1 Fixed Point Iteration and Contraction Mapping Theorem

Notation: For two sets A, B we write $A \subset B$ iff $x \in A \implies x \in B$. So $A \subset A$ is true. Some people use the notation "$\subseteq$" instead.

1.1 Introduction

Consider a function $y = g(x)$ where $x, y \in \mathbb{R}^n$:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} g_1(x_1, \ldots, x_n) \\ \vdots \\ g_n(x_1, \ldots, x_n) \end{bmatrix}$$

We assume that $g(x)$ is defined for $x \in D$ where D is a subset of $\mathbb{R}^n$.

The goal is to find a solution x^* of the **fixed point equation**

$$g(x) = x.$$

A method to find x^* is the **fixed point iteration**: Pick an initial guess $x^{(0)} \in D$ and define for $k = 0, 1, 2, \ldots$

$$x^{(k+1)} := g(x^{(k)})$$

Note that this may not converge. But if the sequence $x^{(k)}$ converges, and the function g is continuous, the limit x^* must be a solution of the fixed point equation.

1.2 Contraction Mapping Theorem

The following theorem is called **Contraction Mapping Theorem** or **Banach Fixed Point Theorem**.

Theorem 1. *Consider a set $D \subset \mathbb{R}^n$ and a function $g\colon D \to \mathbb{R}^n$. Assume*

1. *D is closed (i.e., it contains all limit points of sequences in D)*
2. *$x \in D \implies g(x) \in D$*
3. *The mapping g is a* contraction *on D: There exists $q < 1$ such that*

$$\forall x, y \in D: \qquad \|g(x) - g(y)\| \le q \|x - y\| \tag{1}$$

Then

1. *there exists a unique $x^* \in D$ with $g(x^*) = x^*$*
2. *for any $x^{(0)} \in D$ the fixed point iterates given by $x^{(k+1)} := g(x^{(k)})$ converge to x^* as $k \to \infty$*
3. *$x^{(k)}$ satisfies the **a-priori error estimate***

$$\left\|x^{(k)} - x^*\right\| \le \frac{q^k}{1-q} \left\|x^{(1)} - x^{(0)}\right\| \tag{2}$$

*and the **a-posteriori error estimate***

$$\left\|x^{(k)} - x^*\right\| \le \frac{q}{1-q} \left\|x^{(k)} - x^{(k-1)}\right\| \tag{3}$$

Proof. Pick $x^{(0)} \in D$ and define $x^{(k)}$ for $k = 1,2,\ldots$ by $x^{(k)} := g(x^{(k-1)})$. We have from the contraction property (1)

$$\left\|x^{(k+1)} - x^{(k)}\right\| = \left\|g(x^{(k)}) - g(x^{(k-1)})\right\| \le q\left\|x^{(k)} - x^{(k-1)}\right\| \tag{4}$$

and hence

$$\left\|x^{(k+1)} - x^{(k)}\right\| \le q^k \left\|x^{(1)} - x^{(0)}\right\| \tag{5}$$

Let $d := \left\|x^{(1)} - x^{(0)}\right\|$. We have from the triangle inequality and (5)

$$\begin{aligned}\left\|x^{(k)} - x^{(k+\ell)}\right\| &\le \left\|x^{(k)} - x^{(k+1)}\right\| + \cdots + \left\|x^{(k+\ell-1)} - x^{(k+\ell)}\right\| \\ &\le q^k d + \cdots + q^{k+\ell-1} d = q^k d\,(1 + q + \cdots + q^{\ell-1}) \\ \left\|x^{(k)} - x^{(k+\ell)}\right\| &\le q^k d \frac{1}{1-q} \end{aligned} \tag{6}$$

using the sum of the geometric series $\sum_{j=0}^{\ell-1} q^j \le \sum_{j=0}^{\infty} q^j = 1/(1-q)$. Note that (6) shows that the sequence $x^{(k)}$ is a *Cauchy sequence*. Therefore it must converge to a limit $x^* \in \mathbb{R}^n$ (since the space $\mathbb{R}^n$ is complete). As D is closed, we must have $x^* \in D$.

