar great improvement, if any, in the distance to the solution. Thus for such γ oversolving is inevitable. However, since the convergence of the Newton-Raphson method is quadratic, in later iterations γ will get smaller rapidly and we can choose η_i smaller and smaller without oversolving.

Although the results of this section are for a simplified problem, many of the elements known from using the Newton method in practice can be recognized, and explained by it. If we were to translate equation (8) to the Newton-Raphson method, we would get

$$\|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\| \leq (1+\alpha)\,\eta_i \|\boldsymbol{x}_i - \boldsymbol{x}^*\|. \tag{15}$$

The simplified problem suggests that for any choice of forcing term $\eta_i \in (0, 1)$ and $\alpha > 0$, there is some point in the iteration process from which on forward equation (15) is satisfied. In the following section we present a theorem that proves this idea for the actual Newton method, with just a change of the norm in which the distance to the solution measured.

3 Convergence Theory

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 $\boldsymbol{x}^*$,
- $J(\boldsymbol{x}^*)$ is continuous and non-singular.

In this section we present theory that relates the convergence of the inexact Newton method for the above problem directly to the chosen forcing terms. The following theorem is a variation on the inexact Newton convergence theorem presented in [2, Thm. 2.3].

Theorem 3.1. *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*

$$\|J(\boldsymbol{x}^*)(\boldsymbol{x}_{i+1} - \boldsymbol{x}^*)\| < (1+\alpha)\eta_i \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|. \tag{16}$$

Proof. Define

$$\mu = \max[\|J(\boldsymbol{x}^*)\|, \|J(\boldsymbol{x}^*)^{-1}\|] \geq 1. \tag{17}$$

Recall that $J(\boldsymbol{x}^*)$ is non-singular. Thus μ is well-defined and we can write

$$\frac{1}{\mu}\|\boldsymbol{y}\| \leq \|J(\boldsymbol{x}^*)\,\boldsymbol{y}\| \leq \mu\|\boldsymbol{y}\|. \tag{18}$$

Note that $\mu \geq 1$ because the induced matrix norm is sub-multiplicative.

Let

$$\gamma \in \left(0, \frac{\alpha\eta_i}{5\mu}\right) \tag{19}$$

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

$$\|J(\boldsymbol{y}) - J(\boldsymbol{x}^*)\| \leq \gamma, \tag{20}$$

$$\|J(\boldsymbol{y})^{-1} - J(\boldsymbol{x}^*)^{-1}\| \leq \gamma, \tag{21}$$

$$\|F(\boldsymbol{y}) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{y} - \boldsymbol{x}^*)\| \leq \gamma\|\boldsymbol{y} - \boldsymbol{x}^*\|. \tag{22}$$

That such an ε exists follows from [8, Thm. 2.3.3 & 3.1.5].

First we show that if $\|\boldsymbol{x}_i - \boldsymbol{x}^*\| < \mu^2\varepsilon$, then equation (16) holds.

Write

$$\begin{aligned} J(\boldsymbol{x}^*)(\boldsymbol{x}_{i+1} - \boldsymbol{x}^*) = &\left[I + J(\boldsymbol{x}^*)\left(J(\boldsymbol{x}_i)^{-1} - J(\boldsymbol{x}^*)^{-1}\right)\right] \\ &\cdot \left[\boldsymbol{r}_i + (J(\boldsymbol{x}_i) - J(\boldsymbol{x}^*))(\boldsymbol{x}_i - \boldsymbol{x}^*) - (F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*))\right]. \end{aligned} \tag{23}$$

