Chapter 6

System of Nonlinear Equations

Finding the zeros of a given function f, i.e., finding an argument x for which

$$f(x) = 0 \tag{6.1}$$

where $f : R^n \to R^n$, is a classical problem arising from many areas of applications. Except in linear problems, root-finding invariably proceeds by iteration — starting from some approximate trial solution, a useful algorithm will improve the solution until some predetermined convergence criterion is satisfied.

Unlike most of the iterative methods used for linear systems, having a good first-guess for the solution of a nonlinear system usually is crucial in determining the success of an iterative process. Such methods, among which the Newton-Raphson method is the most celebrated, are called local methods.

Thus far, general–purposed global methods are not available. For special classes of problem, such as solving systems of polynomials, important progress has recently been made. The homotopy method, having connections with differential equations, is one such approach.

Another difficulty often associated with solving nonlinear equations is the detection of existence of one or multiple solutions. A nonlinear set of equations may have no (real) solutions at all. Contrariwise, it may have more than one solution. Ideally, one should resort to some other means, such as the degree theory, to determine that theoretically a nonlinear equation does have a solution before numerically finding the approximate solution. Applying a numerical method blindfold has the danger of being misled to a wrong answer even though the method behaves nicely.

6.1 The Newton-Raphson Method

An iterative scheme for solving the system (6.1) generally takes the form

$$x_{k+1} = \rho(x_k), \ k = 0, 1, \ldots \tag{6.2}$$

where $\rho : R^n \to R^n$ is a fixed function and $x_0 \in R^n$ is a given starting value. It is hoped that the sequence $\{x_k\}$ will converge to a point ξ as $i \to \infty$. If this is so, then ξ must be a fixed point of ρ. It is desirable that $f(\xi) = 0$. It is also important to know how quickly the sequence $\{x_k\}$ converges.

Suppose the system (6.1) has a solution at ξ and suppose that f is sufficiently smooth. Let x_k be an approximation to ξ. Considering the Taylor series expansion of f about x_k, we have

$$f(x) = f(x_k) + f'(x_k)(x - x_k) + 0(\|x - x_k\|^2) \tag{6.3}$$

where $f' \in R^{n\times n}$ denotes the Jacobian matrix of f. If, in particular, we take $x \in \xi$, then

$$0 \approx f(x_k) + f'(x_k)(\xi - x_k). \tag{6.4}$$

From (6.4), we are motivated to think that the quantity

$$x_{k+1} := x_k + s_k \tag{6.5}$$

where s_k satisfies the linear system

$$f'(x_k)x_k = -f(x_k) \tag{6.6}$$

should a better approximation to ξ then x_k. In other words, if f' is nonsingular, then we have derived a special iterative scheme (6.2) with

$$\rho(x) := x - f'(x)^{-1}f(x). \tag{6.7}$$

This is the well-known Newton-Raphson method.

There are many ways to define a sensible iteration function ρ other than (6.7). Besides numerous research papers, a classical reference on this topic is the book "Iterative Solution of Nonlinear Equations in Several Variables" by Ortega and Rheinboldt. An earlier book is "Iterative Methods for the Solution of Equations" by Traub.

Definition 6.1.1 *Let $\rho : R^n \to R^n$ be an iteration function. Let ξ be a fixed point of ρ. The iterative scheme (6.1.1) defined by ρ is said to be a method of order p if for all initial point x_0 in a neighborhood $N(\xi)$, the generated sequence $\{x_k\}$ satisfies*

$$\|x_{k+1} - \xi\| \le C\|x_k - \xi\|^p \tag{6.8}$$

for a certain constant C.

The Newton method is best knwon for its quadratic convergence. Toward this, we first prove a useful lemma.

Lemma 6.1.1 *Let $f : R^n \to R^n$ be continuously differentiable in an open convex set $D \subset R^n$. Suppose a constant γ exists such that $\|f'(x)-f'(y)\| \le \gamma\|x-y\|$ for all $x, y \in D$. Then $\|f(x) - f(y) - f'(y)(x - y)\| \le \frac{\gamma}{2}\|x - y\|^2$.*

(pf): By the line integration, $f(x) - f(y) = \int_0^1 f'(y + t(x-y))(x-y)dt$. So

$$f(x) - f(y) - f'(y)(x-y) = \int_0^1 [f'(y + t(x-y)) - f'(y)](x-y)dt.$$

It follows that

$$\begin{aligned} & \|f(x) - f(y) - f'(y)(x-y)\| \\ \leq & \int_0^1 \|f'(y + t(x-y)) - f'(y)\|\|x-y\|dt \\ \leq & \int_0^1 \gamma t\|x-y\|^2 dt = \frac{\gamma}{2}\|x-y\|^2. \end{aligned}$$

Theorem 6.1.1 *Let $f : R^n \to R^n$ be continuously differentiable in an open convex set $D \subset R^n$. Assume that there exist $\xi \in D$ and $\beta, \gamma > 0$ such that*

(1) $f(\xi) = 0$,

(2) $f'(\xi)^{-1}$ exists,

(3) $\|f'(\xi)^{-1}\| \leq \beta$, and

(4) $\|f'(x) - f'(y)\| \leq \gamma\|x-y\|$ for x, y in a neighborhood of ξ.

