Numerical Analysis Module 5
Solving Nonlinear Algebraic Equations

Sachin C. Patwardhan
Dept. of Chemical Engineering,
Indian Institute of Technology, Bombay
Powai, Mumbai, 400 076, Inda.
Email: sachinp@iitb.ac.in

Contents

1 Introduction

Consider set of n nonlinear simultaneous equations of type

$$f_i(\mathbf{x}) = 0 \quad for \;\; i = 1, 2, ..., n \tag{1}$$

$$\mathbf{F}(\mathbf{x}) = \overline{\mathbf{0}} \tag{2}$$

where $\mathbf{x} \in R^n$ and $\mathbf{F}(.) : R^n \rightarrow R^n$ represents a $n \times 1$ function vector. This problem may have no solution, an infinite number of solutions or any finite number of solutions. In the module on Problem Discretization using Approximation Theory, we have already introduced a basic version of the Newton's method, in which a sequence of approximate linear transformations is constructed to solve equation (2). In this module, we develop this method further and also discuss the conditions under which it converges to the solution. In addition, we discuss the following two approaches that are frequently used for solving nonlinear algebraic equations: (a) method of successive substitutions and (b) unconstrained optimization. Towards the end of the module, we briefly touch upon two fundamental issues related to nonlinear algebraic equations, namely (a) the (local) existence uniqueness of the solutions and (b) the notion of conditioning of nonlinear algebraic equations.

2 Method of Successive Substitutions [4]

In many situations, equation (2) can be rearranged as

$$\mathbf{Ax} = \mathbf{G}(\mathbf{x}) \tag{3}$$

$$\mathbf{G}(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) & g_2(\mathbf{x}) & .... & g_n(\mathbf{x}) \end{bmatrix}^T \tag{4}$$

such that the solution of equation (3) is also solution of equation (2). The nonlinear Equation (3) can be used to formulate iteration sequence of the form

$$\mathbf{Ax}^{(k+1)} = \mathbf{G}\left[\mathbf{x}^{(k)}\right] \tag{5}$$

Given a guess $\mathbf{x}^{(k)}$,the R.H.S. is a fixed vector, say $\mathbf{b}^{(k)} = \mathbf{G}\left[\mathbf{x}^{(k)}\right]$, and computation of the next guess $\mathbf{x}^{(k+1)}$ essentially involves solving the linear algebraic equation

$$\mathbf{Ax}^{(k+1)} = \mathbf{b}^{(k)}$$

at each iteration. Thus, the set of nonlinear algebraic equations is solved by formulating a sequence of linear sub-problems. Computationally efficient method of solving such sequence of linear problems would be to use **LU** decomposition of matrix **A**.

A special case of interest is when matrix $\mathbf{A} = \mathbf{I}$ in equation (5).In this case, if the set of equations given by (3) can be rearranged as follows

$$x_i = g_i(\mathbf{x}) \quad \text{for } i = 1, 2, ...n \tag{6}$$

then, the method of successive substitution can be implemented using either of the following iteration schemes

- **Jacobi-Iterations**

$$x_i^{(k+1)} = g_i[\mathbf{x}^{(k)}] \quad \text{for } i = 1, 2, ...n \tag{7}$$

- **Gauss Seidel Iterations**

$$x_i^{(k+1)} = g_i[x_1^{(k+1)}, ..., x_{i-1}^{(k+1)}, x_i^{(k)}, ..., x_n^{(k)}] \tag{8}$$

$$i = 1, 2, ..., n$$

- **Relaxation Iterations**

$$x_i^{(k+1)} = x_i^{(k)} + \omega_k[g_i(\mathbf{x}^k) - x_i^{(k)}] \tag{9}$$

The iterations can be terminated when

$$\frac{\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|}{\left\|\mathbf{x}^{(k+1)}\right\|} < \varepsilon \tag{10}$$

A major advantage of the iteration schemes discussed in this section is that they do not involve gradient calculations. However, these schemes are likely to converge if a good initial guess can be generated, which is close to the solution. We postpone the discussion on convergence of these methods to the end of this module.

3 Newton's Method

For a general set of simultaneous equations $\mathbf{F}(\mathbf{x}) = \overline{\mathbf{0}}$, it may not be always possible to transform to form $\mathbf{Ax} = \mathbf{G}(\mathbf{x})$ by simple rearrangement of equations. Even when it is possible, the iterations may not converge. When the function vector $\mathbf{F}(\mathbf{x})$ is once differentiable, Newton's method provides a way to transform an arbitrary set of equations $\mathbf{F}(\mathbf{x}) = \overline{\mathbf{0}}$ to form $\mathbf{Ax} = \mathbf{G}(\mathbf{x})$ using Taylor series expansion. In this section we introduce univariate and multivariate Newton's method and their popular variants.

3.1 Univariate Newton Type Methods

In the module on Problem Discretization using Approximation Theory, we have already introduced the Newton's method. For the univariate case, i.e. for solving scalar equation $f(x) = 0$ where $x \in R$, this reduces to the following iteration equation

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{df(x^{(k)})/dx} \tag{11}$$

Secant method is a variation of the univariate Newton's method, which is based on using approximation of the function derivative using Taylor series approximation of $f(x^{(k)})$ in the neighborhood the previous iteration $x^{(k-1)}$. Thus, term $df(x^{(k)})/dx$ appearing in equation (11) is approximated as follows

$$\frac{df(x^{(k)})}{dx} \approx \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$$

and the iteration equation reduces to

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}) \tag{12}$$

To initialize iterations by secant method, we need two initial guesses $x^{(0)}$ and $x^{(1)}$. Note that the secant method does not require $f(x)$ to be a differentiable function. The convergence properties of the secant method can be improved if f(x) is continuous on $[x^{(0)}, x^{(1)}]$ and $f(x^{(0)})$ and $f(x^{(1)})$ have opposite sign. In such a case, there is at least one root of f(x) in the interval $[x^{(0)}, x^{(1)}]$. This bracketing of the root is continued through the iterations. Let us assume that we have

$$f(x^{(k)}) > 0 \text{ and } f(x^{(k-1)}) < 0$$

When we move to compute $x^{(k+2)}$, the iterations proceed in either of the two directions

$$\text{Case } f(x^{(k+1)}) < 0 : x^{(k+2)} = x^{(k+1)} - \frac{x^{(k+1)} - x^{(k)}}{f(x^{(k+1)}) - f(x^{(k)})} f(x^{(k+1)})$$

$$\text{Case } f(x^{(k+1)}) > 0 : x^{(k+2)} = x^{(k+1)} - \frac{x^{(k+1)} - x^{(k-1)}}{f(x^{(k+1)}) - f(x^{(k-1)})} f(x^{(k+1)})$$

This variation of secant method, known as *regula falsi* method, makes sure that at least one root is bracketed in the successive iterates.

3.2 Multi-variate Secant or Wegstein Iterations

This approach can be viewed as a multi-variate extension of the secant method. Alternatively, this method can also be interpreted as relaxation iteration with variable relaxation

parameter. By this approach, the relaxation parameter is estimated by applying a secant method (i.e. local derivative approximation) independently to each component of $\mathbf{x}$ **[2].** Let us define $f_i(\mathbf{x})$ as

$$f_i(\mathbf{x}) = x_i - g_i(\mathbf{x})$$

and slopes $s_i^{(k)}$ as follows

$$s_i^{(k)} = \frac{\left[g_i(\mathbf{x}^{(k)}) - g_i(\mathbf{x}^{(k-1)})\right]}{\left[x_i^{(k)} - x_i^{(k-1)}\right]} \tag{13}$$

At the k^{th} iteration, if we apply a Newton-type method to individual component, then we have

$$x_i^{(k+1)} = x_i^{(k)} - f_i(\mathbf{x}^{(k)}) \frac{x_i^{(k)} - x_i^{(k-1)}}{f_i(\mathbf{x}^{(k)}) - f_i(\mathbf{x}^{(k-1)})} \tag{14}$$

$$\begin{aligned}
x_i^{(k+1)} &= x_i^{(k)} - \left[x_i^{(k)} - g_i(\mathbf{x}^{(k)})\right] \frac{\left(x_i^{(k)} - x_i^{(k-1)}\right)}{\left(x_i^{(k)} - x_i^{(k-1)}\right) - [g_i(\mathbf{x}^{(k)}) - g_i(\mathbf{x}^{(k-1)})]} \\
&= x_i^{(k)} - \left[x_i^{(k)} - g_i(\mathbf{x}^{(k)})\right] \frac{1}{1 - s_i^{(k)}} \\
&= \left[1 - \frac{1}{1 - s_i^{(k)}}\right] x_i^{(k)} + \frac{1}{1 - s_i^{(k)}} g_i(\mathbf{x}^{(k)}) \\
&= \left(1 - \omega_i^{(k)}\right) x_i^{(k)} + \omega_i^{(k)} g_i(\mathbf{x}^{(k)})
\end{aligned}$$

$$for \quad i = 1, 2, ..., n$$

where $\omega_i^{(k)} = 1/\left(1 - s_i^{(k)}\right)$. To avoid large changes, the value of $\omega_i^{(k)}$ is typically clipped between $0 < \omega_i^{(k)} \leq \alpha$ and typical value for α is 5 [3].