Taking norms, and writing $\|\boldsymbol{r}_i\| = \|J(\boldsymbol{x}_i)\,\boldsymbol{s}_i + F(\boldsymbol{x}_i)\|$, we find

$$\begin{aligned} \|J(\boldsymbol{x}^*)(\boldsymbol{x}_{i+1} - \boldsymbol{x}^*)\| &\leq \left[1 + \|J(\boldsymbol{x}^*)\|\|J(\boldsymbol{x}_i)^{-1} - J(\boldsymbol{x}^*)^{-1}\|\right] \\ &\quad \cdot \left[\|\boldsymbol{r}_i\| + \|J(\boldsymbol{x}_i) - J(\boldsymbol{x}^*)\|\|\boldsymbol{x}_i - \boldsymbol{x}^*\| + \|F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|\right], \\ &\leq [1 + \mu\gamma] \cdot \left[\|\boldsymbol{r}_i\| + \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\| + \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\|\right], \\ &= [1 + \mu\gamma] \cdot \left[\eta_i\|F(\boldsymbol{x}_i)\| + 2\gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\|\right], \end{aligned} \tag{24}$$

where we used the definition of μ and equations (20)–(22).

Using that $F(\boldsymbol{x}^*) = \mathbf{0}$ by definition, we further write

$$F(\boldsymbol{x}_i) = [J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)] + [F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)], \tag{25}$$

and again taking norms find that

$$\begin{aligned} \|F(\boldsymbol{x}_i)\| &\leq \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| + \|F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| \\ &\leq \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| + \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\|. \end{aligned} \tag{26}$$

Now, substituting equation (26) into equation (24), and using (18), we get

$$\begin{aligned} &\|J(\boldsymbol{x}^*)(\boldsymbol{x}_{i+1} - \boldsymbol{x}^*)\| \\ &\leq (1+\mu\gamma)\left[\eta_i\left(\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| + \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\|\right) + 2\gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\|\right] \\ &= (1+\mu\gamma)\left[\eta_i(1+\mu\gamma) + 2\mu\gamma\right]\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|. \end{aligned} \tag{27}$$

Finally, using that $\gamma < \frac{\alpha\eta_i}{5\mu}$ and that both $\eta_i < 1$ and $\alpha\eta_i < 1$—the latter being a result from the requirement that $(1+\alpha)\eta_i < 1$—we can write

$$\begin{aligned} (1+\mu\gamma)\left[\eta_i(1+\mu\gamma) + 2\mu\gamma\right] &\leq \left(1 + \frac{\alpha\eta_i}{5}\right)\left[\eta_i\left(1 + \frac{\alpha\eta_i}{5}\right) + \frac{2\alpha\eta_i}{5}\right] \\ &= \left[\left(1 + \frac{\alpha\eta_i}{5}\right)^2 + \left(1 + \frac{\alpha\eta_i}{5}\right)\frac{2\alpha}{5}\right]\eta_i \\ &= \left[1 + \frac{2\alpha\eta_i}{5} + \frac{\alpha^2\eta_i^2}{25} + \frac{2\alpha}{5} + \frac{2\alpha^2\eta_i}{25}\right]\eta_i \\ &< \left[1 + \frac{2\alpha}{5} + \frac{\alpha}{25} + \frac{2\alpha}{5} + \frac{2\alpha}{25}\right]\eta_i \\ &< (1+\alpha)\eta_i. \end{aligned} \tag{28}$$

Equation (24) follows from substituting equation (28) into equation (27).

Given that equation (16) holds if $\|\boldsymbol{x}_i - \boldsymbol{x}^*\| < \mu^2\varepsilon$, we can now proceed to prove Theorem 3.1 by induction.

For the base case we have that

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| < \varepsilon \leq \mu^2\varepsilon. \tag{29}$$

Thus equation (16) holds for $i = 0$.

With the induction hypothesis that equation (16) holds for $i = 0, \ldots, k-1$, and using equation (18), we find that

$$\begin{aligned} \|\boldsymbol{x}_k - \boldsymbol{x}^*\| &\leq \mu\|J(\boldsymbol{x}^*)(\boldsymbol{x}_k - \boldsymbol{x}^*)\| \\ &< \mu(1+\alpha)^k\eta_{k-1}\cdots\eta_0\|J(\boldsymbol{x}^*)(\boldsymbol{x}_0 - \boldsymbol{x}^*)\| \\ &< \mu\|J(\boldsymbol{x}^*)(\boldsymbol{x}_0 - \boldsymbol{x}^*)\| \\ &\leq \mu^2\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \\ &< \mu^2\varepsilon. \end{aligned} \tag{30}$$

Thus equation (16) also holds for $i = k$, completing the proof. □

In other words, Theorem 3.1 states that for any choice of forcing terms $\eta_i \in (0,1)$, we can choose an $\alpha > 0$ arbitrarily small, and equation (16) will hold if the current iterate is close enough to the solution. This does not mean that for a certain iterate $\boldsymbol{x}_i$ we can just choose η_i and α arbitrarily small, and expect equation (16) to hold, as ε depends on the choices of η_i and α.