Then there exists $\epsilon > 0$ such that for every $x_0 \in N(\xi, \epsilon)$, the sequence $\{x_k\}$ define by (6.5) and (6.6) is well defined, converges to ξ and satisfies

$$\|x_{k+1} - \xi\| \leq \beta\gamma\|x_k - \xi\|^2. \tag{6.9}$$

(pf): By continuity of f', choose $\epsilon \leq \min\{\gamma, \frac{1}{2\beta\gamma}\}$ so that $f'(x)$ is nonsingular for all $x \in N(\xi, \epsilon)$. For $k = 0$, we have already known $\|x_0 - \xi\| < \epsilon$. So

$$\begin{aligned} \|f'(\xi)^{-1}(f'(x_0) - f'(\xi))\| & \leq \|f'(\xi)^{-1}\|\|f'(x_0) - f'(\xi)\| \\ & \leq \beta\gamma\|x_0 - \xi\| \leq \frac{1}{2}. \end{aligned}$$

By the Banach lemma (Theorem 2.1.1 in my note .4.14),

$$\begin{aligned} & \|f'(x_0)^{-1}\| = \|[f'(\xi) + (f'(x_0) - f'(\xi))]^{-1}\| \\ \leq & \frac{\|f'(\xi)^{-1}\|}{1 - \|f'(\xi)^{-1}(f'(x\ ) - f'(\xi))\|} \leq 2\|f'(\xi)^{-1}\| \leq 2\beta. \end{aligned}$$

Now $x_1 - \xi = x_0 - \xi - f'(x_0)^{-1}f(x_0) = x_0 - \xi - f'(x_0)^{-1}(f(x_0) - f(\xi)) = f'(x_0)^{-1}[f(\xi) - f(x_0) - f'(x_0)(\xi - x_0)]$. So

$$\begin{aligned} \|x_1 - \xi\| & \leq \|f'(x_0)^{-1}\|\|f(\xi) - f(x_0) - f'(x_0)(\xi - x_0)\| \\ & \leq 2\beta\frac{\gamma}{2}\|\xi - x_0\|^2 = \beta\gamma\|x_0 - \xi\|^2 \text{ (by Lemma 6.1.1)} \\ & \leq \beta\gamma\epsilon\|x_0 - \xi\| \leq \frac{1}{2}\|x_0 - \xi\| \leq \epsilon. \end{aligned}$$

The proof is now completed by induction.

Remarks. (1) The above theorem states that the Newton method converges quadratically if $f'(\xi)$ is nonsingular and if starting point is close enough to ξ.

(2) At each step of the Newton method, an evaluation of the Jacobian matrix $f'(x_k)$ is required. Also, a linear system (6.1.5) needs to be solved. All of these mean that the Newton method is an expensive method. So, modifications of the Newton method to make it more efficient is essential in practice.

6.2 The Broyden Method

Consider that in one-dimensional case, the derivative $f'(x_k)$ may be approximated by the finite difference quotient

$$B_k := \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}. \tag{6.10}$$

This choice results in the so called secant method:

$$x_{k+1} = x_k - B_k^{-1} f(x_k) \tag{6.11}$$

and it can be proved that the rate of convergence is $p = \frac{1+\sqrt{5}}{2} \approx 1.618$. In n-dimensional case, we reformulate the relationship (6.10) as

$$B_k(x_k - x_{k-1}) = f(x_k) - f(x_{k-1}) \tag{6.12}$$

which is known as the quasi-Newton condition. If we further write

$$s_k := x_{k+1} - x_k \tag{6.13}$$

$$\Delta f_k := f(x_{k+1}) - f(x_k) \tag{6.14}$$

$$B_{k+1} = B_k + C_k, \tag{6.15}$$

then (6.12) is equivalent to

$$C_k s_k = \Delta f_k - B_k s_k. \tag{6.16}$$

Let $w_k \in R^n$ be an arbitrary vector such that $w_k^T s_k \neq 0$. Then obviously the matrix

$$C_k := \frac{1}{w_k^T s_k}(\Delta f_k - B_k s_k) w_k^T \tag{6.17}$$

satisfies the quasi-Newton condition (6.16).

Definition 6.2.1 *If $w_k := s_k$, then*

$$C_k := \frac{1}{s_k^T s_k}(\Delta f_k - B_k s_k) s_k^T \tag{6.18}$$

is known as Broyden's first method. If $w_k := B_k^T s_k$, then

$$C_k := \frac{1}{s_k^T B_k s_k}(\Delta f_k - B_k s_k) s_k^T B_k \tag{6.19}$$

is known as Broyden's second method.