3.3 Multivariate Newton's Method

The main idea behind the Newton's method is solving the set of nonlinear algebraic equations (2) by formulating a sequence of linear subproblems of type

$$\begin{aligned}
\mathbf{A}^{(k)} \Delta \mathbf{x}^{(k)} &= \mathbf{b}^{(k)} \\
\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \Delta \mathbf{x}^{(k)} \\
k &= 0, 1, 2, ...
\end{aligned}$$

in such a way that sequence $\left\{\mathbf{x}^{(k)} : k = 0, 1, 2, ...\right\}$ converges to solution of equation (2). The basic version of Newton's method was derived in Section 3.4 of Lecture Notes on Problem Discretization Using Approximation Theory. We obtain the following set of recursive

equations

$$\mathbf{J}^{(k)} = \left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(k)}\right)}{\partial \mathbf{x}}\right] \quad ; \quad \mathbf{F}^{(k)} = \mathbf{F}\left[\mathbf{x}^{(k)}\right]$$

$$\Delta \mathbf{x}^{(k)} = -\left[\mathbf{J}^{(k)}\right]^{-1} \mathbf{F}^{(k)} \tag{15}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta \mathbf{x}^{(k)} \tag{16}$$

Alternatively, iterations can be formulated by solving

$$\left[\mathbf{J}^{(k)T}\mathbf{J}^{(k)}\right] \Delta \mathbf{x}^{(k)} = -\mathbf{J}^{(k)T}\mathbf{F}^{(k)} \tag{17}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta \mathbf{x}^{(k)} \tag{18}$$

where $\left[\mathbf{J}^{(k)T}\mathbf{J}^{(k)}\right]$ is symmetric and positive definite matrix. Iterations can be terminated when either of the following convergence criteria are satisfied

$$||\mathbf{F}^{(k+1)}|| < \varepsilon$$

$$\frac{\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|}{\|\mathbf{x}^{(k+1)}\|} < \varepsilon \tag{19}$$

This approach is called as simplified Newton's *method.*

3.4 Damped Newton Method

Often the simplified Newton's method finds a large step $\Delta \mathbf{x}^{(k)}$ such that the approximation of the function vector $\mathbf{F}(\mathbf{x})$ by the linear term in Taylor series is not valid in interval $\left[\mathbf{x}^{(k)}, \mathbf{x}^{(k)} + \Delta \mathbf{x}^{(k)}\right]$. In order to alleviate this difficulty, we find a corrected $\mathbf{x}^{(k+1)}$ by introducing a relaxation parameter λ as follows

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(k)} \Delta \mathbf{x}^{(k)}$$

where $0 < \lambda^{(k)} \leq 1$ is chosen such that

$$||\mathbf{F}(\mathbf{x}^{(k+1)})|| < ||\mathbf{F}(\mathbf{x}^{(k)})||$$

In practical problems this *damped Newton's algorithm* converges faster than the one that uses $\lambda^{(k)} = 1$. To see rationale behind this approach, consider a scalar function defined as follows [2]

$$\begin{aligned} \phi(\mathbf{x}^{(k+1)}) &= \frac{1}{2}\mathbf{F}(\mathbf{x}^{(k+1)})^T \mathbf{F}(\mathbf{x}^{(k+1)}) \\ &= \frac{1}{2}\mathbf{F}(\mathbf{x}^{(k)} + \lambda \Delta \mathbf{x}^{(k)})^T \mathbf{F}(\mathbf{x}^{(k)} + \lambda \Delta \mathbf{x}^{(k)}) \end{aligned}$$

where $\Delta \mathbf{x}^{(k)} = -\left[\mathbf{J}^{(k)}\right]^{-1} \mathbf{F}^{(k)}$. Using Taylor series expansion in the neighborhood of $\mathbf{x}^{(k)}$, we have

$$\begin{aligned}\phi(\mathbf{x}^{(k+1)}) &= \phi(\mathbf{x}^{(k)}) + \lambda \frac{\partial \phi}{\partial \lambda} + \frac{\lambda^2}{2} \frac{\partial^2 \phi}{\partial \lambda^2} + ... \\ &= \phi(\mathbf{x}^{(k)}) + \lambda \nabla \phi(\mathbf{x}^{(k)})^T \left(\Delta \mathbf{x}^{(k)}\right) + \frac{\lambda^2}{2} \left(\Delta \mathbf{x}^{(k)}\right)^T \nabla^2 \phi(\mathbf{x}^{(k)}) \left(\Delta \mathbf{x}^{(k)}\right) + ...\end{aligned}$$

Now, it is easy to see that

$$\nabla \phi(\mathbf{x}^{(k)}) = \left[\frac{\partial \mathbf{F}(\mathbf{x}^{(k)})}{\partial \mathbf{x}}\right]^T \mathbf{F}(\mathbf{x}^{(k)}) = \left[\mathbf{J}^{(k)}\right]^T \mathbf{F}^{(k)}$$

and

$$\nabla \phi(\mathbf{x}^{(k)})^T \left(\Delta \mathbf{x}^{(k)}\right) = -\left[\mathbf{F}^{(k)}\right]^T \mathbf{F}^{(k)} = -2\phi(\mathbf{x}^{(k)}) < 0$$

Now, if we let $\lambda \to 0$ in the Taylor series expansion, then

$$\phi(\mathbf{x}^{(k+1)}) - \phi(\mathbf{x}^{(k)}) \approx -2\lambda \phi(\mathbf{x}^{(k)}) < 0$$

Thus, for sufficiently small λ, the Newton's step will reduce $\phi(\mathbf{x})$. This property forms the basis of the damped Newton's method. A popular approach to determine the step length is Armijo line search [2]. Algorithm for the damped Newton's method together with Armijo line search (using typical values of tuning parameters β and η) is listed in Table 1.

Remark 1 *Note that the Newton's iterations can be expressed in form (3) as follows*

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \lambda \left[\left(\frac{\partial \mathbf{F}(\mathbf{x}^{(k)})}{\partial \mathbf{x}}\right)\right]^{-1} \mathbf{F}(\mathbf{x}^{(k)}) = \mathbf{G}(\mathbf{x}^{(k)}) \tag{20}$$

i.e.

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - \lambda \left[\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right]^{-1} \mathbf{F}(\mathbf{x})$$

It is easy to see that at the point $\mathbf{x} = \mathbf{x}^*$ *where* $\mathbf{F}(\mathbf{x}^*) = \overline{\mathbf{0}}$, *the iteration equation (20) has a fixed point* $\mathbf{x}^* = \mathbf{G}(\mathbf{x}^*)$.

3.5 Quasi-Newton Method with Broyden Update

A major difficulty with Newton method is that it requires calculation of Jacobian matrix at each iteration. A variation of Newton's method involves use of the Jacobian computed

Table 1: Damped Newton's Algorithm with Armijo Line SEarch

INITIALIZE: $\mathbf{x}^{(0)}, \varepsilon_1, \varepsilon_2, \alpha, k, k_{\max,}\beta, \eta, \delta_2$

$\delta_1 = \left\|\mathbf{F}^{(0)}\right\|$

WHILE $[(\delta_1 > \varepsilon_1)\ AND\ (\delta_2 > \varepsilon_2)\ AND\ (k < k_{max})]$

$\Delta\mathbf{x}^{(k)} = -\left[\mathbf{J}^{(k)T}\mathbf{J}^{(k)}\right]^{-1}\left[\mathbf{J}^{(k)}\right]^{T}\mathbf{F}^{(k)})$

Set $\lambda = 1,\ \beta = 0.1, \eta = 0.1$

$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda\Delta\mathbf{x}^{(k)}$

$\delta_1 = \frac{1}{2}\left[\left\|\mathbf{F}\left(\mathbf{x}^{(k+1)}\right)\right\|_2^2 - \left\|\mathbf{F}^{(k)})\right\|_2^2 + 2\beta\lambda\left\|\mathbf{F}^{(k)})\right\|_2^2\right]$

WHILE$[\delta_1 > 0]$

$$\lambda_q = \frac{\lambda\left\|\mathbf{F}^{(k)}\left(\mathbf{x}^{(k)}\right)\right\|_2^2}{(2\lambda - 1)\left\|\mathbf{F}^{(k)}\left(\mathbf{x}^{(k)}\right)\right\|_2^2 + \left\|\mathbf{F}\left(\mathbf{x}^{(k)} + \lambda\Delta\mathbf{x}^{(k)}\right)\right\|_2^2}$$

$\gamma = \max\{\eta, \lambda_q\}$

$\lambda = \gamma\lambda$

$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda\Delta\mathbf{x}^{(k)}$

$\delta_1 = \frac{1}{2}\left[\left\|\mathbf{F}\left(\mathbf{x}^{(k+1)}\right)\right\|_2^2 - \left\|\mathbf{F}^{(k)})\right\|_2^2 + 2\beta\lambda\left\|\mathbf{F}^{(k)})\right\|_2^2\right]$

END WHILE

$\delta_{2=}||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|| \,/\, ||\mathbf{x}^{(k+1)}||$

$k = k + 1$

END WHILE

only at the beginning of the iterations. By this approach, the Newton's step is computed as follows [6]

$$\mathbf{J}^{(0)} = \left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right] \tag{21}$$

$$\Delta \mathbf{x}^{(k)} = -\left[\mathbf{J}^{(0)}\right]^{-1} \mathbf{F}^{(k)} \tag{22}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta \mathbf{x}^{(k)} \tag{23}$$

There is another variation of this approach where the Jacobian is changed periodically but at much less frequency than the iterations.