Another way of looking at it is that a given iterate $\boldsymbol{x}_i$—close enough to the solution to guarantee convergence—imposes the restriction that for Theorem 3.1 to hold the forcing terms η_i cannot be chosen too small. This is nothing new, as we already noted that choosing η_i too small leads to oversolving. Thus, using equation (19), we can consider $\eta_i > \frac{5\mu\gamma}{\alpha}$ a theoretical bound on 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 3.1, then we can characterize the theorem by saying that a convergence factor $(1+\alpha)\,\eta_i$ is attained if η_i is chosen such that there is no oversolving.

Corollary 3.1. *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*

$$\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| < (1+\alpha)^i \eta_{i-1}\cdots\eta_0 \|J(\boldsymbol{x}^*)(\boldsymbol{x}_0 - \boldsymbol{x}^*)\|. \tag{31}$$

Proof. The stated follows readily from the repeated application of Theorem 3.1. □

Using the same principles as in Theorem 3.1, we derive a relation between the non-linear residual norm $\|F(\boldsymbol{x}_i)\|$ and the error norm $\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|$ within the neighborhood of the solution where Theorem 3.1 holds.

Theorem 3.2. *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 $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| < \varepsilon$, then*

$$\left(1 - \frac{\alpha\eta_i}{5}\right)\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| < \|F(\boldsymbol{x}_i)\| < \left(1 + \frac{\alpha\eta_i}{5}\right)\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|. \tag{32}$$

Proof. Using that $F(\boldsymbol{x}^*) = \boldsymbol{0}$ by definition, we can again write

$$F(\boldsymbol{x}_i) = [J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)] + [F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)]. \tag{33}$$

Taking the norm on both sides and using equations (22) and (18) we find

$$\begin{aligned}
\|F(\boldsymbol{x}_i)\| &\leq \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| + \|F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| \\
&\leq \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| + \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\| \\
&\leq \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| + \mu\gamma\|J(\boldsymbol{x}^*)\boldsymbol{x}_i - \boldsymbol{x}^*\| \\
&= (1+\mu\gamma)\,\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|.
\end{aligned} \tag{34}$$

Similarly, we can derive

$$\begin{aligned}
\|F(\boldsymbol{x}_i)\| &\geq \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| - \|F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*) - J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| \\
&\geq \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| - \gamma\|\boldsymbol{x}_i - \boldsymbol{x}^*\| \\
&\geq \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\| - \mu\gamma\|J(\boldsymbol{x}^*)\boldsymbol{x}_i - \boldsymbol{x}^*\| \\
&= (1-\mu\gamma)\,\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|.
\end{aligned} \tag{35}$$

The theorem now follows from (19). □

A similar result as presented for the error $\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i - \boldsymbol{x}^*)\|$ in Theorem 3.1 can now be derived for $\|F(\boldsymbol{x}_i)\|$.

Theorem 3.3. *Let $\eta_i \in (0,1)$ and choose $\alpha > 0$ such that $(1+2\alpha)\,\eta_i < 1$. Then there exists an $\varepsilon > 0$ such that, if $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| < \varepsilon$, the sequence $\|F(\boldsymbol{x}_i)\|$ converges to 0, with*

$$\|F(\boldsymbol{x}_{i+1})\| < (1+2\alpha)\,\eta_i \|F(\boldsymbol{x}_i)\|. \tag{36}$$

Proof. Note that the conditions of Theorem 3.3 are such that Theorems 3.1 and 3.2 hold. Define μ and γ again as in Theorem 3.1.