Theorem 6.2.1 *Let $s_k, \Delta f_k \in R^n$ be given. The matrix C_k given by (6.18) is the minimal change of B_k in the Frobenius norm such that $B_{k+1} = B_k + C_k$ satisfies the quasi-Newton condition $B_{k+1}s_k = \Delta f_k$,.*

(pf): Let $\tilde{C}_k$ denote another possible change of B_k such that $\tilde{B}_{k+1} := B_k + \tilde{C}_k$ satisfies $\tilde{B}_{k+1}s_k = \Delta f_k$. Then

$$\begin{aligned} \|C_k\| = \|B_{k+1} - B_k\|_F &= \frac{1}{s_k^T s_k}(\Delta f_k - B_k s_k)s_k^T\|_F \\ &= \|\frac{1}{s_k^T s_k}(\tilde{B}_{k+1}s_k - B_k s_k)s_k^T\| \\ &\leq \|\tilde{B}_{k+1} - B_k\|_F\|\frac{s_k s_k^T}{s_k^T s_k}\|_F = \|\tilde{C}_k\|_F. \end{aligned}$$

Algorithm 6.2.1 (Broyden's Method)
Given an initial guesst x_0,
Approximate $f'(x_0)$ by a matrix B_0 (say, by a finite difference method),
For $k = 0, 1, \ldots$
Solve $B_k d_k = +f(x_k)$ for d_k,
Determine λ_k for which $\|f(x_k - \lambda_k d_k)\|^2$ is approximately minimized,

$$\begin{aligned} x_{k+1} : &= x_k - \lambda_k d_k, \\ s_k : &= x_{k+1} - x_k, \\ \Delta f_k : &= f(x_{k+1}) - f(x_k), \\ B_{k+1} : &= B_k + \tfrac{1}{s_k^T s_k}(\Delta f_k - B_k s_k)s_k^T. \end{aligned}$$

Lemma 6.2.1 *Let $f : R^n \to R^n$ be continuously differentiable on an open convex set $D \subset R^n$. Suppose there exists a constant γ exists such that $\|f'(x) - f'(y)\| \leq \gamma\|x - y\|$ for $x, y \in D$. Then it holds that for any $x, y, \xi \in D$, $\|f(x) - f(y) - f'(\xi)(u - v)\| \leq \frac{\gamma}{2}(\|x - \xi\| + \|y - \xi\|)\|x - y\|$.*

(pf): The proof is parallel to that of (6.1.1). We have, by the line integral, $\|f(x) - f(y) - f'(\xi)(x - y)\| = \|\int_0^1 [f'(y + t(x - y)) - f'(\xi)](x - y)dt\| \leq \gamma\|x - y\| \int_0^1 \|y + t(x - y) - \xi\|dt \leq \gamma\|x - y\| \int_0^1 \{t\|x - \xi\| + (1 - t)\|y - \xi\}dt.$ $\oplus$

Lemma 6.2.2 *Let $f : R^n \to R^n$ be continuously differentiable on an open convex set $D \subset R^n$. Suppose there exists a constant γ exists such that $\|f'(x) - f'(y)\| \leq \gamma\|x - y\|$ for $x, y \in D$. Then for $x_{k+1}, x_k \in D$, holds that $\|B_{k+1} - f'(\xi)\| \leq \|B_k - f'(\xi)\| + \frac{\gamma}{2}(\|x_{k+1} - \xi\| + \|x_k - \xi\|)$.*

(pf): By definition,

$$\begin{aligned} B_{k+1} - f'(\xi) &= B_k - f'(\xi) + \frac{(\Delta f_k - B_k s_k)s_k^T}{s_k^T s_k} \\ = \quad B_k \left(I = \frac{s_k s_k^T}{s_k^T s_k}\right) &- f'(\xi)\left(I - \frac{s_k s_k^T}{s_k^T s_k}\right) + \frac{(\Delta f_k - f'(\xi)s_k s_k^T}{s_k^T s_k}. \end{aligned}$$

Taking norm, we have

$$\|B_{k+1} - f'(\xi)\| \leq B_k - f'(\xi)\|\|I - \frac{s_k s_k^T}{s_k^T s_k}\| + \|\frac{(\Delta f_k - f'(\xi)s_k)s_k^T}{s_k^T s_k}\|.$$

Observe that $\|I - \frac{s_k s_k^T}{s_k^T s_k}\| \leq 1$. The third term is estimated by

$$\begin{aligned}\|\frac{(\Delta f_k - f'(\xi)s_k)s_k^T}{s_k^T s_k}\| &= \|\frac{\{[f(x_{k+1}) - f(x_k) - f'(\xi)(x_{k+1} - x_k)]\}s_k^T}{s_k^T s_k}\| \\ &\leq \frac{\gamma}{2}(\|x_{k+1} - \xi\| + \|x_k - \xi\|)\end{aligned}$$

by the preceding lemma. $\oplus$

Theorem 6.2.2 *Let $f : R^n \to R^n$ be continuously differentiable on an open convex set $D \subset R^n$. Suppose there exists $\xi \in R^n, \beta, \gamma > 0$ such that*