Alternatively, the quasi-Newton methods try to overcome the difficulty associated with Jacobian calculations by generating *approximate* successive Jacobians using function vectors evaluated at previous iterations. While moving from iteration k to $(k+1)$, if $||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}||$ is not too large, then it can be argued that $\mathbf{J}^{(k+1)}$ is "**close**" to $\mathbf{J}^{(k)}$. Under such situation ,we can use the following rank-one update of the Jacobian matrix

$$\mathbf{J}^{(k+1)} = \mathbf{J}^{(k)} + \mathbf{y}^{(k)}[\mathbf{z}^{(k)}]^T \tag{24}$$

Here, $\mathbf{y}^{(k)}$ and $\mathbf{z}^{(k)}$ are two vectors that depend on $\mathbf{x}^{(k)}$, $\mathbf{x}^{(k+1)}$, $\mathbf{F}^{(k)}$ and $\mathbf{F}^{(k+1)}$. To arrive at the update formula, consider Jacobian $\mathbf{J}^{(k)}$ that produces step $\Delta \mathbf{x}^{(k)}$ as

$$\mathbf{J}^{(k)} \Delta \mathbf{x}^{(k)} = -\mathbf{F}^{(k)} \tag{25}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta \mathbf{x}^{(k)} \tag{26}$$

Step $\mathbf{x}^{(k+1)}$ predicts a function change

$$\Delta \mathbf{F}^{(k)} = \mathbf{F}^{(k+1)} - \mathbf{F}^{(k)} \tag{27}$$

We impose the following two conditions to obtain estimate of $\mathbf{J}^{(k+1)}$.

1. In the direction perpendicular to $\Delta \mathbf{x}^{(k)}$, our knowledge about $\mathbf{F}$ is maintained by new Jacobian estimate $\mathbf{J}^{(k+1)}$. This means for a vector, say $\mathbf{r}$, if $[\Delta \mathbf{x}^{(k)}]^T \mathbf{r} = 0$, then

   $$\mathbf{J}^{(k)} \mathbf{r} = \mathbf{J}^{(k+1)} \mathbf{r} \tag{28}$$

   In other words, both $\mathbf{J}^{(k)}$and $\mathbf{J}^{(k+1)}$ will predict some change in direction perpendicular to $\Delta \mathbf{x}^{(k)}$.

2. $\mathbf{J}^{(k+1)}$ predicts for $\Delta \mathbf{x}^{(k)}$, the same $\Delta \mathbf{F}^{(k)}$ in linear expansion, i.e.,

   $$\mathbf{F}^{(k+1)} = \mathbf{F}^{(k)} + \mathbf{J}^{(k+1)} \Delta \mathbf{x}^{(k)} \tag{29}$$

   or

   $$\mathbf{J}^{(k+1)} \Delta \mathbf{x}^{(k)} = \Delta \mathbf{F}^{(k)} \tag{30}$$

Now, for vector $\mathbf{r}$ perpendicular to $\Delta\mathbf{x}^{(k)}$, we have

$$\mathbf{J}^{(k+1)}\ \mathbf{r} = \mathbf{J}^{(k)}\mathbf{r} + \mathbf{y}^{(k)}[\mathbf{z}^{(k)}]^T\ \mathbf{r} \tag{31}$$

As

$$\mathbf{J}^{(k+1)}\ \mathbf{r} = \mathbf{J}^{(k)}\ \mathbf{r} \tag{32}$$

We have

$$\mathbf{y}^{(k)}[\mathbf{z}^{(k)}]^T\ \mathbf{r} = 0 \tag{33}$$

Since $\Delta\mathbf{x}^{(k)}$ is perpendicular to $\mathbf{r}$, we can choose $\mathbf{z}^{(k)} = \Delta\mathbf{x}^{(k)}$. Substituting this choice of $\mathbf{z}^{(k)}$ in equation (24) and post multiplying equation (24) by $\Delta\mathbf{x}^{(k)}$, we get

$$\mathbf{J}^{(k+1)}\Delta\mathbf{x}^{(k)} = \mathbf{J}^{(k)}\Delta\mathbf{x}^{(k)} + \mathbf{y}^{(k)}[\Delta\mathbf{x}^{(k)}]^T\Delta\mathbf{x}^{(k)} \tag{34}$$

Using equation (30), we have

$$\Delta\mathbf{F}^{(k)} = \mathbf{J}^{(k)}\Delta\mathbf{x}^{(k)} + \mathbf{y}^{(k)}[\Delta\mathbf{x}^{(k)}]^T\Delta\mathbf{x}^{(k)} \tag{35}$$

which yields

$$\mathbf{y}^{(k)} = \frac{\left[\Delta\mathbf{F}^{(k)} - \mathbf{J}^{(k)}\Delta\mathbf{x}^{(k)}\right]}{[[\Delta\mathbf{x}^{(k)}]^T\Delta\mathbf{x}^{(k)}]} \tag{36}$$

Thus, the Broyden's update formula for the Jacobian is

$$\mathbf{J}^{(k+1)} = \mathbf{J}^{(k)} + \frac{[\Delta\mathbf{F}^{(k)} - \mathbf{J}^{(k)}\Delta\mathbf{x}^{(k)}][\Delta\mathbf{x}^{(k)}]^T}{[[\Delta\mathbf{x}^{(k)}]^T\Delta\mathbf{x}^{(k)}]} \tag{37}$$

This can be further simplified as

$$\Delta\mathbf{F}^{(k)} - \mathbf{J}^{(k)}\Delta\mathbf{x}^{(k)} = \mathbf{F}^{(k+1)} - (\mathbf{F}^{(k)} + \mathbf{J}^{(k)}\Delta\mathbf{x}^{(k)}) = \mathbf{F}^{(k+1)} \tag{38}$$

Thus, Jackobian can be updated as

$$\mathbf{J}^{(k+1)} = \mathbf{J}^{(k)} + \frac{1}{[[\Delta\mathbf{x}^{(k)}]^T\Delta\mathbf{x}^{(k)}]}\left[\mathbf{F}^{(k+1)}[\Delta\mathbf{x}^{(k)}]^T\right] \tag{39}$$

Broyden's update can be derived by an alternate approach, which yields the following update formula [8]

$$\mathbf{J}^{(k+1)} = \mathbf{J}^{(k)} - \frac{1}{[[\mathbf{p}^{(k)}]^T\mathbf{J}^{(k)}\Delta\mathbf{F}^{(k)}]}\left[\mathbf{J}^{(k)}\Delta\mathbf{F}^{(k)} - \mathbf{p}^{(k)}\right][\mathbf{p}^{(k)}]^T\mathbf{J}^{(k)} \tag{40}$$

Example 2 ***Continuous Stirred Tank Reactor*** *The system under consideration is a Continuous Stirred Tank Reactor (CSTR) in which a non-isothermal, irreversible first order reaction*

$$A \longrightarrow B$$

is taking place. The steady state model for a non-isothermal CSTR is given as follows :

$$\begin{aligned} 0 &= \frac{F}{V}(C_{A0} - C_A) - k_0 \exp(-\frac{E}{RT})C_A \\ 0 &= \frac{F}{V}(T_0 - T) + \frac{(-\Delta H_r)\, k_0}{\rho C_p} \exp(-\frac{E}{RT})C_A - \frac{Q}{V\rho C_p} \\ Q &= \frac{aF_c^{b+1}}{F_c + \left(\frac{aF_c^b}{2\rho_c C_{pc}}\right)}(T - T_{cin}) \end{aligned}$$

Model parameters are listed in Table 1, Example 4 in Module on Solving Nonlinear ODE-IVPs. The set of model parameters considered here correspond to the stable steady state. The problem at hand is to find steady state $C_A = 0.265$ *and* $T = 393.954$ *starting from an initial guess for the steady state. Figure 1 to 4 show progress of Newton's method, Newton's method with initial Jacobian, Newton's method with Broyden update and Damped Newton's method with different initial conditions. It may be observed that the Damped Newton's method exhibits most* well behaved iterations *among all the variants considered. The iterations remain within physically meaningful ranges of values. It also ensures smoother convergence to the solution (see Figure 4). The Newton's method with initial Jacobian also seems to converge in all the cases with significantly slow rate of convergence and large number of iterations. It may be noted that Newton's method with Broyden's update and base Newton's method can move into infeasible regions of the state space, i.e. -ve concentrations or absurdly large intermediate temperature guesses (see Figure 3 and 4). For the cases demonstrated in the figures, these methods are somehow able to recover and finally converge to the solution. However, such a behavior can potentially lead to divergence of iterations.*

4 Solutions of Nonlinear Algebraic Equations Using Optimization

To solve the set of equation (2) using numerical optimization techniques, we define a scalar objective function

$$\phi(\mathbf{x}) = \frac{1}{2}\left[\mathbf{F}(\mathbf{x})\right]^T \mathbf{F}(\mathbf{x}) = \left[f_1(\mathbf{x})\right]^2 + \left[f_2(\mathbf{x})\right]^2 + ... + \left[f_n(\mathbf{x})\right]^2 \tag{41}$$

and finding solution to equation (2) is formulated as minimization of $\phi(\mathbf{x})$ with respect to $\mathbf{x}$. The necessary condition for unconstrained optimality at $\mathbf{x} = \overline{\mathbf{x}}$ is

$$\frac{\partial \phi(\overline{\mathbf{x}})}{\partial \mathbf{x}} = \left[\frac{\partial \mathbf{F}(\overline{\mathbf{x}})}{\partial \mathbf{x}}\right]^T \mathbf{F}(\overline{\mathbf{x}}) = \overline{\mathbf{0}} \tag{42}$$