We need to show that $x^* = g(x^*)$: We have $x^{(k+1)} = g(x^{(k)})$, hence

$$\lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} g(x^{(k)})$$

The limit of the left hand side is x^*. Note that because of (1) the function g must be continuous. Therefore

$$\lim_{k\to\infty} g(x^{(k)}) = g(\lim_{k\to\infty} x^{(k)}) = g(x^*).$$

Next we need to show that the fixed point x^* is unique. Assume that we have fixed points $x^* = g(x^*)$ and $y^* = g(y^*)$. Then we obtain using the contraction property (1)

$$\left\|x^* - y^*\right\| = \left\|g(x^*) - g(y^*)\right\| \le q\left\|x^* - y^*\right\|$$

implying $(1-q)\left\|x^* - y^*\right\| \le 0$ and therefore $\left\|x^* - y^*\right\| = 0$, i.e., $x^* = y^*$.

The a-priori estimate (2) follows from (6) by letting ℓ tend to infinity. For the a-posteriori estimate use (2) with $k = 1$ for $\tilde{x}^{(0)} := x^{(k)}$, $\tilde{x}^{(1)} = x^{(k+1)}$. □

1.3 Proving the Contraction Property

The contraction property is related to the Jacobian $g'(x)$ which is an $n \times n$ matrix for each point $x \in D$. If the matrix norm satisfies $\|g'(x)\| \le q < 1$ then the mapping g must be a contraction:

Theorem 2. *Assume the set $D \subset \mathbb{R}^n$ is convex and the function $g\colon D \to \mathbb{R}^n$ has continuous partial derivatives $\dfrac{\partial g_j}{\partial k}$ in D. If for $q < 1$ the matrix norm of the Jacobian satisfies*

$$\forall x \in D\colon \qquad \left\|g'(x)\right\| \le q \tag{7}$$

the mapping g is a contraction in D and satisfies (1).

Proof. Let $x, y \in D$. Then the points on the straight line from x to y are given by $x + t(y-x)$ for $t \in [0,1]$. As D is convex all these points are contained in D. Let $G(t) := g\big(x + t(y-x)\big)$, then by the chain rule we have $G'(t) = g'\big(x + t(y-x)\big)\,(y-x)$ and

$$g(y) - g(x) = G(1) - G(0) = \int_0^1 G'(t)dt = \int_0^1 g'\big(x + t(y-x)\big)\,(y-x)\,dt$$

As an integral of a continuous function is a limit of Riemann sums the triangle inequality implies $\left\|\int_a^b F(t)dt\right\| \le \int_a^b \|F(t)\|\,dt$:

$$\|g(y) - g(x)\| \le \int_0^1 \left\|g'\big(x + t(y-x)\big)\,(y-x)dt\right\| \le \int_0^1 \underbrace{\left\|g'\big(x + t(y-x)\big)\right\|}_{\le q} \|y-x\|\,dt \le q\,\|y-x\|$$

□

This is usually the easiest method to prove that a given mapping g is a contraction, see the examples in sections 1.5, 1.6.

1.4 A-priori and a-posteriori error estimates

The error estimates (2), (3) are useful for figuring out how many iterations we need. For this we need to know the contraction constant q (typically we get this from (7)).

A-priori estimate: For an initial guess $x^{(0)}$ we can find $x^{(1)}$. Without computing anything else we then have the error bound $\|x^{(k)}-x^*\| \le \frac{q^k}{1-q}\|x^{(1)}-x^{(0)}\|$ for all future iterates $x^{(k)}$, before ("a-priori") we actually compute them. We can e.g. use this to find a value k such that $\|x^{(k)}-x^*\|$ is below a given tolerance.