1. *$f(\xi) = 0$,*
2. *$f'(\xi)^{-1}$ exists,*
3. *$\|f'(\xi)^{-1}\| \leq \beta$, and*
4. *$\|f'(x) - f'(y)\| \leq \gamma\|x - y\|$ for x, y in a neighborhood of ξ.*

Then there exist positive constants δ_1, δ_2 such that if $\|x_0 - \xi\| < \delta_1$ and $\|B_0 - f'(\xi)\| \leq \delta_2$, then the Broyden's method is well defined, converges to ξ, and satisfies

$$\|x_{k+1} - \xi\| \leq c_k\|x_k - \xi\| \tag{6.20}$$

with $\lim_{k\to\infty} c_k = 0$. (This behavior is called superlinear convergence.)

(pf): Choose $\delta_2 \leq \frac{1}{6\beta}$ and $\delta_1 \leq \frac{2\delta}{5\gamma}$. Then $\|f'(\xi)^{-1}B_0 - I\| \leq \beta\delta_2 \leq \frac{1}{6}$. By the Banach lemma, B_0^{-1} exists. So x_1 can be defined. Furthermore,

$$\begin{aligned}\|B_0^{-1}\| &= \|(f'(\xi) + (B_0 - f'(\xi)))^{-1}\| \\ &\leq \frac{\|f'(\xi)^{-1}\|}{1 - \|f'(\xi)^{-1}\|\|B_0 - f'(\xi)\|} \leq \frac{\beta}{1 - \beta\delta_2}.\end{aligned} \tag{6.21}$$

Thus

$$\begin{aligned}\|e_1\| &:= \|x_1 - \xi\| = \|x_0 - B_0^{-1}(f(x_0) - f(\xi)) - \xi\| \\ &= \| - B_0^{-1}[f(x_0) - f(\xi) - B_0(x_0 - \xi)]\| \\ &= \|B_0^{-1}[f(x_0) - f(\xi) - f'(\xi)(x_0 - \xi) + (f'(\xi) - B_0)(x_0 - \xi)]\| \\ &\leq \frac{\beta}{1 - \beta\delta_2}\left[\frac{\gamma}{2}\|e_0\|^2 + \delta_2\|e_0\|\right] \leq \frac{\beta}{1 - \beta\delta_2}\left[\gamma, voer 2\delta_1 + \delta_2\right]\|e_0\| \\ &\leq \frac{\beta}{1 - \beta\delta_2}\frac{6\delta_2}{5}\|e_0\| \leq \frac{\frac{1}{6}}{1 - \frac{1}{6}}\frac{6}{5}\|e_0\| < \frac{1}{2}\|e_0\|.\end{aligned}$$

By (6.2.2), we know

$$\begin{aligned}\|B_1 - f'(\xi)\| &\leq \|B_0 - f'(\xi)\| + \frac{\gamma}{2}(\|x_1 - \xi\| + \|x_0 - \xi\|) \\ &\leq \delta_2 + \frac{\gamma}{2}\left(\frac{3}{2}\|e_0\|\right) \leq \delta_2\left(1 + \frac{\gamma}{2}\frac{3}{2}\frac{2}{5\gamma}\right) \\ &= \left(1 + \frac{3}{10}\right)\delta_2 \leq \frac{3}{2}\delta_2.\end{aligned}$$

Thus $\|f'(\xi)^{-1}B_1 - I\| \leq 2\beta\delta_2 \leq \frac{1}{3}$. By the Banach lemma, B_1^{-1} exists and

$$\|B_1^{-1}\| \leq \frac{\|f'(\xi)^{-1}\|}{1 - \|f'(\xi)^{-1}\|\|B_1 - f'(\xi)\|} \leq \frac{\beta}{1 - 2\beta\delta_2} \leq \frac{3}{2}\beta. \tag{6.22}$$

We can now estimate

$$\begin{aligned}\|e_2\| : &= \|x_2 - \xi\| = \|x_1 - B_1^{-1}(f(x_1) - f(\xi)) - \xi\| \\ &= \| - B_1^{-1}[f(x_1) - f(\xi) - B_1 e_1]\| \\ &= \|B_1^{-1}[f(x_1) - f(\xi) - f'(\xi)e_1 + (f'(\xi) - B_1)e_1]\| \\ &\leq \tfrac{3\beta}{2}\left[\tfrac{\gamma}{2}\|e_1\|^2 + \tfrac{3}{2}\delta_2\|e_1\|\right] \leq \tfrac{3\beta}{2}\left[\tfrac{\gamma}{2}\tfrac{\delta_1}{2} + \tfrac{3}{2}\delta_2\right]\|e_1\| \\ &\leq \tfrac{3\beta\delta_2}{2}\left[\tfrac{\gamma}{2}\tfrac{1}{2}\tfrac{2}{5\gamma} + \tfrac{3}{2}\right]\|e_1\| \leq \tfrac{1}{4}\tfrac{16}{10}\|e_0\| < \tfrac{1}{2}\|e_0\|.\end{aligned}$$