Applying equation (34), Theorem 3.1, and equation (35), we find

$$\begin{aligned}\|F(\boldsymbol{x}_{i+1})\| &\leq (1+\mu\gamma)\,\|J(\boldsymbol{x}^*)(\boldsymbol{x}_{i+1}-\boldsymbol{x}^*)\| \\ &< (1+\mu\gamma)(1+\alpha)\,\eta_i \|J(\boldsymbol{x}^*)(\boldsymbol{x}_i-\boldsymbol{x}^*)\| \\ &\leq \frac{(1+\mu\gamma)}{(1-\mu\gamma)}(1+\alpha)\,\eta_i \|F(\boldsymbol{x}_i)\|. \end{aligned} \tag{37}$$

Now, using (19), we can write

$$\frac{1+\mu\gamma}{1-\mu\gamma} < \frac{1+\frac{\alpha\eta_i}{5}}{1-\frac{\alpha\eta_i}{5}} = \frac{1-\frac{\alpha\eta_i}{5}+\frac{2}{5}\alpha\eta_i}{1-\frac{\alpha\eta_i}{5}} = 1 + \frac{\frac{2}{5}\alpha\eta_i}{1-\frac{\alpha\eta_i}{5}} < 1 + \frac{\frac{2}{5}\alpha\eta_i}{\frac{4}{5}} = 1 + \frac{\alpha\eta_i}{2},$$

and using that both $\eta_i < 1$ and $2\alpha\eta_i < 1$—the latter being a result from the requirement that $(1+2\alpha)\,\eta_i < 1$—we find

$$\frac{1+\mu\gamma}{1-\mu\gamma}(1+\alpha) < \left(1+\frac{\alpha\eta_i}{2}\right)(1+\alpha) = 1 + \left(1+\frac{\eta_i}{2}\right)\alpha + \frac{1}{2}\eta_i\alpha^2 < 1+2\alpha.$$

□

3.1 When the Linear Solver Convergences Linearly

The inexact Newton method uses an iterative process in each Newton iteration to solve the linear Jacobian system $J(\boldsymbol{x}_i)\,\boldsymbol{s}_i = -F(\boldsymbol{x}_i)$ up to accuracy $\|J(\boldsymbol{x}_i)\,\boldsymbol{s}_i + F(\boldsymbol{x}_i)\| \leq \eta_i \|F(\boldsymbol{x}_i)\|$. In many practical applications the convergence of such a linear solver turns out to be approximately linear, i.e.,

$$\|\boldsymbol{r}_i^k\| \approx 10^{-\beta k}\|F(\boldsymbol{x}_i)\| \tag{38}$$

for some convergence rate $\beta > 0$. Here $\boldsymbol{r}_i^k = F(\boldsymbol{x}_i) + J(\boldsymbol{x}_i)\,\boldsymbol{s}_i^k$ is the linear residual after k iterations of the linear solver in Newton iteration i.

Let us suppose that the linear solver indeed converges linearly with the same rate β in each Newton iteration. Let K_i be the number of linear iterations performed in Newton iteration i, i.e., K_i is minimum integer such that $10^{-\beta K_i} \leq \eta_i$. Then we can apply Corollary 3.1 to find

$$\begin{aligned}\|J(\boldsymbol{x}^*)(\boldsymbol{x}_i-\boldsymbol{x}^*)\| &< (1+\alpha)^i\,\eta_{i-1}\cdots\eta_0\|J(\boldsymbol{x}^*)(\boldsymbol{x}_0-\boldsymbol{x}^*)\| \\ &= (1+\alpha)^i\,10^{-\beta N_i}\|J(\boldsymbol{x}^*)(\boldsymbol{x}_0-\boldsymbol{x}^*)\|, \end{aligned} \tag{39}$$

where $N_i = \sum_{j=0}^{i-1} K_j$ is the total number of linear iterations performed.

For small enough α, when the linear solver converges linearly, we thus find that the inexact Newton method converges approximately linearly in the total number of linear iterations. Note that this result is independent of the Newton convergence in the Newton iterations. If the forcing terms η_i are chosen properly, the method will converge quadratically in the number of Newton iterations, while it converges linearly in the number of linear iterations performed throughout those Newton iterations.