A-posteriori estimate: After we have actually computed $x^{(k)}$ ("a-posteriori") we would like to know where the true solution x^* is located. Let

$$\delta_k := \frac{q}{1-q}\|x^{(k)}-x^{(k-1)}\|, \qquad D_k := \{x \mid \|x-x^{(k)}\| \le \delta_k\}$$

The a-posteriori estimate states that x^* is contained in the set D_k. Note:

- the "radius" δ_k of D_k decreases at least by a factor of q with each iteration: $\delta_{k+1} \le q\delta_k$
- the sets D_k are nested: $D_1 \supset D_2 \supset D_3 \supset \cdots$

To show $D_{k+1} \subset D_k$ assume $x \in D_{k+1}$. Then

$$\|x-x^{(k)}\| \le \underbrace{\|x-x^{(k+1)}\|}_{\le \delta_{k+1}} + \|x^{(k+1)}-x^{(k)}\| \le \left(\frac{q}{1-q}+1\right)\|x^{(k+1)}-x^{(k)}\| \overset{(4)}{\le} \frac{1}{1-q}\,q\|x^{(k)}-x^{(k-1)}\| = \delta_k \tag{8}$$

If we use the ∞-norm: $\|x^{(k)}-x^*\|_\infty \le \delta_k$ means that for each component x_j^* we have a bracket

$$x_j^* \in [x_j^{(k)}-\delta_k, x_j^{(k)}+\delta_k],$$

i.e., the set D_k is a square/cube/hypercube with side length $2\delta_k$ centered in $x^{(k)}$.

1.5 Example

We want to find x_1, x_2 satisfying the nonlinear system

$$10x_1 + x_2 + \sin(x_1+x_2) = 1 \tag{9}$$
$$x_1 + 10x_2 - \cos(x_1-x_2) = 2 \tag{10}$$

We first have to rewrite this system in fixed point form $x = g(x)$. If we solve the first equation for x_1 in $10x_1$, and we solve the second equation for x_2 in $10x_2$ we get the following system

$$x_1 = \tfrac{1}{10}\left[1 - x_2 - \sin(x_1+x_2)\right] \tag{11}$$
$$x_2 = \tfrac{1}{10}\left[2 + x_1 + \cos(x_1-x_2)\right] \tag{12}$$

This is in fixed point form $x = g(x)$ with $g(x) = \frac{1}{10}\begin{bmatrix} 1-x_2-\sin(x_1+x_2) \\ 2+x_1+\cos(x_1-x_2) \end{bmatrix}$.

First we want to show that g is a contraction using Theorem 2. Therefore we first have to find the Jacobian $g'(x)$:

$$g'(x) = \tfrac{1}{10}\begin{bmatrix} -\cos(x_1+x_2) & -1-\cos(x_1+x_2) \\ 1-\sin(x_1-x_2) & \sin(x_1-x_2) \end{bmatrix}$$

Let $A := g'(x)$. Let us use the ∞-norm. We need to find an upper bound for $\|A\|_\infty = \max\{|a_{11}|+|a_{12}|, |a_{21}|+|a_{22}|\}$. We obtain for any $x_1, x_2 \in \mathbb{R}$

$$|a_{11}| = \tfrac{1}{10}|-\cos(x_1+x_2)| \le \tfrac{1}{10}, \qquad |a_{12}| = \tfrac{1}{10}|-1-\cos(x_1+x_2)| \le \tfrac{1}{10}(1+1)$$
$$|a_{21}| = \tfrac{1}{10}|1-\sin(x_1-x_2)| \le \tfrac{1}{10}(1+1), \qquad |a_{22}| \le \tfrac{1}{10}|\sin(x_1-x_2)| \le \tfrac{1}{10}$$

Therefore for any $x \in \mathbb{R}^2$ we have

$$\left\|g'(x)\right\|_\infty \le \frac{3}{10} = q < 1.$$

By Theorem 2 we therefore obtain that g is a contraction for all of $\mathbb{R}^2$.