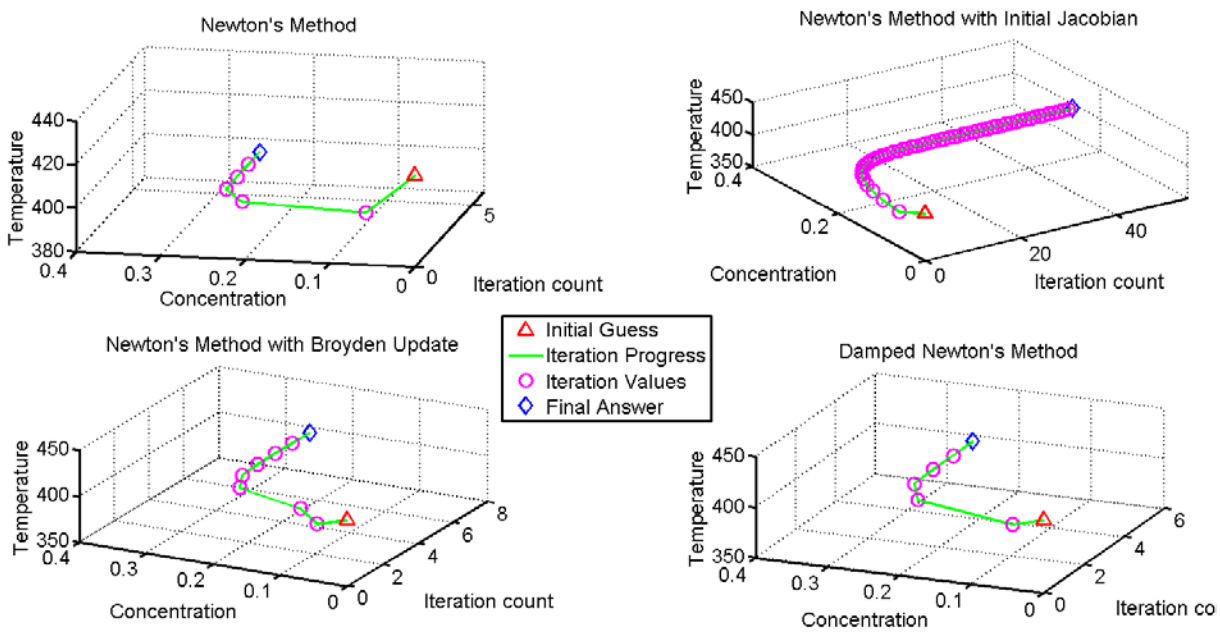

Figure 1: CSTR Example: Progress of iterations of variants of Newton's method - Initial Condition 1

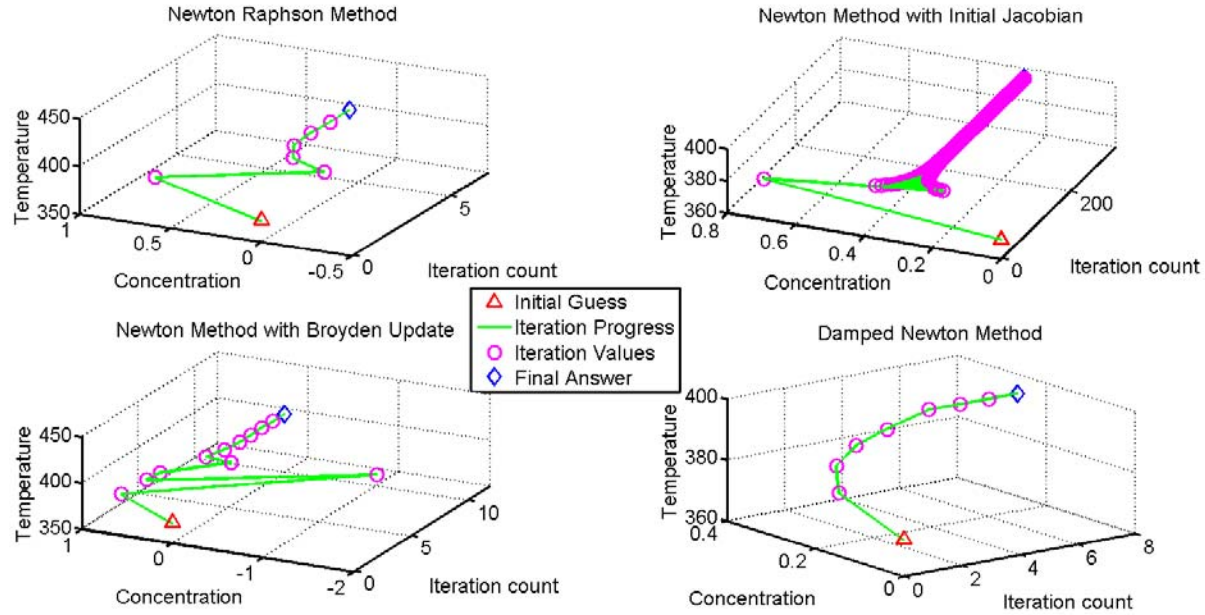

Figure 2: CSTR Example: Progress of iterations of variants of Newton's method - Initial Condition 2

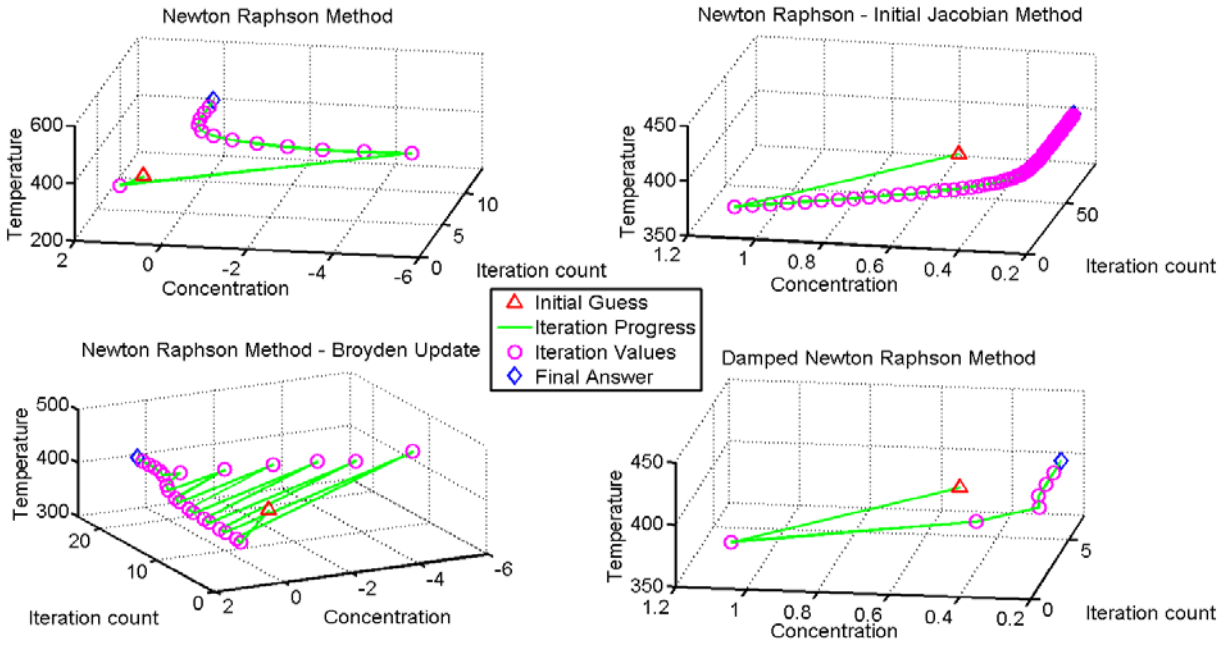

Figure 3: CSTR Example: Progress of iterations of variants of Newton's method - Initial Condition 3