4 Numerical Experiments

Both the established Newton-Raphson convergence theory [8, 3], and the inexact Newton convergence theory by Dembo et al. [2], require the current iterate to be "close enough" to the solution. The bounds for what is "close enough" presented in these theorems are very hard to calculate in practice. However, decades of practice has shown that the convergence behavior presented in these theorems is reached within a few Newton steps for most problems. Thus the theory is not just of theoretical, but also of practical importance.

In this section, practical experiments are presented that illustrate that Theorem 3.1 also has practical merit, despite the restriction that the current iterate has to be "close enough" to the solution. Moreover, instead of convergence relation (16), an idealized version is tested, in which the error norm is changed to the 2-norm, and α is neglected:

$$\|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\| < \eta_i \|\boldsymbol{x}_i - \boldsymbol{x}^*\|. \tag{40}$$

If relation (40) holds, that means that any improvement of the linear residual error in a certain Newton iteration, improves the error in the non-linear iterate by an equal factor.

The test case used is a power flow problem on a power system with around 136k busses and 230k lines. The resulting non-linear system has approximately 256k equations, and the Jacobian matrix has around 2M non-zeros. The linear Jacobian systems are solved using GMRES [9], with a high quality ILU factorization of the Jacobian as preconditioner. For more information on the test case and solution method see [6].

In Figures 4–6 the test case is solved with different amounts of GMRES iterations per Newton iteration. In all cases two Newton steps with just a single GMRES iteration were performed at the start, but not represented in the figures. In each figure the solid line depicts the actual error $\|\boldsymbol{x}_i - \boldsymbol{x}^*\|$ found while solving the problem, while the dashed line represents the expected error following the idealized theoretical relation (40).

Figure 4 shows the distribution of GMRES iterations for a typical choice of forcing terms that leads to a fast solution of the problem. The practical convergence nicely follows the idealized theory, suggesting that the two initial Newton iterations, with a single GMRES iteration each, lead to an iterate $\boldsymbol{x}_2$ close enough to the solution for the theory to hold for the chosen forcing terms η_i. Note that $\boldsymbol{x}_2$ is in actuality still very far from the solution, and that it is unlikely that it satisfies the theoretical bound on the proximity to the solution required in Theorem 3.1.

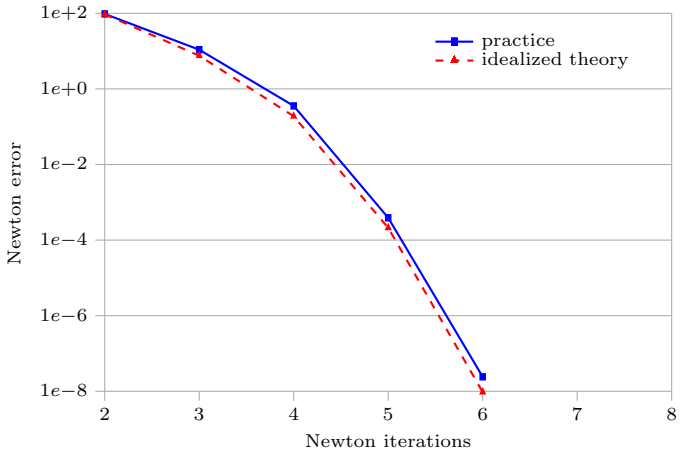

Figure 4: Convergence for GMRES iteration distribution 1,1,4,6,10,14

Figure 5 has a more exotic distribution of GMRES iterations per Newton iterations, illustrating

that practice can also follow theory nicely for such a scenario.