Continuing, we see that

$$\begin{aligned}\|B_2 - f'(\xi)\| &\leq B_1 - f'(\xi)\| + \frac{\gamma}{2}(\|e_2\| + \|e_1\|) \\ &\leq \frac{13}{10}\delta_2 + \frac{\gamma}{2}\left(\frac{3}{2}\|e_1\|\right) \leq \delta_2\left(1 + \frac{3}{10} + \frac{\gamma}{2}\frac{3}{2}\frac{1}{2}\frac{2}{5\gamma}\right) \\ &= \left(1 + \frac{3}{10} + \frac{1}{2}\frac{3}{10}\right)\delta_2 \leq \left(2 - \left(\frac{1}{2}\right)^2\right)\delta_2 \leq 2\delta_2.\end{aligned}$$

The proof is now completed by induction.

Remark. In addition to saving functional evaluation of $f'(x)$, Broyden's method has another important advantage, that is, the matrix factorization of B_{k+1} can easily be updated.

For simplicity, we consider the basic form

$$B_+ = B_c + uv^T \tag{6.23}$$

where u and v represent two column vectors in R^n. Suppose

$$B_c = Q_c R_c \tag{6.24}$$

is already known. We want to find the QR decomposition for B_+. Assume

$$B_+ = Q_+ R_+. \tag{6.25}$$

Let $w := Q_c^T u$. Then $B_+ = Q_c(R_c + wv^T)$. If the QR decomposition of $R_c + wv^T$ is $\tilde{Q}\tilde{R}$, then $Q_+R_+ = (Q_+\tilde{Q})\tilde{R}$ and we are done. But how to find $\tilde{Q}\tilde{R}$? The point is that uv^T is only a rank one matrix. So the QR decomposition of $R_c + wv^T$ would be much cheaper if we perform the orthogonalization process carefully. We first recall the 2-dimensional rotation matrix.

Definition 6.2.2 *A 2-dimensional rotation matrix is a matrix $R(\theta)$ of the form*

$$R(\theta) = \begin{bmatrix} \cos\theta, & \sin\theta \\ -\sin\theta, & \cos\theta \end{bmatrix}, \tag{6.26}$$

or equivalently, a matrix $R(\alpha, \beta)$ of the form

$$R(\alpha, \beta) = \frac{1}{\sqrt{\alpha}^2 + \beta^2} \begin{bmatrix} \alpha, \beta \\ -\beta, \alpha \end{bmatrix}. \tag{6.27}$$

Definition 6.2.3 *A Jacobi rotation matrix is a matrix $J(s, t; \alpha, \beta) \in R^{n\times n}$ of the form, for $\Delta := \sqrt{\alpha}^2 + \beta^2, \overline{\alpha} := \frac{\alpha}{\Delta}, \overline{\beta} := \frac{\beta}{\Delta}$,*

$$J = \begin{bmatrix} 1 & 0 & & & 0 & \\ 0 & \overline{\alpha} & 0 & 0 & \overline{\beta} & 0 \\ 0 & & 1 & & & \\ 0 & & 1 & & & \\ 0 & \overline{-\beta} & & \overline{\alpha} & 0 & \\ 0 & & & 0 & 1 & \end{bmatrix} \begin{matrix} \\ \\ \rightarrow s - throw \\ \rightarrow t - throw \\ \\ \\ \end{matrix} \tag{6.28}$$

Remarks. (1) It is easy to see that

$$J(s, t; \alpha, \beta) \begin{bmatrix} a_1 \\ \\ \\ a_n \end{bmatrix} = \begin{bmatrix} a_1 \\ \overline{\alpha}a_s + \overline{\beta}a_t \\ -\overline{\beta}a_s + \overline{\alpha}a_t \\ a_n \end{bmatrix}. \tag{6.29}$$

(2) Let $v = [v_1 v_2]$ be a 2-dimensional vector. Then $R(\theta)v$ rotates v by an angle θ counterclockwise. Indeed, $R(v_1, v_2)v = \begin{bmatrix} \|v\| \\ 0 \end{bmatrix}$ and $R(v_2, -v_1)v =$.