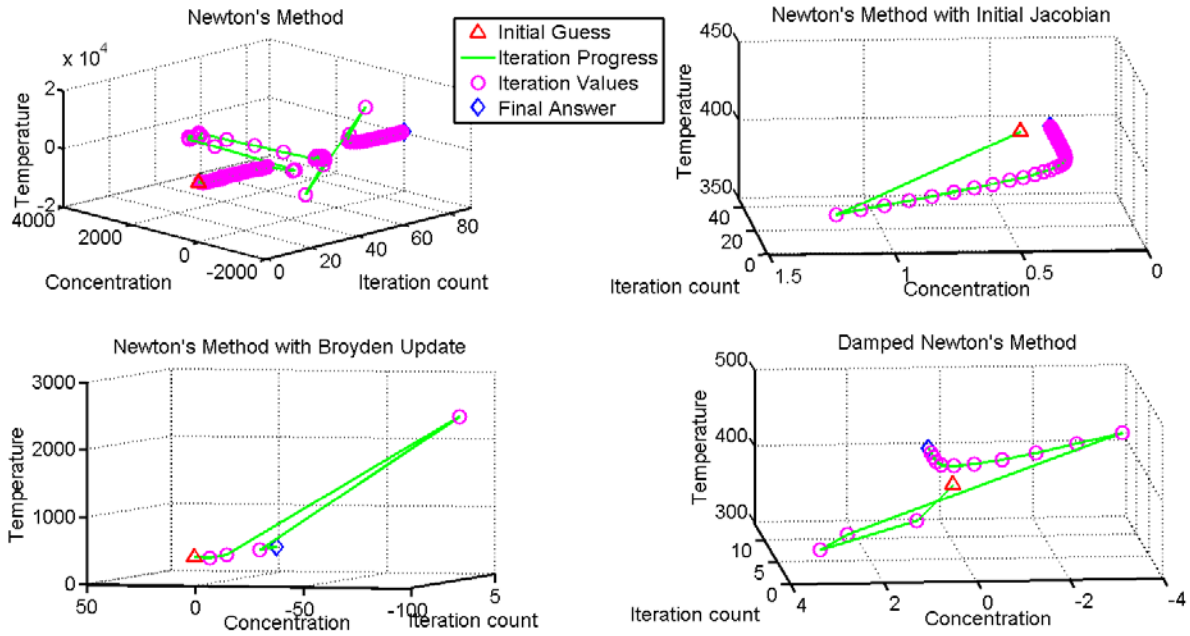

Figure 4: CSTR Example: Progress of iterations of variants of Newton's method - Initial Condition 4

Let us ignore the degenerate case where matrix $\left[\dfrac{\partial \mathbf{F}(\overline{\mathbf{x}})}{\partial \mathbf{x}}\right]$ is singular and vector $\mathbf{F}(\overline{\mathbf{x}})$ belongs to the null space of this matrix. Then, the necessary condition for optimality is satisfied when

$$\mathbf{F}(\overline{\mathbf{x}}) = \overline{\mathbf{0}} \tag{43}$$

which also corresponds to the global minimum of $\phi(\overline{\mathbf{x}})$. The resulting nonlinear optimization problem can be solved using the the conjugate gradient method or Newton's optimization method. Another popular approach to solve the optimization problem is Leverberg-Marquardt method, which is known to work well for solving nonlinear algebraic equations. In this section, we present details of the conjugate gradient method and the Leverberg-Marquardt method, which is based on the Newton's method.

4.1 Conjugate Gradient Method

The conjugate gradient method discussed in the module on Solving $\mathbf{Ax} = \mathbf{b}$ can be used for minimization of $\phi(\mathbf{x})$ with the following modifications

- Negative of the gradient direction is computed as follows

$$\mathbf{g}^{(k)} = -\left[\frac{\partial \mathbf{F}(\mathbf{x}^{(k)})}{\partial \mathbf{x}}\right]^T \mathbf{F}(\mathbf{x}^{(k)}) = -\left[\mathbf{J}^{(k)}\right]^T \mathbf{F}(\mathbf{x}^{(k)})$$

- Conjugate gradient directions are generated with respect to $\mathbf{A} = \mathbf{I}$, i.e.

$$\begin{aligned} \beta_k &= -\frac{\left[\mathbf{g}^{(k)}\right]^T \mathbf{s}^{(k-1)}}{\left[\mathbf{s}^{(k-1)}\right]^T \mathbf{s}^{(k-1)}} \\ \mathbf{s}^{(k)} &= \beta_k \mathbf{s}^{(k-1)} + \mathbf{g}^{(k)} \end{aligned}$$

- Step length is computed by numerically solving one dimensional minimization problem of the form

$$\lambda_k = \min_{\lambda} \phi(\mathbf{x}^{(k)} + \lambda\, \mathbf{s}^{(k)})$$

The gradient based methods have a definite advantage over the Newton's method in the situations where $\left[\mathbf{J}^{(k)}\right]$ become singular. The computation of the search direction does not involve inverse of the Jacobian matrix.

4.2 Newton and Quasi-Newton Methods

The gradient based methods tend to become slow in the neighborhood of the optimum. This difficulty can be alleviated if local Hessian can be used for computing the search direction. The necessary condition for optimization of a scalar function $\phi(\mathbf{x})$ is

$$\nabla\phi(\overline{\mathbf{x}}) = \overline{\mathbf{0}} \tag{44}$$

if $\mathbf{x} = \overline{\mathbf{x}}$ is the optimum. Note that equation (44) defines a system of n equations in n unknowns. If $\nabla\phi(\mathbf{x})$ is continuously differentiable in the neighborhood of $\mathbf{x} = \overline{\mathbf{x}}$, then, using Taylor series expansion, we can express the optimality condition (44) as

$$\nabla\phi(\overline{\mathbf{x}}) = \nabla\phi\left[\mathbf{x}^{(k)} + (\overline{\mathbf{x}} - \mathbf{x}^k)\right] \simeq \nabla\phi[\mathbf{x}^{(k)}] + \left[\nabla^2\phi(\mathbf{x}^{(k)})\right]\Delta\mathbf{x}^{(k)} = \overline{\mathbf{0}} \tag{45}$$

Defining Hessian matrix $\mathbf{H}^{(k)}$ as

$$\mathbf{H}^{(k)} = \left[\nabla^2\phi(\mathbf{x}^{(k)})\right]$$

an iteration scheme can be developed by solving equation (45)

$$\Delta\mathbf{x}^{(k)} = -\left[\mathbf{H}^{(k)}\right]^{-1}\nabla\phi[\mathbf{x}^{(k)}] = -\left[\mathbf{H}^{(k)}\right]^{-1}\left[\mathbf{J}^{(k)}\right]^T\mathbf{F}(\mathbf{x}^{(k)})$$

and generating new guess as follows

$$\mathbf{x}^{(k+1)} = \mathbf{x}^k + \lambda_k\Delta\mathbf{x}^k$$

where λ_k is the step length parameter. In order that $\nabla\mathbf{x}^{(k)}$ is a descent direction it should satisfy the condition

$$\left[\nabla\phi[\mathbf{x}^{(k)}]\right]^T\Delta\mathbf{x}^{(k)} < 0 \tag{46}$$

or

$$\left[\nabla\phi[\mathbf{x}^{(k)}]\right]^T\left[\mathbf{H}^{(k)}\right]^{-1}\nabla\phi[\mathbf{x}^{(k)}] > 0 \tag{47}$$

i.e. in order that $\Delta\mathbf{x}^{(k)}$ is a descent direction, Hessian $\mathbf{H}^{(k)}$ should be a positive definite matrix. This method has good convergence but demands large amount of computations i.e. solving a system of linear equations and evaluation of Hessian at each step. The steps involved in the Newton's unconstrained optimization scheme are summarized in Table 2.

Major disadvantage of Newtons method is the need to compute the Hessian of each iteration. The quasi-Newton methods overcome this difficulty by constructing an approximate Hessian from the gradient information available at the successive iteration. One of the widely used algorithm is of this type is **variable metric** (or Devidon - Fletcher- Powell) method. Let us define matrix

$$\mathbf{L} = \mathbf{H}^{-1}$$

Table 2: Newton's Method for Unconstrained Optimization

INITIALIZE: $\mathbf{x}^{(0)}, \varepsilon, k_{\max}, \lambda^{(0)}$
$k = 0$
$\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon) \;\; AND \;\; (k < k_{\max})]$
$\mathbf{H}^{(k)}\mathbf{s}^{(k)} = -\nabla\phi^{(k)}$
$\lambda_k = \min_{\lambda} \phi\left(\mathbf{x}^{(k)} - \lambda \mathbf{s}^{(k)}\right)$
$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \lambda_k \, \mathbf{s}^{(k)}$
$\delta = \left\|\nabla\phi\left[\mathbf{x}^{(k+1)}\right]\right\|_2$
END WHILE

Then, matrix $\mathbf{L}^{(k)}$ is iteratively computed as follows [11]

$$\mathbf{L}^{(k+1)} = \mathbf{L}^{(k)} + \mathbf{M}^{(k)} - \mathbf{N}^{(k)} \tag{48}$$

$$\mathbf{q}^{(k)} = \nabla\phi^{(k+1)} - \nabla\phi^{(k)} \tag{49}$$

$$\mathbf{M}^{(k)} = \left(\frac{\lambda_k}{\left[\Delta\mathbf{x}^{(k)}\right]^T \mathbf{q}^{(k)}}\right)\left[\left[\Delta\mathbf{x}^{(k)}\right]\left[\Delta\mathbf{x}^{(k)}\right]^T\right] \tag{50}$$

$$\mathbf{N}^{k} = \left(\frac{1}{\left[\mathbf{q}^{(k)}\right]^T \mathbf{L}^{(k)}\mathbf{q}^{(k)}}\right)\left[\mathbf{L}^{(k)}\mathbf{q}^{(k)}\right]\left[\mathbf{L}^{(k)}\mathbf{q}^{(k)}\right]^T \tag{51}$$

starting from some initial guess, usually a positive definite matrix. Typical choice is $\mathbf{L}^{(0)} = \mathbf{I}$.