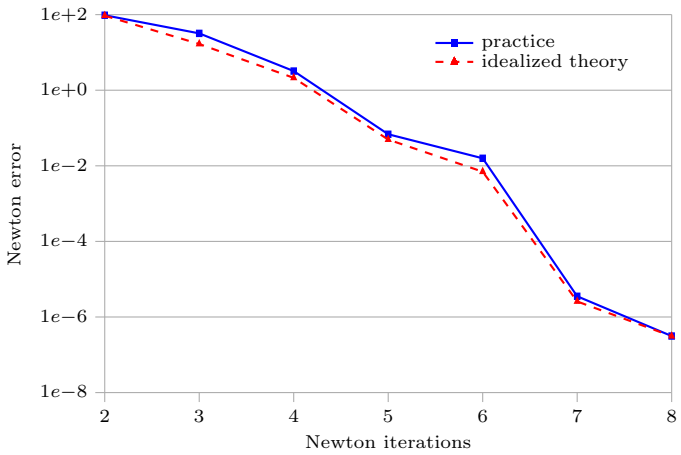

Figure 5: Convergence for GMRES iteration distribution 1,1,3,4,6,3,11,3

In Figure 6 the practical convergence is nowhere near the idealized theory. In terms of Theorem 3.1, this means that the iterates $\boldsymbol{x}_i$ are not close enough to be able to take the forcing terms η_i as small as they were chosen in this example. In practical terms this means that we are oversolving; too many GMRES iterations are performed per Newton iteration.

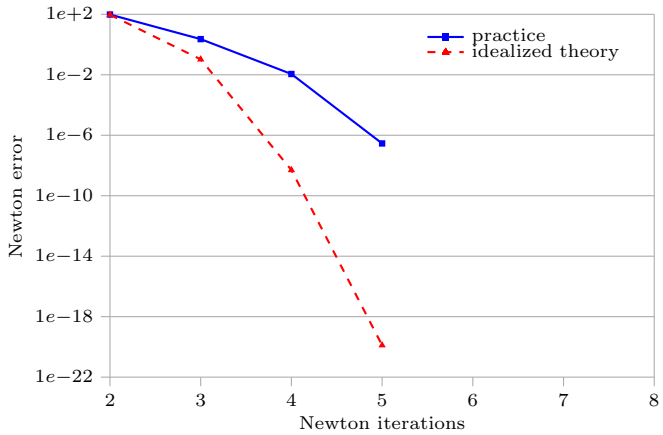

Figure 6: Convergence for GMRES iteration distribution 1,1,9,19,30

In Figures 7 and 8 we compare the convergence in the number of Newton iterations with the convergence in the number of GMRES iterations. In our test case the convergence of the GMRES solves is approximately linear, with similar rate of convergence in each Newton iteration. Thus we can use the same figures to also illustrate the theory from Section 3.1.

Figure 7 shows the convergence of the true error in the number of Newton iterations, for 5 different distributions of GMRES iterations per Newton iteration, i.e., for 5 different sets of forcing terms. The graphs are precisely as expected; the more GMRES iteration are performed in a Newton iteration, the better the convergence. A naive interpretation might conclude that option (A) is the best of the considered choices, and that option (E) is by far the worst. However, this is too simple a conclusion, as illustrated below.

Figure 8 shows the convergence of the true error in the total number of GMRES iterations for the same 5 distributions. The convergence of option (A) is worse than that of option (E) in this

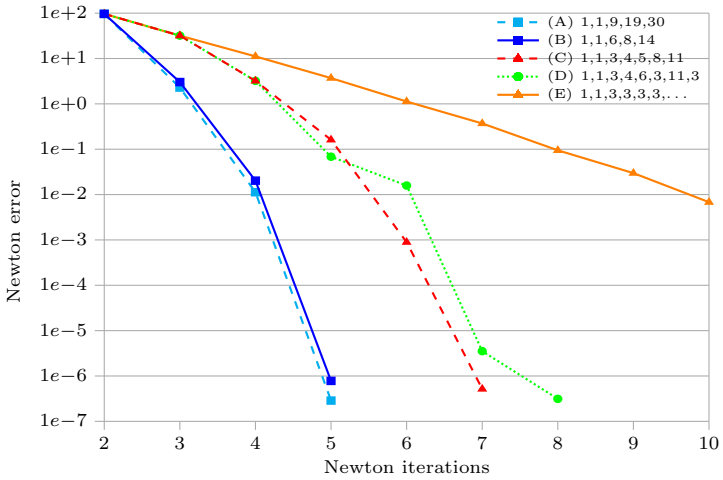

Figure 7: Convergence in the Newton iterations

figure, revealing that option (A) imposes a lot of oversolving. Option (E) is still the worst of the options that do not oversolve much, but it no longer seems as bad as suggested by Figure 7. Options (B), (C), and (D) show approximately linear convergence, as predicted by the theory of Section 3.1. As the practical GMRES convergence is not exactly linear, nor exactly the same in each Newton iteration, the convergence of these options is not identical, and option (E) is still quite a bit worse, but the strong influence of the near linear GMRES convergence is nonetheless very clear.

