4 Solution of systems of nonlinear equations

Given a system of nonlinear equations

$$\mathbf{f}(\mathbf{x}) = 0, \qquad \mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m \tag{12}$$

for which we assume that there is (at least) one solution **xi**. The idea is to rewrite this system into the form

$$\mathbf{x} = \mathbf{g}(\mathbf{x}), \qquad \mathbf{g} : \mathbb{R}^m \to \mathbb{R}^m. \tag{13}$$

The solution ξ of (12) should satisfy $\xi = \mathbf{g}(\xi)$, and is thus called a *fixed point* of $\mathbf{g}$. The iteration schemes becomes: given an initial guess $\mathbf{x}^{(0)}$, the *fixed point iterations* becomes

$$\mathbf{x}^{(k+1)} = \mathbf{g}(\mathbf{x}^{(k)}), \qquad k = 1, 2, \dots . \tag{14}$$

The following questions arise:

(i) How to find a suitable function $\mathbf{g}$?
(ii) Under what conditions will the sequence $\mathbf{x}^{(k)}$ converge to the fixed point ξ?
(iii) How quickly will the sequence $\mathbf{x}^{(k)}$ converge?

Point (ii) can be answered by Banach fixed point theorem:

Theorem 4.1. *Let $D \subseteq \mathbb{R}^m$ be a and closed set. If*

$$\mathbf{g}(D) \subseteq D \tag{15a}$$

and

$$\|\mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{v})\| \le L\|\mathbf{y} - \mathbf{v}\|, \qquad \textit{with } L < 1 \textit{ for all } \mathbf{y}, \mathbf{v} \in D, \tag{15b}$$

then G has a unique fixed point in D and the fixed point iterations (14) *converges for all $\mathbf{x}^{(0)} \in D$. Further,*

$$\|\mathbf{x}^{(k)} - \xi\| \le \frac{L^k}{1-L}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|. \tag{15c}$$

Proof. The proof is based on the *Cauchy Convergence theorem*, saying that a sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converges to some ξ if and only if for every $\varepsilon > 0$ there is an N such that

$$\|\mathbf{x}^{(l)} - \mathbf{x}^{(k)}\| < \varepsilon \qquad \text{for all} \quad l, k > N. \tag{16}$$

Assumption (15a) ensures $\mathbf{x}^{(k)} \in D$ as long as $\mathbf{x}^{(0)} \in D$. From (14) and (15b) we get:

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| = \|\mathbf{g}(\mathbf{x}^{(k)}) - \mathbf{g}(\mathbf{x}^{(k-1)})\| \le L\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \le L^k\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|.$$

We can write $\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)} = \sum_{i=1}^{p}(\mathbf{x}^{(k+i)} - \mathbf{x}^{(k+i-1)})$, thus

$$\begin{aligned}\|\mathbf{x}^{(k+p)} - \mathbf{x}^{(k)}\| &\le \sum_{i=1}^{p}\|\mathbf{x}^{(k+i)} - \mathbf{x}^{(k+i-1)}\| \\ &= (L^{p-1} + L^{p-2} + \cdots + 1)\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le \frac{L^k}{1-L}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|,\end{aligned}$$

since $L < 1$. For the same reason, the sequence satisfy (16), so the sequence converges to some $\xi \in D$. Since the inequality is true for all $p > 0$ it is also true for ξ, proving (15c).

To prove that the fixed point is unique, let ξ and η be two different fixed points in D. Then

$$\|\xi - \eta\| = \|\mathbf{g}(\xi) - \mathbf{g}(\eta)\| < \|\xi - \eta\|$$

which is impossible. □

For a given problem, it is not necessarily straightforward to justify the two assumptions of the theorem. But it is sufficient to find some L satisfying the condition $L < 1$ in some norm to prove convergence.

Let $\mathbf{x} = [x_1, \dots, x_m]^T$ and $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \dots, g_m(\mathbf{x})]^T$. Let $\mathbf{y}, \mathbf{v} \in D$, assume D to be convex,[2] and let $\mathbf{x}(\theta) = \theta\mathbf{y} + (1-\theta)\mathbf{v}$ be the straight line between $\mathbf{y}$ and $\mathbf{v}$. According to the mean value theorem for functions, for each g_i there exist at $\tilde{\theta}_i$ such that

$$\begin{aligned} g_i(\mathbf{y}) - g_i(\mathbf{v}) = g_i(\mathbf{x}(1)) - g_i(\mathbf{x}(0)) &= \frac{dg_i}{d\theta}(\tilde{\theta}_i)(1-0), && \tilde{\theta}_i \in (0,1) \\ &= \sum_{j=1}^{m} \frac{\partial g_i}{\partial x_j}(\tilde{\mathbf{x}}_i)(y_j - v_j), && \tilde{\mathbf{x}}_i = \tilde{\theta}_i\mathbf{y} + (1-\tilde{\theta}_i)\mathbf{v} \end{aligned}$$

since $dx_j(\theta)/d\theta = y_j - v_j$. Then

$$|g_i(\mathbf{y}) - g_i(\mathbf{v})| \le \sum_{j=1}^{m} |\frac{\partial g_i}{\partial x_j}(\tilde{\mathbf{x}}_i)| \cdot |y_j - v_j| \le \left(\sum_{j=1}^{m} |\frac{\partial g_i}{\partial x_j}(\tilde{\mathbf{x}}_i)| \right) \max_l |y_l - v_l|.$$