Consider now the QR decomposition of the matrix $R_c + wv^T$. Note that $wv^T = \begin{bmatrix} w_1 v^T \\ w_n v^T \end{bmatrix}$. Let $c_1 := \sqrt{w_{n-1}^2 + w_n^2}$. Then with $\hat{Q}_1 := J(n-1, n; w_{n-1}, w_n)$ we have $\hat{Q}_1 wv^T = \begin{bmatrix} w_1 v^T \\ \vdots \\ w_{n-2} v^T \\ c_1 v^T \\ 0 \end{bmatrix}$. Let $c_2 := \sqrt{w_{n-2}^2 + c_1^2}$ and $\hat{Q}_2 := J(n-2, n-$

$1; w_{n-2}, c_1)$, we have $\hat{Q}_1\hat{Q}_2 w v^T = \begin{bmatrix} w_1 v^T \\ \vdots \\ w_{n-3} v^T \\ c_2 v^T \\ 0 \\ 0 \end{bmatrix}$. Continuing this process for $n-1$ iterations, we obtain

$$\hat{Q}_{n-1}\dots\hat{Q}_1(R_c + wv^T) = \hat{Q}_{n-1}\dots\hat{Q}_1 R_c + \begin{bmatrix} \|w\|v^T \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{6.30}$$

We note $\hat{Q}_{n-1}\dots\hat{Q}_1 R_c$ to the worst is an upper Hessenberg matrix. So the matrix $R_c + wv^T$ has been reduced by rotation matrices to an upper Hessenberg matrix. We can now continue to do a sequence of plane rotations to change the upper Hessenberg matrix into an upper triangular matrix (Recall how the QR algorithm works!).

6.3 Sturm Sequences

The fundamental theorem of algebra asserts that a polynomial $p(x)$ of degree n,

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \dots + a_1 x + a_0, \tag{6.31}$$

has exactly n (complex) roots, if counting multiplicity. If all the coefficients are real numbers, it is often desirable to determine the number of real roots of $p(x)$ in a specified region (a, b) where either a or b may be infinite. Toward this end, the concept of Sturm sequences offers a very useful technique here.

Definition 6.3.1 *A sequence*

$$p(x) = f_1(x), \dots, f_m(x) \tag{6.32}$$

of real polynomials is called a Sturm sequence on an interval (a, b) if

1. *The last polynomial $f_m(x)$ does not vanish in (a, b);*
2. *At any zero ξ of $f_k(x), k = 2, \dots, m-1$, the two adjacent polynomials are nonzero and have opposite signs, that is*

$$f_{k-1}(\xi) f_{k+!}(\xi) < 0. \tag{6.33}$$

Definition 6.3.2 *Let $\{f_k(x)\}$ be a Sturm sequence on (a, b), and let $x_0 \in (a, b)$ at which $f_1(x) \neq 0$. We define $V(x_0)$ to be the number of changes of sign of $\{f_k(x_0)\}$, zero values being ignored. If a is finite, then $V(a)$ is defined as $V(a + \epsilon)$, where ϵ is such that no $f_k(x)$ vanishes in $(a, a + \epsilon)$. If $a = -\infty$, then $V(a)$ is defined to be the number of changes of signs of $\{\lim_{x\to-\infty} f_k(x)\}$. Similarly, $V(b)$ is defined.*

Definition 6.3.3 *Let $R(x)$ be any rational function. The Cauchy index of $R(x)$ on (a,b), denoted by $I_a^b R(x)$, is defined to be the difference between the number of jumps of $R(x)$ from $-\infty$ to $+\infty$ and the number of jumps from $+\infty$ to $-\infty$ as x goes from a to b, excluding the endpoints. That is, at every pole of $R(x)$ in (a,b) add 1 to the Cauchy index if $R(x) \to -\infty$ on the left of the pole and $R(x) \to +\infty$ on the right of the pole, and subtract 1 if vice versa.*

Theorem 6.3.1 *(Sturm) If $f_k(x), k = 1, \ldots, m$ is a Sturm sequence on an interval (a,b), then if neither $f_1(a)$ nor $f_1(b)$ equals 0, we have*

