Nonlinear Equations

1 Introduction

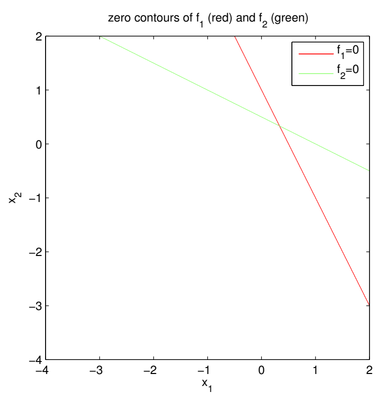
In applications we usually need to find several unknown values $x_1,\dots,x_n$. We have n equations for $x_1,\dots,x_n$

$$\begin{aligned} f_1(x_1,\dots,x_n) &= 0 \\ &\vdots \\ f_n(x_1,\dots,x_n) &= 0 \end{aligned}$$

and we want to find the solutions.

In many cases the problem can be (approximatively) described by **linear equations**. In this case we have n linear equations for n unknowns. We will get a unique solution if the matrix is nonsingular.
Example with $n = 2$: Find x_1, x_2 such that

$$\begin{aligned} 2x_1 + x_2 - 1 &= 0 \\ x_1 + 2x_2 - 1 &= 0 \end{aligned}$$

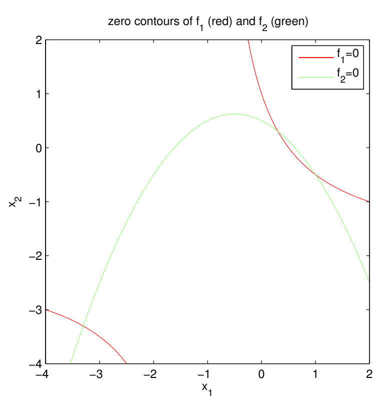


Here we have one solution $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$ which is the intersection of the red and the green line.

In other cases the problem is nonlinear, and we obtain n **nonlinear equations**.

Example with $n = 2$: Find x_1, x_2 such that

$$\begin{aligned} 2x_1 + x_2 + x_1x_2 - 1 &= 0 \\ x_1 + 2x_2 + x_1^2 - 1 &= 0 \end{aligned}$$



Here we have three solutions $\begin{bmatrix} .3028 \\ .3028 \end{bmatrix}, \begin{bmatrix} 1 \\ -.5 \end{bmatrix}, \begin{bmatrix} -3.3028 \\ -3.3028 \end{bmatrix}$.

```
f = @(x) [ 2*x(1)+x(2)+x(1)*x(2)-1 ; x(1)+2*x(2)+x(1)^2-1 ]  % Define function f
xs = fsolve(f,[0;0])                                          % Find solution near [0;0]
```

2 One nonlinear equation

2.1 Introduction

2.2 Bisection Method

Assume that the function f is continuous. If we have two function values $f(a), f(b)$ with opposite signs then the intermediate value theorem guarantees that there must be a point $x_* \in (a,b)$ with $f(x_*) = 0$. This motivates the bisection method:

Algorithm: Bisection method

The algorithm gives a sequence of intervals $[a_k, b_k]$. There exists a solution $x_*$with

- Initial guesses a_0, b_0 where $f(a_0)$ and $f(b_0)$ have different signs
- For $k = 0, 1, 2, \ldots$:
  $c_k := (a_k + b_k)/2$
  If $f(c_k), f(a_k)$ have different sign: $[a_{k+1}, b_{k+1}] := [a_k, c_k]$
  If $f(c_k), f(a_k)$ have same sign: $[a_{k+1}, b_{k+1}] := [c_k, b_k]$
  If $f(c_k) = 0$: stop

Theorem 2.1. *Assume that the function f is continuous on $[a_0, b_0]$. If $f(a_0)$ and $f(b_0)$ have different sign, then the bisection method converges:*

$$\lim_{k\to\infty} a_k = \lim_{k\to\infty} b_k = x_* \qquad \textit{with } f(x_*) = 0.$$

Note that the midpoint c_k satisfies $|c_k - x_*| \leq (b_k - a_k)/2$, therefore we have decreasing error bounds E_k

$$|c_k - x_*| \leq E_k, \qquad E_{k+1} = \frac{1}{2} E_k$$

If we have error bounds E_k with $E_{k+1} \leq C \cdot E_k$ (where $C < 1$) we say we have **convergence of order 1**.