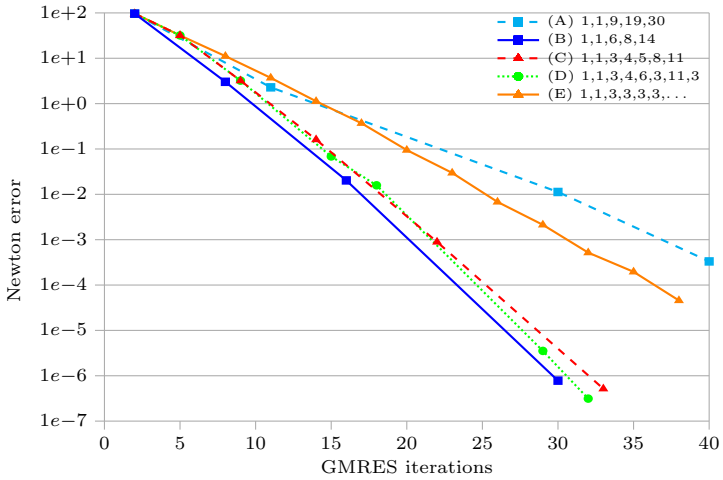

Figure 8: Convergence in the GMRES iterations

Neither Figure 7 nor Figure 8 tell the entire story. If the set-up time of a Newton iteration, generally the calculation of $J(\boldsymbol{x}_i)$ and $F(\boldsymbol{x}_i)$, is very high compared to the computational cost of iterations of the linear solver, then Figure 7 approximates the convergence in the solution time. However, if those set-up costs are negligible compared to the linear solves, then it is Figure 8 that approximates the convergence in the solution time. The practical truth is generally in between, but knowing which of these extremes a certain problem is closer to can be important to make the correct choice of forcing terms.

5 Applications

In this section we explore how we can use to knowledge from the previous sections to design better inexact Newton algorithm. First we investigate ideas to optimize the choice of the forcing terms, and after that we look at possible adaptations of the linear solver within the Newton process.

5.1 Forcing Terms

The ideas for the choice of the forcing terms η_i rely on the expectation that in Newton iteration i—provided that we are not oversolving—both the unknown true error and its known measure $\|F_i\|$ should reduce with an approximate factor η_i, as indicated by Theorem 3.1.

We can use this knowledge to choose the forcing terms adaptively by recalculating $\|F\left(\boldsymbol{x}_i + \boldsymbol{s}_i^k\right)\|$ in every linear iteration k, and checking whether the reduction in the norm of F is close enough to the reduction in the linear residual. Once the reduction in the norm of F starts lagging that of the linear residual, we know we are introducing oversolving, and should go to the next Newton iteration. Obviously, the above adaptive method only makes sense if the evaluation of $\|F\left(\boldsymbol{x}_i + \boldsymbol{s}_i^k\right)\|$ is cheap compared to doing extra linear iterations, which is often not the case. Otherwise we are back to making an educated guess of the forcing term at the start of each Newton iteration.

Theorem 3.1 can also be used to set a lower bound for the forcing terms. Assume that we want to solve up to a tolerance $\|F\| \leq \tau$. We expect that a forcing term $\eta_i = \frac{\tau}{\|F(\boldsymbol{x}_i)\|}$ is enough to approximately reach the asked non-linear tolerance, provided that there is no oversolving. Choosing η_i significantly smaller than that, will always lead to a waste of computational effort. Therefore, it makes sense to use $\frac{\tau}{\|F(\boldsymbol{x}_i)\|}$ as a lower bound for the forcing term guess in each Newton iteration.