$$I_a^b \frac{f_2(x)}{f_1(x)} = V(a) - V(b). \tag{6.34}$$

(pf): we first claim that the value of $V(x)$ does not change when x passes through a zero of $f_k(x), k = 2, \ldots, m-1$. To see this, suppose $f_k(\xi) = 0$. Then by (6.33), $f_{k+!}(\xi) f_{k-1}(\xi) < 0$. If $f_k(x)$ changes sign at $x = \xi$, then for a sufficiently small perturbation $h > 0$, the signs of the polynomials $f_{k-1}(x), f_k(x)$ and $f_{k+1}(x)$ display the behavior in one of the following four patterns: $\begin{matrix} --- \\ -0+ \\ +++ \end{matrix}$, $\begin{matrix} +++ \\ -0+ \\ --- \end{matrix}$, $\begin{matrix} --- \\ +0- \\ +++ \end{matrix}$, $\begin{matrix} +++ \\ +0- \\ --- \end{matrix}$. In each case, $V(\xi - h) = V(\xi) = V(\xi + h)$. This is also true if $f_k(x)$ does not change sign at $x = \xi$. Thus $V(x)$ can change only when $f_1(x)$ goes through 0. If ξ is a zero of $f_1(x)$, it is not a zero of $f_2(x)$ because of property 2 of Sturm sequences. Therefore, $f_2(x)$ has the same sign on both sides of ξ. If ξ is a zero of $f_1(x)$ of even multiplicity, when $V(x)$ does not change as x increases through ξ and there is not contribution to the Cauchy index. If the zero is of odd multiplicity, then $V(x)$ will increase by 1 if $f_1(x)$ and $f_2(x)$ have the same sign to the left of ξ, (i.e., $\begin{matrix} -0+ \\ --- \end{matrix}$, or $\begin{matrix} +0- \\ +++ \end{matrix}$) and will decrease by 1 if the signs to the left are different, (i.e., $\begin{matrix} -0+ \\ +++ \end{matrix}$, or $\begin{matrix} +0- \\ --- \end{matrix}$). Correspondingly for zeros of odd multiplicity, there is a -1 contribution to the Cauchy indes if the signs of $f_1(x)$ and $f_2(x)$ are the same to the left of ξ and +1 contribution if they are different. This establishes the theorem. $\oplus$

We now apply the Sturm theorem to find the real roots of $p(x)$ in an interval (a,b). Consider the sequence of functions $f_k(x), k = 1, \ldots, m$ where

$$\begin{array}{lcl} f_1(x) & = & p(x) \\ f_2(x) & = & p'(x) \\ f_{j-1}(x) & = & q_{j-1}(x) f_j(x) - f_{j+1}(x), j = 2, \ldots, m-1 \\ f_{m-1}(x) & = & q_{m-1}(x) f_m(x) \end{array} \tag{6.35}$$

where $q_{j-1}(x)$ is the quotient and $f_{j+1}(x)$ is the negative of the remainder when $f_{j-1}(x)$ is divided by $f_j(x)$. Thus $\{f_k(x)\}$ is a sequence of polynomials of decreasing degree which eventually must terminate in a polynomial $f_m(x), m \le n+1$, which divides $f_{m-1}(x)$ (why?). The polynomial $f_m(x)$ is the

greatest common divisor of $f_1(x)$ and $f_2(x)$ and also of every other member of the sequence (6.35). (This is the so called the Euclidean algorithm.)

Suppose $f_m(x)$ does not vanish in (a, b) so that the first condition of Definition 6.3.1 is satisfied. If $f_k(\xi) = 0$ for any $k, k = 2, \ldots, m-1$, then $f_{k-1}(\xi) = -f_{k+1}(\xi)$ by (6.35). Moreover, when $f_{k+1}(\xi) \neq 0$ since otherwise $f_m(\xi)$ would also be 0 (Why?). Thus the sequence $\{f_k(x)\}$ is a Sturm sequence when $f_m(x)$ does not vanish in (a, b).

Suppose $f_m(x)$ is not of constant sign in (a, b), then we use the sequence $\{f_k(x)/f_m(x)\}$. Then not only is this a Sturm sequence but also both sides of (6.34) are the same for this sequence and for the sequence $\{f_k(x)\}$. Therefore, we can use these two sequences interchangeably in applying Sturm's theorem.

Now for the sequence $\{f_k(x)\}$ define by (6.35), we write

$$\frac{f_2(x)}{f_1(x)} = \frac{p'(x)}{p(x)} = \sum_{j=1}^{p} \frac{n_j}{x - a_j} + R_1(x) \tag{6.36}$$

where the $a_j, j = 1, \ldots, p$, are the distinct real zeros of $p(x)$, n_j is the multiplicity of the zeros a_j, and $R_1(x)$ has no poles on the real axis. Since the n_j are all positive, $I_a^b(p'(x)/p(x))$ is equal to the number of distinct real zeros of $p(x)$ in the interval (a, b). Therefore, we have the following theorem:

Theorem 6.3.2 *The number of distinct real zeros of the polynomial $p(x)$ in the interval (a, b) is equal to $V(a) - V(b)$ if neither $f(a)$ nor $f(b)$ is equal to 0.*

Example. Consider the polynomial

$$p(x) = x^6 + 4x^5 + 4x^4 - x^2 4x - 4. \tag{6.37}$$

Using (6.35), we calculate

$$\begin{aligned} f_1(x) &= x^6 + 4x^5 + 4x^4 - x^2 - 4x - 4 \\ f_2(x) &= 6x^5 + 20x^4 + 16x^3 - 2x - 4 \\ f_3(x) &= 4x^4 + 8x^3 + 3x^2 + 14x + 16 \\ f_4(x) &= x^3 + 6x^2 + 12x + 8 \\ f^5(x) &= -17x^2 - 58x - 48 \\ f_6(x) &= -x - 2 \end{aligned} \tag{6.38}$$

where the coefficients have been made integers by multiplying by suitable positive constants. For some sample values of x the signs of the $f_k(x)$ are:

	$-\infty$	∞	0	-1	+1	-24/17
$f_1(x)$	+	+	-	0	0	+
$f_2(x)$	-	+	-	-	+	-
$f_3(x)$	+	+	+	+	+	-
$f_4(x)$	-	+	+	+	+	+
$f_5(x)$	-	-	-	-	-	0
$f_6(x)$	+	-	-	-	-	-
# of sign change	4	1	2	2	1	3

(6.39)

Thus we conclude there are three distinct real zeros, two negative and one positive. Moreover, there are two distinct zeros in $(-\infty, -1]$ and three in $(-\infty, +1]$.