4.3 Leverberg-Marquardt Method

It is known from the experience that the steepest descent method produces large reduction in objective function when $\mathbf{x}^{(0)}$ is far away from, $\overline{\mathbf{x}}$,i.e. the optimal solution. However, the steepest descent method becomes notoriously slow near the optimum. On the other hand, Newton's method generates ideal search directions near the optimum. The Leverberg-Marquardt approach combines advantages of the gradient method and the Newton's method by starting with gradient search initially and switching to Newton's method as iterations progress. This is achieved as follows

$$\begin{aligned}
\mathbf{g}^{(k)} &= -\nabla\phi[\mathbf{x}^{(k)}] = -\left[\mathbf{J}^{(k)}\right]^T \mathbf{F}(\mathbf{x}^{(k)}) \\
\mathbf{s}^{(k)} &= -\left[\mathbf{H}^{(k)} + \beta_k \mathbf{I}\right]^{-1}\left[\mathbf{J}^{(k)}\right]^T \mathbf{F}(\mathbf{x}^{(k)}) \\
\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \mathbf{s}^{(k)}
\end{aligned}$$

Table 3: Table Caption

INITIALIZE: $\mathbf{x}^{(0)}, \varepsilon, k_{\max}, \beta^{(0)}$
$k = 0$
$\delta = 100 * \varepsilon$
WHILE $[(\delta > \varepsilon) \;\; AND \;\; (k < k_{\max})]$
    STEP 1 : Compute $\mathbf{H}^{(k)}$ and $\nabla\phi[\mathbf{x}^{(k)}]$
    STEP 2 : $Solve \;\; \left[\mathbf{H}^{(k)} + \beta_k I\right] \mathbf{s}^{(k)} = -\nabla\phi[\mathbf{x}^{(k)}]$
    IF $\left(\phi[\mathbf{x}^{(k+1)}] < \phi[\mathbf{x}^{(k)}]\right)$
        $\beta^{(k+1)} = \frac{1}{2}\beta^{(k)}$
        $\delta = \left\|\nabla\phi[\mathbf{x}^{(k+1)}]\right\|$
        $k = k + 1$
    ELSE
        $\beta^{(k)} = 2\beta^{(k)}$
    GO TO STEP 2
END WHILE

Here β is used to set the search direction. To begin the search, a large value of $\lambda(\widetilde{=}10^4)$is selected so that

$$\left[\mathbf{H}^{(0)} + \beta_0\mathbf{I}\right] \;\widetilde{=}\; [\beta_0\mathbf{I}]$$

Thus, for sufficiently large β, the search direction $\mathbf{s}^{(k)}$is in the negative of the gradient direction i.e. $-\nabla\phi^{(0)}/\beta_0$. On the other hand, when $\beta_k \to 0$ we have $\mathbf{s}^{(k)}$ goes from steepest descent to Newtons method. With intermediate values of β, we get a step that is intermediate between the Newton's step and gradient step. The steps involved in the implementation of the basic version of Leverberg Marquardt algorithm are summarized in Table 3. The main advantage of Leverberg-Marquardt method is simplicity and excellent convergence in the neighborhood of the solution. However, it is necessary to compute Hessian matrix, $\mathbf{H}^{(k)}$and set of linear equations has to be solved many times at each iteration before fixing $\mathbf{s}^{(k)}$. Also, when β_k is large initially, the step size $\mathbf{s}^{(k)} = -\nabla\phi^{(k)}/\beta_k$ is small and this can result in slow progress of the iterations.

5 Condition Number of Nonlinear Set of Equations [7]

Concept of condition number can be easily extended to analyze numerical conditioning of set on nonlinear algebraic equations. Consider nonlinear algebraic equations of the form

$$\mathbf{F}(\mathbf{x}, \mathbf{u}) = \overline{0} \ ; \quad \mathbf{x} \in R^n \, , \quad \mathbf{u} \in R^m$$

where $\mathbf{F}$ is $n \times 1$ function vector and $\mathbf{u}$ is a set of known parameters or independent variables on which the solution depends. The condition number measures the worst possible effect on the solution $\mathbf{x}$ caused by small perturbation in $\mathbf{u}$. Let $\delta\mathbf{x}$ represent the perturbation in the solution caused by perturbation $\delta\mathbf{u}$,i.e.

$$\mathbf{F}(\mathbf{x}+\delta\mathbf{x}, \mathbf{u}+\delta\mathbf{u}) = \overline{\mathbf{0}}$$

Then the condition number of the system of equations is defined as

$$\begin{aligned} C(\mathbf{x}) &= \sup_{\delta\mathbf{u}} \frac{||\delta\mathbf{x}||/||\mathbf{x}||}{||\delta\mathbf{u}||/||\mathbf{u}||} \\ &\Rightarrow \frac{||\delta\mathbf{x}||/||\mathbf{x}||}{||\delta\mathbf{u}||/||\mathbf{u}||} \leq C(\mathbf{x}) \end{aligned}$$

If the solution does not depend continuously on $\mathbf{u}$, then the $C(\mathbf{x})$ becomes infinity and such systems are called as (numerically) unstable systems. Systems with large condition numbers are more susceptible to computational errors.

Example 3 *[7] Consider equation*

$$x - e^u = 0$$

Then,

$$\begin{aligned} \delta x/x &= \frac{e^{u+\delta u} - e^u}{e^u} = e^{\delta u} - 1 \\ C(x) &= \sup_{\delta u} \left| u \frac{e^{\delta u} - 1}{\delta u} \right| \end{aligned}$$

For small δu, we have $e^{\delta u} = 1 + \delta u$ and

$$C(x) = |u|$$

6 Existence of Solutions and Convergence of Iterative Methods [12]

If we critically view the methods presented for solving equation (2), it is clear that this problem, in general, cannot solved in its original form. To generate a numerical approximation to the solution of equation (2), this equation is further transformed to formulate an iteration sequence as follows

$$\mathbf{x}^{(k+1)} = \mathbf{G}\left[\mathbf{x}^{(k)}\right] \quad ; \quad k = 0, 1, 2, ... \tag{52}$$

where $\left\{\mathbf{x}^{(k)} : k = 0, 1, 2, ...\right\}$ is sequence of vectors in vector space under consideration. The iteration equation is formulated in such a way that the solution $\mathbf{x}^*$ of equation **(52)** also solves equation **(2),** i.e.

$$\mathbf{x}^* = \mathbf{G}\left[\mathbf{x}^*\right] \Rightarrow \mathbf{F}(\mathbf{x}^*) = \overline{\mathbf{0}}$$

For example, in the Newton's method, we have

$$\mathbf{G}\left[\mathbf{x}\right] \leftrightarrow \mathbf{F}(\mathbf{x}) - \left[\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right]^{-1} \mathbf{F}(\mathbf{x})$$

Thus, we concentrate on the existence and (local) uniqueness of solutions of $\mathbf{x}^* = \mathbf{G}\left[\mathbf{x}^*\right]$ rather than that of $\mathbf{F}(\mathbf{x})$.

Contraction mapping theorem develops sufficient conditions for convergence of general nonlinear iterative equation (5). Consider general nonlinear iteration equation of the form

$$\mathbf{x}^{(k+1)} = \mathbf{G}(\mathbf{x}^{(k)}) \tag{53}$$

which defines a mapping from a Banach space $\mathbf{X}$ into itself, i.e. $\mathbf{G}(.) : X \rightarrow X$.

Definition 4 ***(Contraction Mapping):*** *An operator* $\mathbf{G} : X \rightarrow X$ *given by equation (5), mapping a Banach space* X *into itself, is called a contraction mapping of closed ball* $U(\mathbf{x}^{(0)}, r) = \left\{\mathbf{x} \in X : \left\|\mathbf{x} - \mathbf{x}^{(0)}\right\| \leq r\right\}$, *if there exists a real number* θ $(0 \leq \theta < 1)$ *such that*

$$\left\|\mathbf{G}\left(\mathbf{x}^{(1)}\right) - \mathbf{G}\left(\mathbf{x}^{(2)}\right)\right\| \leq \theta \left\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\right\|$$

for all $\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in U(\mathbf{x}^{(0)}, r)$. *The quantity* θ *is called contraction constant of* $\mathbf{G}$ *on* $U(\mathbf{x}^{(0)}, r)$.

In other words, a function $\mathbf{x} = \mathbf{G}(\mathbf{x})$ is said to be a contraction mapping with respect to a norm $\|.\|$ on a closed region $\mathbf{S}$ if

Definition 5

- $\mathbf{x} \in \mathbf{S}$ *implies that* $\mathbf{G}(\mathbf{x}) \in \mathbf{S}$, *i.e.* $\mathbf{G}$ *maps* $\mathbf{S}$ *onto itself*
- $\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\widetilde{\mathbf{x}})\| \leq \theta \|\mathbf{x} - \widetilde{\mathbf{x}}\|$ *with* $0 \leq \theta < 1$ *for all* $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathbf{S}$

When the map $\mathbf{G}(.)$ is differentiable, an exact characterization of the contraction property can be developed.

Lemma 6 *Let the operator $\mathbf{G}(.)$ on a Banach space X be differentiable in $U(\mathbf{x}^{(0)}, r)$. Operator $\mathbf{G}(.)$ is a contraction of $U(\mathbf{x}^{(0)}, r)$ if and only if*

$$\left\| \frac{\partial \mathbf{G}}{\partial \mathbf{x}} \right\| \leq \theta < 1 \quad \textit{for every } \mathbf{x} \in U(\mathbf{x}^{(0)}, r)$$

where $\|.\|$ is any induced operator norm.