2.3 Secant Method

Assume that we have two function values $f(a)$ and $f(b)$. Based on this information we want to find a good guess c for the solution x_*: We can approximate $f(x)$ by the linear interpolation

$$p(x) = f(b) + f[a,b](x - b)$$

where $f[a,b] = \frac{f(b)-f(a)}{b-a}$. Then we find c such that $p(c) = 0$: Solving $f(b) + f[a,b](c - b) = 0$ for c gives

$$c = b - f(b)/f[a,b].$$

If we have two initial guesses x_0, x_1 we can use this to find an improved guess x_2. Using x_1, x_2 we find x_3, etc.

Algorithm: Secant Method

- Initial guesses x_0, x_1
- For $k = 1, 2, 3, \ldots$:
  $x_{k+1} := x_k - f(x_k)/f[x_{k-1}, x_k]$

During the algorithm we have $a = x_{k-1}$ and $b = x_k$. We then compute $c = x_{k+1}$ using the secant. We want to show that the new error $|c - x_*|$ is small:

From the interpolation error we know that

$$f(x_*) - p(x_*) = R(x_*), \qquad R(x_*) = \frac{1}{2} f''(t) \cdot (x_* - a)(x_* - b)$$

(where t is somewhere between a,b,x_*). If $|f''(t)| \le C_2$ we have $|R(x_*)| \le \frac{C_2}{2}|x_* - a| \cdot |x_* - b|$.

Note that $f(x_*) = 0 = p(c)$. Hence

$$\underbrace{p(c) - p(x_*)}_{f[a,b]\cdot(c-x_*)} = R(x_*)$$

since $p(x)$ is a linear function with slope $f[a,b]$. Therefore

$$c - x_* = \frac{R(x_*)}{f[a,b]}$$

We have $f[a,b] = f'(s)$ with $s \in [a,b]$. If $|f'(s)| \ge C_1 > 0$ we therefore have with $D := \frac{C_2}{2C_1}$

$$\boxed{|c - x_*| \le D\,|a - x_*| \cdot |b - x_*|} \tag{1}$$

Since $a = x_{k-1}$, $b = x_k$, $c = x_{k+1}$ we obtain

$$|x_{k+1} - x_*| \le D\,|x_{k-1} - x_*| \cdot |x_k - x_*|$$

Let $e_k := D\,|x_k - x_*|$. Multiplying by D gives

$$e_{k+1} \le e_{k-1} e_k.$$

Now assume that

$$e_0 \le q, \quad e_1 \le q \qquad \text{with } q < 1$$

Then we obtain

$$e_0 \le q^1, \quad e_1 \le q^2, \quad e_2 \le q^3, \quad e_3 \le q^5, \quad \ldots \quad e_k \le q^{F_k}$$

with the **Fibonacci number** F_k (defined by $F_0 = 1$, $F_1 = 1$, $F_{k+1} = F_k + F_{k-1}$). Since $q < 1$ and $F_k \to \infty$ for $k \to \infty$ we obtain convergence $e_k = D \cdot |x_k - x_*| \to 0$ if our assumptions

$$\left|f''(t)\right| \le C_2, \qquad \left|f'(t)\right| \ge C_1 > 0 \tag{2}$$

are satisfied. The order of convergence corresponds to the ratio F_k/F_{k-1} which converges to the golden ratio $\frac{\sqrt{5}+1}{2}$.

Theorem 2.2. *Assume that $f(x_*) = 0$ and*

- *$f'(x)$ and $f''(x)$ exist and are continuous near x_**
- *$f'(x_*) \ne 0$.*

Then there exists $\delta > 0$, $C > 0$ such that for $|x_0 - x_*| \le \delta$, $|x_1 - x_*| \le \delta$ we have

- *$\lim_{k\to\infty} x_k = x_*$ (convergence)*
- *$|x_k - x_*| \le E_k$ and $E_{k+1} \le C E_k^\alpha$ with $\alpha = \frac{\sqrt{5}+1}{2}$ (convergence with order $\alpha > 1$)*

Proof. Pick $\varepsilon > 0$ such that on the interval $B_\varepsilon = [x_* - \varepsilon, x_* + \varepsilon]$ we have that $f'(x) > 0$ and f'' is continuous:

$$\text{For } x \in B_\varepsilon: \qquad \left|f'(x)\right| \ge C_1 > 0, \quad \left|f''(x)\right| \le C_2 \tag{3}$$

with some constants C_1, C_2. Let $D = \frac{C_2}{2C_1}$. Pick $q < 1$ such that $\delta := q/D \le \varepsilon$.