If we let $\bar{g}_{ij}$ be some upper bound for each of the partial derivatives, that is

$$|\frac{\partial g_i}{\partial x_j}(\mathbf{x})| \le \bar{g}_{ij}, \quad \text{for all } \mathbf{x} \in D.$$

then

$$\|\mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{v})\|_\infty = \left(\max_i \sum_{j=1}^{m} \bar{g}_{ij} \right) \|\mathbf{y} - \mathbf{v}\|_\infty.$$

We can then conclude that (15b) is satisfied if

$$\max_i \sum_{j=1}^{m} \bar{g}_{ij} < 1.$$

Newton's method

Newton's method is a fixed point iterations for which

$$\mathbf{g}(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)} - J_f(\mathbf{x}^{(k)})^{-1}\mathbf{f}(\mathbf{x}^{(k)}), \tag{17}$$

[2] D is convex if $\theta y + (1-\theta)v \in D$ for all $y, v \in D$ and $\theta \in [0,1]$.

where the *Jacobian* is the matrix function

$$J_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_m}(\mathbf{x}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_m}(\mathbf{x}) \end{pmatrix}.$$

The Newton method can be derived as follow: Consider element i in $\mathbf{f}$, that is $f_i(\mathbf{x})$. Do a multidimensional Taylor expansion of $f_i(\xi)$ around the vector $\mathbf{x}^{(k)}$, using $\mathbf{e}^{(k)} = \xi - \mathbf{x}^{(k)}$ This gives

$$0 = f_i(x_1^{(k)} + e_1^{(k)}, \ldots, x_m^{(k)} + e_m^{(k)}) = f_i + \frac{\partial f_i}{\partial x_1} e_1^{(k)} + \cdots + \frac{\partial f_i}{\partial x_m} e_m^{(k)} + R_i$$

The function and all the derivatives are evaluated in $\mathbf{x}^{(k)}$. The remainder term R_i consists of quadratic terms like $\mathcal{O}(e_i^{(k)} e_j^{(k)})$. If the error is small, this term is even smaller, so let us now ignore it and replace the errors $e_i^{(k)}$ with an approximation to the error $\Delta x_i^{(k)}$ to compensate. Doing so for each $i = 1, 2, \ldots, m$ gives us the following system of linear equations,

$$f_i + \frac{\partial f_i}{\partial x_1} \Delta x_1^{(k)} + \cdots + \frac{\partial f_i}{\partial x_m} \Delta x_m^{(k)} = 0, \qquad i = 1, 2, \ldots, m.$$

which is

$$\mathbf{f}(\mathbf{x}^{(k)}) + J_f(\mathbf{x}^{(k)}) \cdot \Delta x^{(k)} = \mathbf{0}.$$

Solve this with repect to $\Delta\mathbf{x}^{(k)}$. Remember that $\Delta\mathbf{x}^{(k)} \approx \xi - \mathbf{x}_k^{(k)}$ it seems reasonable to update our iterate with this amount, thus

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

which finally results in (17).

It is possible to prove, e.g. [1, Sec. 7.1] that if *i)* (12) has a solution ξ, *ii)* $J_f(\mathbf{x})$ is nonsingular in some open neighbourhood around ξ and *iii)* the initial guess $\mathbf{x}^{(0)}$ is sufficiently close to ξ, the Newton iterations will converge to ξ and

$$\|\xi - \mathbf{x}^{(k+1)}\| \leq K \|\xi - \mathbf{x}^{(k)}\|^2$$

for some positive constant K. We say that the convergence is *quadratic*.

Steepest descent

Steepest descent is an algorithm that search for a (local) minimum of a given function $\psi : \mathbb{R}^m \to \mathbb{R}$. The idea is as follows.

a) Given some point $\mathbf{x} \in \mathbb{R}^m$.
b) Find the direction of steepest decline of ψ from $\mathbf{x}$ (steepest descent direction)
c) Walk steady in this direction till ψ starts to increase again.
d) Repeat from a).

The direction of steepest descent is $-\nabla\psi(\mathbf{x})$, where the gradient $\nabla\psi$ is given by

$$\nabla\psi(\mathbf{x}) = \left[\frac{\partial\psi}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial\psi}{\partial x_m}(\mathbf{x})\right]^T.$$

And the steepest descent algorithm reads

---

function STEEPEST DESCENT(ψ, $\mathbf{x}^{(0)}$)
    for k=0,1,2,.... **do**
        $\mathbf{p} = -\nabla\psi(\mathbf{x}^{(k)})/\|\nabla\psi(\mathbf{x}^{(k)})\|$ ▷ The steepest descent direction.
        Minimize $\psi(\mathbf{x}^{(k)} + \alpha\mathbf{p})$, giving $\alpha = \alpha^\star$.
        $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^\star\mathbf{p}$
    end for
end function

This algorithm will always converge to some point ξ in which $\nabla\psi(\xi) = 0$, usually a local minimum, if one exist. But the convergence can be very slow.

This can be used to find solution of the nonlinear system of equations (12) by defining

$$\psi(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x}) = \|\mathbf{f}(\mathbf{x})\|_2^2.$$

Thus, ξ is a minimum of $\psi(\mathbf{x})$ if and only if ξ is a solution of $\mathbf{f}(\mathbf{x}) = 0$. In this case, we can show that

$$\nabla\psi(\mathbf{x}) = 2J_f(\mathbf{x})^T\mathbf{f}(\mathbf{x}).$$

References

[1] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer-Verlag, Berlin, second edition, 2007.