The contraction mapping theorem is stated next. Here, $\mathbf{x}^{(0)}$ refers to the initial guess vector in the iteration process given by equation (53).

Theorem 7 *[12, 9] If $\mathbf{G}(.)$ maps $U(\mathbf{x}^{(0)}, r)$ into itself and $\mathbf{G}(.)$ is a contraction mapping on the set with contraction constant θ, for*

$$\begin{aligned} r &\geq r_0 \\ r_0 &= \frac{1}{1-\theta} \left\| \mathbf{G}\left[\mathbf{x}^{(0)}\right] - \mathbf{x}^{(0)} \right\| \end{aligned}$$

then:

1. *$\mathbf{G}(.)$ has a fixed point $\mathbf{x}^*$ in $U(\mathbf{x}^{(0)}, r_0)$ such that $\mathbf{x}^* = \mathbf{G}(\mathbf{x}^*)$*
2. *$\mathbf{x}^*$ is unique in $U(\mathbf{x}^{(0)}, r)$*
3. *The sequence $\mathbf{x}^{(k)}$ generated by equation $\mathbf{x}^{(k+1)} = \mathbf{G}\left[\mathbf{x}^{(k)}\right]$ converges to $\mathbf{x}^*$ with*

$$\left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \leq \theta^k \left\| \mathbf{x}^{(0)} - \mathbf{x}^* \right\|$$

4. *Furthermore, the sequence $\overline{\mathbf{x}}^{(k)}$ generated by equation*

$$\overline{\mathbf{x}}^{(k+1)} = \mathbf{G}\left[\overline{\mathbf{x}}^{(k)}\right] \quad \textit{starting from any initial guess } \overline{\mathbf{x}}^{(0)} \in U(\mathbf{x}^{(0)}, r_0)$$

converges to $\mathbf{x}^*$ with

$$\left\| \overline{\mathbf{x}}^{(k)} - \mathbf{x}^* \right\| \leq \theta^k \left\| \overline{\mathbf{x}}^{(0)} - \mathbf{x}^* \right\|$$

The proof of this theorem can be found in Rall [12] and Linz [9].

Example 8 *[9] Consider simultaneous nonlinear equations*

$$z + \frac{1}{4}y^2 = \frac{1}{16} \tag{54}$$

$$\frac{1}{3}\sin(z) + y = \frac{1}{2} \tag{55}$$

We can form an iteration sequence

$$z^{(k+1)} = \frac{1}{16} - \frac{1}{4}\left(y^{(k)}\right)^2 \tag{56}$$

$$y^{(k+1)} = \frac{1}{2} - \frac{1}{3}\sin(z^{(k)}) \tag{57}$$

Using $\infty-$norm In the unit ball $U(\mathbf{x}^{(0)} = \overline{0}, 1)$ in the neighborhood of origin, we have

$$\left\|\mathbf{G}\left(\mathbf{x}^{(i)}\right) - \mathbf{G}\left(\mathbf{x}^{(j)}\right)\right\|_\infty = \max\left(\frac{1}{4}\left|\left(y^{(i)}\right)^2 - \left(y^{(j)}\right)^2\right|, \frac{1}{3}\left|\sin(x^{(i)}) - \sin(x^{(j)})\right|\right) \tag{58}$$

$$\leq \max\left(\frac{1}{4}\left|y^{(i)} - y^{(j)}\right|, \frac{1}{3}\left|x^{(i)} - x^{(j)}\right|\right) \tag{59}$$

$$\leq \frac{1}{2}\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right\|_\infty \tag{60}$$

Thus, $\mathbf{G}(.)$ is a contraction map with $\theta = 1/2$ and the system of equation has a unique solution in the unit ball $U(\mathbf{x}^{(0)} = \overline{0}, 1)$ i.e. $-1 \leq x \leq 1$ and $-1 \leq y \leq 1$. The iteration sequence converges to the solution.

Example 9 *[9] Consider system*

$$x - 2y^2 = -1 \tag{61}$$

$$3x^2 - y = 2 \tag{62}$$

which has a solution (1,1). The iterative method

$$x^{(k+1)} = 2\left(y^{(k)}\right)^2 - 1 \tag{63}$$

$$y^{(k+1)} = 3\left(x^{(k)}\right)^2 - 2 \tag{64}$$

is not a contraction mapping near (1,1) and the iterations do not converge even if we start from a value close to the solution. On the other hand, the rearrangement

$$x(k+1) = \sqrt{(y^{(k)} + 2)/3} \tag{65}$$

$$y^{(k+1)} = \sqrt{(x^{(k)} + 1)/2} \tag{66}$$

is a contraction mapping and solution converges if the starting guess is close to the solution.

6.1 Convergence of Successive Substitution Schemes [4]

Either by successive substitution approach or Newton's method, we generate an iteration sequence

$$\mathbf{x}^{(k+1)} = \mathbf{G}(\mathbf{x}^{(k)}) \tag{67}$$

which has a fixed point

$$\mathbf{x}^* = \mathbf{G}(\mathbf{x}^*) \tag{68}$$

at solution of $\mathbf{F}(\mathbf{x}^*) = \overline{0}$. Defining error

$$\mathbf{e}^{(k+1)} = \mathbf{x}^{(k+1)} - \mathbf{x}^* = \mathbf{G}(\mathbf{x}^{(k)}) - \mathbf{G}(\mathbf{x}^*) \tag{69}$$

and using Taylor series expansion, we can write

$$\begin{aligned} \mathbf{G}(\mathbf{x}^*) &= \mathbf{G}[\mathbf{x}^{(k)} - (\mathbf{x}^{(k)} - \mathbf{x}^*)] & (70) \\ &\simeq \mathbf{G}(\mathbf{x}^{(k)}) - \left[\frac{\partial \mathbf{G}}{\partial \mathbf{x}}\right]_{x=x^{(k)}} (\mathbf{x}^{(k)} - \mathbf{x}^*) & (71) \end{aligned}$$

Substituting in (69)

$$\mathbf{e}^{(k+1)} = \left[\frac{\partial \mathbf{G}}{\partial \mathbf{x}}\right]_{x=x^{(k)}} \mathbf{e}^{(k)} \tag{72}$$

where

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}^*$$

and using definition of induced matrix norm, we can write

$$\frac{||\mathbf{e}^{(k+1)}||}{||\mathbf{e}^{(k)}||} < \left\| \left[\frac{\partial \mathbf{G}}{\partial \mathbf{x}}\right]_{x=x^{(k)}} \right\| \tag{73}$$

It is easy to see that the successive errors will reduce in magnitude if the following condition is satisfied at each iteration i.e.

$$\left\| \left[\frac{\partial \mathbf{G}}{\partial \mathbf{x}}\right]_{x=x^{(k)}} \right\| < 1 \text{ for } k = 1, 2, ... \tag{74}$$

Applying *contraction mapping theorem* (refer to Appendix A for details), a sufficient condition for convergence of iterations in the neighborhood $\mathbf{x}^*$can be stated as

$$\left\| \left[\frac{\partial \mathbf{G}}{\partial \mathbf{x}}\right] \right\|_1 \leq \theta_1 < 1$$

or

$$\left\| \left[\frac{\partial \mathbf{G}}{\partial \mathbf{x}}\right] \right\|_\infty \leq \theta_\infty < 1$$

Note that this is only a sufficient conditions. If the condition is not satisfied, then the iteration scheme may or may not converge. Also, note that introduction of step length parameter $\lambda^{(k)}$ in Newton's step as

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda^{(k)} \Delta \mathbf{x}^{(k)} \tag{75}$$

such that $\left\|\mathbf{F}^{(k+1)}\right\| < \left\|\mathbf{F}^{(k)})\right\|$ ensures that $\mathbf{G}(\mathbf{x})$ is a contraction map and ensures convergence.

Consider equation of type $\mathbf{x} = \mathbf{G}(\mathbf{x})$ where $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{G}(\mathbf{x})$ represents a function vector of type

$$\mathbf{G}(\mathbf{x}) = \left[\begin{array}{cccc} \mathbf{g}_1(\mathbf{x}) & \mathbf{g}_2(\mathbf{x}) & .... & \mathbf{g}_n(\mathbf{x}) \end{array} \right]^T$$

Let us suppose that $\partial G/\partial x_i$ are continuous in some region **S.** Let us define a matrix $\mathbf{J}$ such that (i.j)'th element of $\mathbf{J}$ is defined as follows

$$\mathbf{J}_{i,j} = \begin{array}{c} \sup \\ x \in S \end{array} \left| \frac{\partial \mathbf{g}_i(\mathbf{x})}{\partial x_j} \right|$$

Then, it can be shown that

$$\left\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\widetilde{\mathbf{x}})\right\|_p \leq \left\|\mathbf{J}\right\|_p \left\|\mathbf{x} - \widetilde{\mathbf{x}}\right\|_p$$

and if $\left\|\mathbf{J}\right\|_p < 1$ holds in the region of interest, then $\mathbf{G}(.)$ is a contraction mapping with $L = \left\|\mathbf{J}\right\|_p$. Also, note that, for 2 norm, the following inequality can be used

$$\begin{aligned} \left\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\widetilde{\mathbf{x}})\right\|_2 &\leq \left\|\mathbf{J}\right\|_{FRO} \left\|\mathbf{x} - \widetilde{\mathbf{x}}\right\|_2 \\ \left\|\mathbf{J}\right\|_{FRO} &= \left[\sum_{i,j} \left(\mathbf{J}_{i,j}\right)^2 \right]^{1/2} \end{aligned}$$

where $\left\|\mathbf{J}\right\|_{FRO}$ is called as *Frobenius norm* of matrix $\mathbf{J}$.