Now assume $|x_{k-1} - x_*| \le \delta$, $|x_k - x_*| \le \delta$. Since $\delta \le \varepsilon$ we have $x_{k-1}, x_k, x_* \in B_\varepsilon$. We now have

$$|x_{k+1} - x_*| = \frac{|f''(t)|}{2\,|f'(s)|}\,|x_k - x_*| \cdot |x_{k-1} - x_*|$$

where the intermediate points s,t are located between x_0,x_1,x_*. Hence we have $s,t\in B_\varepsilon$ and (3) gives

$$|x_{k+1}-x_*|\le D|x_k-x_*|\cdot|x_{k-1}-x_*|\le \underbrace{D\delta}_{q<1}\cdot\delta<\delta$$

so that we also have $|x_{k+1}-x_*|\le\delta$.

Therefore we obtain by induction that $|x_k-x_*|\le\delta$ for $k=0,1,2,\ldots$, and that

$$|x_{k+1}-x_*|\le D|x_k-x_*|\cdot|x_{k-1}-x_*|$$

As we saw above, this implies that $e_k:=D|x_k-x_*|$ satisfies $e_k\le q^{F_k}$ where F_k are the Fibonacci numbers. Since $q<1$ and $F_k\to\infty$ we obtain convergence $\lim_{k\to\infty}x_k=x_*$.

It remains to prove convergence of order $\alpha=\frac{\sqrt{5}+1}{2}$: We have shown $e_k\le\tilde{E}_k:=q^{F_k}$. Since the Fibonacci numbers satisfy $F_{k+1}-\alpha F_k=(1-\alpha)^{k+1}$ we have

$$\begin{aligned}
& F_{k+1}\ge\alpha F_k-1\\
\Rightarrow\quad & q^{F_{k+1}}\le q^{\alpha F_k}\cdot q^{-1}\\
\Rightarrow\quad & \tilde{E}_{k+1}\le\tilde{E}_k^{\alpha}\cdot q^{-1}
\end{aligned}$$

□

2.4 Newton Method

For the secant method we used the interpolating polynomial with the nodes a,b. Now assume that $a=b$, and that we know $f(a)$ and $f'(a)$. We can approximate $f(x)$ by the linear interpolation

$$p(x)=f(a)+f[a,a](x-a)$$

where $f[a,a]=f'(a)$. Then we find c such that $p(c)=0$: Solving $f(a)+f[a,a](c-a)=0$ for c gives

$$c=b-f(a)/f[a,a].$$

If we have an initial guesses x_0 we can use this to find an improved guess x_1, etc.:

Algorithm: Newton Method

- Initial guess x_0
- For $k=1,2,3,\ldots$:
  $x_{k+1}:=x_k-f(x_k)/f[x_{k-1},x_k]$

For the errors we obtain from (1) with $a=b=x_k$, $c=x_{k+1}$ that

$$|x_{k+1}-x_*|\le D|x_k-x_*|^2$$

if the assumptions (2) hold. Multiplying this by D gives with $e_k:=D|x_k-x_*|$ that

$$e_{k+1}\le e_k^2$$

If $e_0\le q<1$ we therefore obtain $e_1\le q^2$, $e_2\le q^4$, $e_3\le q^8,\ldots$

$$e_k\le q^{(2^k)}$$

This means that the error converges to zero as $k\to\infty$, and we obtain the following theorem:

Theorem 2.3. *Assume that $f(x_*)=0$ and*

- *$f'(x)$ and $f''(x)$ exist and are continuous near x_**

- $f'(x_*) \neq 0$.

Then there exists $\delta > 0$, $C > 0$ *such that for* $|x_0 - x_*| \leq \delta$ we have

- $\lim_{k\to\infty} x_k = x_*$ *(convergence)*
- $|x_{k+1} - x_*| \leq C\,|x_k - x_*|^2$ *(convergence of order 2)*

Proof. Exactly like the proof of Theorem 2.2. □

3 Nonlinear system

We have n nonlinear equations $f_1(x_1,\ldots,x_n) = 0,\ldots, f_n(x_1,\ldots,x_n) = 0$. We define the vector-valued function $f(x)$ as

$$f(x) = \begin{bmatrix} f_1(x_1,\ldots,x_n) \\ \vdots \\ f_n(x_1,\ldots,x_n) \end{bmatrix}$$

The Jacobian $f'(x)$ (often denoted by $Df(x)$) is the $n \times n$ matrix of first partial derivatives

$$f'(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Then Taylor's theorem for functions $g(x_1,\ldots,x_n)$ gives that

$$f(x) = \underbrace{f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)})}_{p(x)} + R(x)$$

We assume that the second order partial derivatives $\dfrac{\partial^2 f_i}{\partial x_j \partial x_k}(x)$ exist and are continous. Then the remainder term $R(x) = f(x) - p(x)$ satisfies

$$\|R(x)\|_\infty \leq C \left\| x - x^{(0)} \right\|_\infty^2$$

If $\left|\frac{\partial^2 f_i}{\partial x_j \partial x_k}(x)\right| \leq c_2$ for $i,j,k = 1,\ldots,n$ we obtain $C = n^2 c_2$.