6.4 Bairstow's Method

A real polynomial may have complex conjugate roots. In order to find the complex roots, most methods would have to begin at a complex starting point and be carried out in complex arithmetic. Bairstow's method avoids complex arithmetic.

The roots of a quadratic polynomial

$$d(x) = x^2 - rx - q \tag{6.40}$$

are obviously known. For a polynomial $p(x)$, we write

$$p(x) = p_1(x)(x^2 - rx - q) + Ax + B. \tag{6.41}$$

The coefficients of the remainder depends upon r and q. The idea of Bairstow's method is to determine r and q so that

$$A(r, q) = 0 \tag{6.42}$$
$$B(r, q) = 0 \tag{6.43}$$

Applying Newton's method to (6.42), we need to compute

$$\begin{bmatrix} r_{i+1} \\ q_{i+1} \end{bmatrix} = \begin{bmatrix} r_i \\ q_i \end{bmatrix} - \begin{bmatrix} \frac{\partial A}{\partial r} & \frac{\partial A}{\partial q} \\ \frac{\partial B}{\partial r} & \frac{\partial B}{\partial q} \end{bmatrix}^{-1}_{(r_i, q_i)} \begin{bmatrix} A(r_1, q_i) \\ B(r_1, q_i) \end{bmatrix}. \tag{6.44}$$

Upon differentiating, we observe that

$$0 \equiv \frac{\partial p(x)}{\partial r} = \frac{\partial p_1}{\partial r}(x)(x^2 - rx - q) - p_1(x)x + \frac{\partial A}{\partial r}x + \frac{\partial B}{\partial r} \tag{6.45}$$
$$0 \equiv \frac{\partial p(x)}{\partial q} = \frac{\partial p_1}{\partial q}(x)(x^2 - rx - q) - p_1(x) + \frac{\partial A}{\partial q}x + \frac{\partial B}{\partial q}. \tag{6.46}$$

Let $p_1(x)$ be further divided by $d(x)$ and denote

$$p_1(x) = p_2(x)(x^2 - rx - q) + \tilde{A}x + \tilde{B}. \tag{6.47}$$

Assuming the two roots x_0 and x_1, of $d(x)$ are distinct, it follows that

$$p_1(x_i) = \tilde{A}x_i + \tilde{B}. \tag{6.48}$$

Substituting (6.46) we obtain

$$-x_i(\tilde{A}x_i + \tilde{B}) + \frac{\partial A}{\partial r}x_i + \frac{\partial B}{\partial r} = 0 \tag{6.49}$$
$$-(\tilde{A}x_i + \tilde{B}) + \frac{\partial A}{\partial q}x_i + \frac{\partial B}{\partial q} = 0. \tag{6.50}$$

Form the second equations we have (since $x_0 \neq x_1$)

$$\frac{\partial A}{\partial q} = \tilde{A}, \frac{\partial B}{\partial q} = \tilde{B}. \tag{6.51}$$

Form the first equations, we have

$$-\tilde{A}(rx_i + q) + x_i(\frac{\partial A}{\partial r} - \tilde{B}) + \frac{\partial B}{\partial r} = 0. \tag{6.52}$$

Therefore, we know

$$\frac{\partial A}{\partial r} = \tilde{B} = \tilde{A}r \tag{6.53}$$

$$\frac{\partial B}{\partial r} = \tilde{A}q. \tag{6.54}$$

The values of A, B can be obtained without difficulty: Suppose

$$p(x) = a_n x^n + \ldots + a_1 x + a_0 \tag{6.55}$$

$$p_1(x) = b_{n-2}x^{n-2} + \ldots + b_0. \tag{6.56}$$

By comparing coefficients of (6.41), we obtain

$$b_{n-2} = a_n, \tag{6.57}$$

$$b_{n-3} = b_{n-2}r + a_{n-1}, \tag{6.58}$$

$$b_{n-k} = b_{n-k+2}q + b_{n-k+1}r + a_{n-k+2} \tag{6.59}$$

$$A = b_1 q + b_0 r + a_1 \tag{6.60}$$

$$B = b_0 q + a_0. \tag{6.61}$$

Similarly, by using (6.47), we can obtain the values of $\tilde{A}$ and $\tilde{B}$.