Example 10 *Consider the following system of equations*

$$\begin{aligned} x_1 &= \frac{1}{12}(-1 + \sin(x_2) + \sin(x_3)) \\ x_2 &= \frac{1}{3}(x_1 - \sin(x_2) + \sin(x_3)) \\ x_3 &= \frac{1}{12}(1 - \sin(x_1) + x_2) \end{aligned}$$

which is of the form, $\mathbf{x} = \mathbf{G}(\mathbf{x})$, *in the closed and bounded region* $\mathbf{S}$ *defined as* $-1 \leq x_1, x_2, x_3 \leq 1$. *Then, the matrix* $\mathbf{J}$ *can be shown to be*

$$\mathbf{J} = \frac{1}{12} \left[\begin{array}{ccc} 0 & 1 & 1 \\ 4 & 4 & 4 \\ 1 & 1 & 0 \end{array} \right]$$

Further, $\|\mathbf{J}\|_1 = \frac{1}{2}, \quad \|\mathbf{J}\|_\infty = 1$ *and* $\|\mathbf{J}\|_{FRO} = \sqrt{13}/6$. *The* $\mathbf{G}(.)$ *is a contraction mapping for norms* $\|.\|_1$ *and* $\|.\|_2$ *in region* $\mathbf{S}$. *From contraction mapping theorem, it follows that the system of equations* $\mathbf{x} = \mathbf{G}(\mathbf{x})$ *has a unique solution in region* $\mathbf{S}$.

6.2 Convergence of Newton's Method

Sufficient conditions for the convergence of Newton's method have been established by Kantorovic' Theorem.

Theorem 11 *(Kantorovic'): Consider equation* $\mathbf{F}(\mathbf{x}) = \overline{\mathbf{0}}$ *where operator* $\mathbf{F} : R^n \rightarrow R^n$ *is twice differentiable and the following conditions hold*

- *There is a* $\mathbf{x}^{(0)} \in R^n$ *such that* $\left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right]^{-1}$ *exists with*

$$\left\| \left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right]^{-1} \right\| = \beta_0 \text{ and } \left\| \left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right]^{-1} \mathbf{F}\left(\mathbf{x}^{(0)}\right) \right\| \leq \eta_0$$

- $\left\| \left[\frac{\partial^2 \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}^2}\right] \right\| \leq \kappa$ *in a closed ball* $U(\mathbf{x}^{(0)}, 2\eta_0)$
- $h_0 = \beta_0 \eta_0 \kappa < 1/2$

Then the sequence

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(k)}\right)}{\partial \mathbf{x}}\right]^{-1} \mathbf{F}\left(\mathbf{x}^{(k)}\right) \tag{76}$$

exists for all $k \geq 0$ *and converges to the solution of* $\mathbf{F}(\mathbf{x}) = \overline{\mathbf{0}}$, *which exists and is unique in* $U(\mathbf{x}^{(0)}, 2\eta_0)$.

Proof. Proof of this theorem can be found in Demidovich[6]. ■

Economou [10] has given an interesting interpretation of this theorem. Using multivariable Taylor series expansion of $\left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right]$ in the neighborhood of $\mathbf{x}^{(0)}$, we have

$$\begin{aligned} \left\| \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}} - \frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}} \right\| &\leq \sup_{0 \leq \lambda \leq 1} \left\| \frac{\partial^2 \mathbf{F}\left(\lambda \mathbf{x} + (1-\lambda)\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}^2} \right\| \left\| \mathbf{x} - \mathbf{x}^{(0)} \right\| \\ &\leq \kappa \left\| \mathbf{x} - \mathbf{x}^{(0)} \right\| \end{aligned}$$

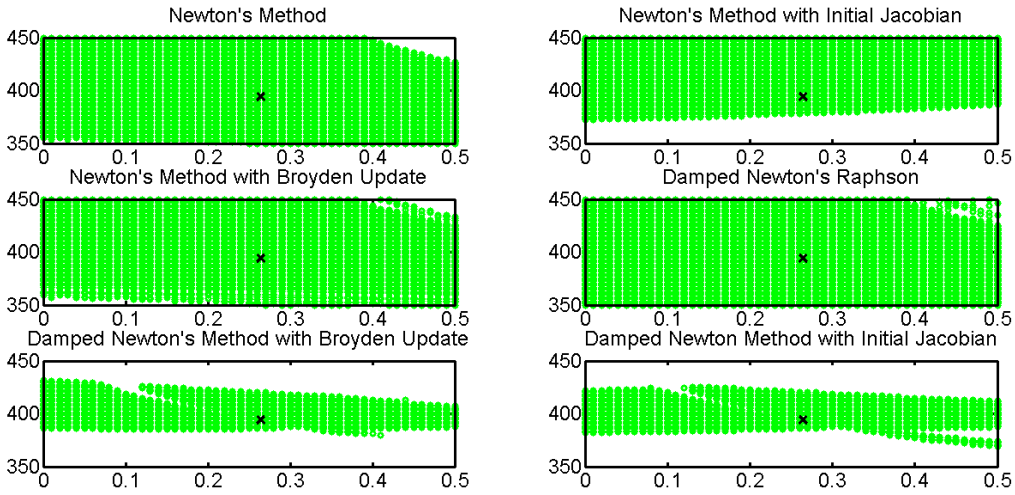

Figure 5: CSTR Example: Basins of convergence for different variants of Newton's method. Green dots represent initial conditions that lead to convergence of iterations while the black cross represents the solution.

Multiplying by $\left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right]^{-1}$ on both the sides, we have

$$\left\|\left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right]^{-1}\right\|\left\|\frac{\partial \mathbf{F}\left(\mathbf{x}\right)}{\partial \mathbf{x}}-\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right\| \leq \left\|\left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right]^{-1}\right\| \kappa \left\|\mathbf{x}-\mathbf{x}^{(0)}\right\|$$

When conditions of the Kantorovic' theorem are satisfied, it follows that

$$\left\|\left[\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right]^{-1}\right\|\left\|\frac{\partial \mathbf{F}\left(\mathbf{x}\right)}{\partial \mathbf{x}}-\frac{\partial \mathbf{F}\left(\mathbf{x}^{(0)}\right)}{\partial \mathbf{x}}\right\| \leq 2\beta_0\eta_0\kappa < 1$$

The term on the L.H.S. of the inequality represents the magnitude of relative change in the Jacobian of operator $\mathbf{F}(.)$ in ball $U(\mathbf{x}^{(0)}, 2\eta_0)$. The Kantorovic' theorem asserts that Newton's method converges if the relative change of the Jacobian in ball $U(\mathbf{x}^{(0)}, 2\eta_0)$ is less than 100%.

Example 12 ***Basins of Attraction for CSTR Example:*** *The CSTR system described in Example 2 was studied for understanding convergence behavior of variants of Newton's method. Iterations were started from various initial conditions in a box around the steady state solution and progress of iterations towards the solutions was recorded. Figure 5 compares sets of initial conditions starting from which the respective methods converge to the solution. In each box, green dots represent initial conditions that lead to convergence of iterations. As evident from this figure, Newton's method with the initial Jacobian has a smaller basin of convergence. Damped Newton's method appears to have largest basin of convergence.*

7 Summary

In these lecture notes, we have developed methods for efficiently solving nonlinear algebraic equations. These methods can be classified as derivative free and derivative based methods. Issues related to existence and uniqueness of the solutions and convergence of the iterative schemes have also been discussed briefly.

References

[1] Bazara, M.S., Sherali, H. D., Shetty, C. M., Nonlinear Programming, John Wiley, 1979.

[2] Biegler, L. T., I. E. Grossman, Westerberg, A. W., Systematic Method of Chemical Process Design, Prentice-Hall International, 1997.

[3] Gupta, S. K.; Numerical Methods for Engineers. Wiley Eastern, New Delhi, 1995.

[4] Gourdin, A. and M Boumhrat; Applied Numerical Methods. Prentice Hall India, New Delhi.

[5] Strang, G.; Linear Algebra and Its Applications. Harcourt Brace Jevanovich College Publisher, New York, 1988.

[6] Demidovich, B. P. and I. A. Maron; Computational Mathematics. Mir Publishers, Moskow, 1976.

[7] Atkinson, K. E.; An Introduction to Numerical Analysis, John Wiley, 2001.

[8] Linfield, G. and J. Penny; Numerical Methods Using Matlab, Prentice Hall, 1999.

[9] Linz, P.; Theoretical Numerical Analysis, Dover, New York, 1979.

[10] Economou, C. G. An operator Theory Approch to Nonlinear Controller Design. *Ph.D. Dissertation.* California Institute of Technology, 1985.

[11] Rao, S. S., Optimization: Theory and Applications, Wiley Eastern, New Delhi, 1978.

[12] Rall, L. B.; Computational Solutions of Nonlinear Operator Equations. John Wiley, New York, 1969.