We start with an inital guess $x^{(0)}$. Then we approximate the function $f(x)$ by the Taylor approximation $p(x) = b + A(x - x^{(0)})$ with $b := f(x^{(0)})$ and $A := f'(x^{(0)})$. Instead of $f(x) = \vec{0}$ we solve solve $p(x) = \vec{0}$ as follows: Let $d = x - x^{(0)}$, solve the linear system $Ad = -b$, then let $x^{(1)} := x^{(0)} + d$.

Algorithm: Newton Method

- Initial guess $x^{(0)}$
- For $k = 0,1,2,\ldots$:
  $b := f(x^{(k)})$
  $A := f'(x^{(k)})$
  solve $Ad = -b$ for d (use Gaussian elimination with pivoting)
  $x^{(k+1)} := x^{(k)} + d$

Let us investigate the errors. For $p(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)})$ Taylor's theorem gives for $x = x^*$

$$f(x^*) - p(x^*) = R(x^*)$$

Since $f(x^*) = \vec{0} = p(x^{(k+1)})$ we get

$$p(x^{(k+1)}) - p(x^*) = R(x^*)$$

From $p(x) = b + A(x - x^{(0)})$ we get $p(x^{(k+1)}) - p(x^*) = A\left(x^{(k+1)} - x^*\right)$ so that

$$x^{(k+1)} - x^* = A^{-1}R(x^*)$$
$$\left\|x^{(k+1)} - x^*\right\| \le \left\|A^{-1}\right\| \|R(x^*)\|$$
$$\left\|x^{(k+1)} - x^*\right\| \le \left\|A^{-1}\right\| C \left\|x^{(k)} - x^*\right\|^2$$

If we have $\left\|f'(x)^{-1}\right\| \le c_1$ and $\left|\dfrac{\partial^2 f_i(x)}{\partial x_j \partial x_k}\right| \le c_2$ we get $C_2 = n^2 c_2$ and $D = c_1 n^2 c_2$ yielding

$$\left\|x^{(k+1)} - x^*\right\| \le D \left\|x^{(k)} - x^*\right\|^2$$

Therefore we obtain the following theorem:

Theorem 3.1. *Assume that $f(x^*) = 0$ and*

- $\dfrac{\partial f_i}{\partial x_j}$ *and* $\dfrac{\partial^2 f_i}{\partial x_j \partial x_k}$ *exist and are continuous near x_* for $i, j, k = 1, \ldots, n$*
- *the matrix $f'(x^*)$ is nonsingular.*

Then there exists $\delta > 0$, $C > 0$ such that for $\left\|x^{(0)} - x_*\right\| \le \delta$ we have

- $\lim_{k\to\infty} x^{(k)} = x^*$ *(convergence)*
- $\left\|x^{(k+1)} - x^*\right\| \le C \left\|x^{(k)} - x^*\right\|^2$ *(convergence of order 2)*

Proof. Since $f'(x^*)$ is nonsingular and $f'(x)$ is continuous, we can find $\varepsilon > 0$ such that on $B_\varepsilon := \{x \mid \|x - x^*\| \le \varepsilon\}$ we have

$$\left\|f'(x)^{-1}\right\| \le c_1.$$

We can then determine c_2 such that $\left|\dfrac{\partial^2 f_i(x)}{\partial x_j \partial x_k}\right| \le c_2$ on B_ε. Then we have for $x^{(k)} \in B_\varepsilon$ that $\left\|x^{(k+1)} - x^*\right\| \le D \left\|x^{(k)} - x^*\right\|^2$. Now we proceed exactly as in the proof of Theorem 2.3. □

4 Nonlinear least squares problem

We have N functions $f_1(x_1, \ldots, x_n), \ldots, f_N(x_1, \ldots, x_n)$ for n unknowns with $N > n$. We define the vector-valued function $f(x)$ as

$$f(x) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_N(x_1, \ldots, x_n) \end{bmatrix}$$

We cannot expect to find $x \in \mathbb{R}^n$ such that $f(x) = \vec{0}$ since we have more equations than unknowns. But we can try to find $x \in \mathbb{R}^n$ such that the vector $f(x)$ becomes "as small as possible":

$$\text{Find } x \in \mathbb{R}^n \text{ such that } \|f(x)\|_2 \text{ is minimal}$$