Knowledge of the computational cost to set up a new Newton iteration, and of the convergence behavior of the used iterative linear solver, can help to choose better forcing terms. If the set-up cost of a Newton iteration is very high, it then makes sense to choose smaller forcing terms to get the most out of each Newton iteration. Similarly, if the linear solver converges superlinearly, slightly smaller forcing terms can be used to maximize the benefit of this superlinear convergence. On the other hand if the set-up cost of a Newton iteration is low, then it may yield better results to keep the forcing terms a bit larger to prevent oversolving, especially if the linear solver has only linear convergence, or worse.

5.2 Linear Solver

Given a forcing term η_i, we may adapt which linear solver we use to this value of this forcing term. For example, if we expect that only a few linear iterations are needed we may choose GMRES [9], whereas if many linear iterations are anticipated it could be better to use Bi-CGSTAB [12, 10] or IDR(s) [11]. In some cases, we may even choose to switch to a direct solver, if η_i becomes very small in later Newton iterations.

Instead of changing the entire linear solver between Newton iterations, we can also just change the preconditioning. For example, we may choose to take a higher quality preconditioners in later iterations. Or, alternatively, we can make a single preconditioner in the first Newton iteration and keep it through multiple Newton iterations, updating it depending on η_i.

For example, suppose that we have an iterate not too far from the solution, that we have some preconditioner, and that we found the linear solver to converge approximately linearly, as in our numerical experiments in Section 4. Then it is likely for the linear solver to converge approximately linear with similar convergence rate in the next iteration. This means that—given a forcing term—

we can make an approximation of the computational work needed to solve the linear problem. If we then know the amount of work needed to create a better preconditioner, and can approximate the work saved by this new preconditioner, we have an idea of whether we should keep the old preconditioner or build a new one.

Note that the presented ideas to adapt the linear solver during the Newton process do not necessarily rely on the convergence theory presented in this paper, but do rely on Newton-Raphson convergence in general.

6 Conclusions

In this paper the relation between the forcing terms and the convergence behavior of inexact Newton method was investigated.

The notion of oversolving was treated by means of a simplified problem, resulting in the conclusion that under certain conditions the convergence factor in a single inexact Newton iteration is approximately equal to the forcing term.

The same result was proven for the actual inexact Newton method, and the interpretation was discussed. Further some derived theory was treated, including the possibility of having linear convergence in the total number of linear iteration, while having the quadratic convergence in the number of non-linear iterations that is typical of Newton methods.

References

[1] P. N. Brown and Y. Saad. Hybrid Krylov methods for nonlinear systems of equations. *SIAM J. Sci. Stat. Comput.*, 11(3):450–481, 1990.

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

[3] J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM Classics. SIAM, Philadelphia, 1996.

[4] 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.

[5] R. Idema, D. J. P. Lahaye, C. Vuik, and L. van der Sluis. Fast Newton load flow. In *Transmission and Distribution Conference and Exposition, 2010 IEEE PES*, pages 1–7, April 2010.

[6] R. Idema, D. J. P. Lahaye, C. Vuik, and L. van der Sluis. Scalable Newton-Krylov solver for very large power flow problems. *IEEE Transactions on Power Systems*, Accepted for publication, 2011.

[7] D. A. Knoll and D. E. Keyes. Jacobian-free Newton-Krylov methods: a survey of approaches and applications. *J. Comp. Phys.*, 193:357–397, 2004.

[8] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. SIAM Classics. SIAM, Philadelphia, 2000.

[9] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7:856–869, 1986.

[10] G. L. G. Sleijpen, H. A. van der Vorst, and D. R. Fokkema. BiCGstab(ℓ) and other hybrid Bi-CG methods. *Numerical Algorithms*, 7:75–109, 1994.

[11] P. Sonneveld and M. B. van Gijzen. IDR(s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations. *SIAM J. Sci. Comput.*, 31(2):1035–1062, 2008.

[12] H. A. van der Vorst. Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 13:631–644, 1992.