The Jacobian $f'(x)$ (often denoted by $Df(x)$) is the $N \times n$ matrix (more rows than columns) of first partial derivatives

$$f'(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_N}{\partial x_1} & \cdots & \frac{\partial f_N}{\partial x_n} \end{bmatrix}$$

We start with an inital guess $x^{(0)}$. Then we approximate the function $f(x)$ by the Taylor approximation $p(x) = b + A(x - x^{(0)})$ with $b := f(x^{(0)})$ and $A := f'(x^{(0)})$. Instead of $\|f(x)\| = \min$ we solve $\|p(x)\| = \min$ as follows: Let $d = x - x^{(0)}$, solve the linear least squares problem $\|Ad + b\| = \min$, then let $x^{(1)} := x^{(0)} + d$.

Algorithm: Gauss-Newton Method

- Initial guess $x^{(0)}$
- For $k = 0, 1, 2, \ldots$:
  $b := f(x^{(k)})$
  $A := f'(x^{(k)})$
  find d such that $\|Ad + b\|$ is minimal (use normal equations or QR decomposition)
  $x^{(k+1)} := x^{(k)} + d$

Convergence of the Gauss-Newton method: We assume that $F(x) := \|f(x)\|_2^2 = f_1(x)^2 + \cdots + f_N(x)^2$ has a local minimum at $x^* \in \mathbb{R}^n$. Therefore $\frac{\partial F}{\partial x_j}(x^*) = 0$ for $i = 1, \ldots, n$, i.e., with $A_* := f'(x^*)$ we have the normal equations

$$A_*^\top f(x^*) = \vec{0} \tag{4}$$

If our current approximation is $x^{(k)}$ we consider the Taylor approximation $p(x) = b + A(x - x^{(k)})$ with $b = f(x^{(k)})$ and $A = f'(x^{(k)})$. Then we determine $x^{(k+1)}$ such that $\left\|p(x^{(k+1)})\right\|_2$ is minimal, hence we have the normal equations

$$A^\top p(x^{(k+1)}) = \vec{0} \tag{5}$$

For the Taylor approximation we know that

$$f(x^*) - p(x^*) = r(x^*), \qquad \|r(x^*)\| \le C_2 \left\|x^* - x^{(k)}\right\|^2 \tag{6}$$

where C_2 depends on the size of the second order partial derivatives $\frac{\partial^2 f_i}{\partial x_j \partial x_k}$. We also have

$$\|A_* - A\| = \left\|f'(x^*) - f'(x^{(k)})\right\| \le C_2 \left\|x^* - x^{(k)}\right\|$$

From (6) we obtain

$$A^\top f(x^*) - A^\top p(x^*) = A^\top r(x^*)$$

Now (4), (5) give for the first term

$$\begin{aligned} A^\top f(x^*) &= \underbrace{A_*^\top f(x^*)}_{0} + (A - A_*)^\top f(x^*) \\ &= A^\top p(x^{(k+1)}) + (A - A_*)^\top f(x^*) \end{aligned}$$

yielding

$$A^\top \underbrace{\left(p(x^{(k+1)}) - p(x^*)\right)}_{A\left(x^{(k+1)} - x^*\right)} = (A_* - A)^\top f(x^*) + A^\top r(x^*)$$

and

$$x^{(k+1)} - x^* = (A^\top A)^{-1}(A_* - A)^\top f(x^*) + (A^\top A)^{-1} A^\top r(x^*)$$

$$\left\|x^{(k+1)} - x^*\right\| \le \left\|(A^\top A)^{-1}\right\| \left(C_2 \left\|x^{(k)} - x^*\right\| \|f(x^*)\| + \left\|A^\top\right\| C_2 \left\|x^{(k)} - x^*\right\|^2\right)$$

$$\left\|x^{(k+1)} - x^*\right\| \le D\left(c\,\|f(x^*)\| \cdot \left\|x^{(k)} - x^*\right\| + \left\|x^{(k)} - x^*\right\|^2\right)$$

with $D := C_2 \left\|(A^\top A)^{-1}\right\| \left\|A^\top\right\|$. If the residual $\|f(x^*)\|$ is zero (usually not satisfied) we get quadratic convergence. If $\varepsilon := c\,\|f(x^*)\|$ is small the error $\left\|x^{(k)} - x^*\right\|$ will at first decrease as with quadratic convergence, until $\left\|x^{(k)} - x^*\right\| \approx \varepsilon$. From then on we will **only have convergence of order 1** (if the residual $\|f(x^*)\|$ is sufficiently small). If the residual $\|f(x^*)\|$ is too large the **Gauss-Newton method may not be locally convergent**.