We now want to use Theorem 1. We need to pick a set D such that the three assumptions of the theorem are satisfied. We consider two choices:

First choice $D = \mathbb{R}^2$: We can use the set $D = \mathbb{R}^2$. This set is closed. For any $x \in \mathbb{R}^2$ we certainly have that $g(x) \in \mathbb{R}^2$. We have also shown that g is a contraction for all of $\mathbb{R}^2$. Therefore we obtain from Theorem 1 that the nonlinear system $g(x) = x$ has exactly one solution x^* in all of $\mathbb{R}^2$.

Second choice $D = [-1,1]\times[-1,1]$: We can use for D the square with $-1 \le x_1 \le 1$ and $-1 \le x_2 \le 1$. This is a closed set (the boundary of the square is included). We now have to check that for $x \in D$ we have that $y = g(x) \in D$: We have using $-1 \le \sin\alpha \le 1$, $-1 \le \cos\alpha \le 1$

$$-\tfrac{2}{10} = \tfrac{1}{10}(1-1-1) \le y_1 = \tfrac{1}{10}[1-x_2-\sin(x_1+x_2)] \le \tfrac{1}{10}(1+1+1) = \tfrac{3}{10}$$
$$0 = \tfrac{1}{10}(2-1-1) \le y_2 = \tfrac{1}{10}[2+x_1+\cos(x_1-x_2)] \le \tfrac{1}{10}(2+1+1) = \tfrac{4}{10}$$

therefore $y \in D$ and the second assumption of the theorem is satisfied. We already showed that g is a contraction for all of $\mathbb{R}^2$, so the third assumption definitely holds for $x, y \in D$. We can now apply Theorem 1 and obtain that the nonlinear system has exactly one solution x^* which is located in the square $D = [-1,1]\times[-1,1]$.

Numerical Computation: We start with the initial guess $x^{(0)} = (0,0)^\top$. After each iteration we find δ_k and the square D_k containing x^*:

k	$x^{(k)}$	δ_k	D_k
1	$(.1,\ .3)^\top$	$1.3\cdot10^{-1}$	$[-.02857,\ .2286]\times[.1714,\ .4286]$
2	$(.03106,\ .3080)^\top$	$3.0\cdot10^{-2}$	$[.00151,\ .06060]\times[.2785,\ .3376]$
3	$(.03594,\ .2993)^\top$	$3.7\cdot10^{-3}$	$[.03221,\ .03967]\times[.2956,\ .3030]$
4	$(.03717,\ .3001)^\top$	$5.3\cdot10^{-4}$	$[.03664,\ .03770]\times[.2996,\ .3007]$
5	$(.03689,\ .3003)^\top$	$1.2\cdot10^{-4}$	$[.03677,\ .03701]\times[.3001,\ .3004]$

Note: (i) δ_k decreases at least by a factor of $q = 0.3$ with each iteration.

(ii) The sets D_k are nested: $D_1 \supset D_2 \supset D_3 \supset \cdots$

1.6 Using the Fixed Point Theorem *without* the Assumption $g(D) \subset D$

The tricky part in using the contraction mapping theorem is to find a set D for which *both* the 2nd and 3rd assumption of the fixed point theorem hold:

- $x \in D \implies g(x) \in D$
- g is a contraction on D

Typically we can prove that $\|g'(x)\| \le q < 1$ for x in some convex region $\tilde{D}$. We suspect that there is a solution x^* of the fixed point equation in $\tilde{D}$. But it may not be true that $g(x) \in \tilde{D}$ for all $x \in \tilde{D}$.

In this case we may be able to prove a result by computing a few iterates $x^{(k)}$: Start with $k = 0$ and an initial guess $x^{(0)} \in \tilde{D}$. Then repeat

- let $k := k+1$ and compute $x^{(k)} := g(x^{(k-1)})$
- compute $\delta_k := \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\|$, let $D_k := \{x \mid \|x - x^{(k)}\| \le \delta_k\}$

until either $D_k \subset \tilde{D}$ or $x^{(k)} \notin \tilde{D}$.

If the iterates exit from the set $\tilde{D}$ we cannot conclude anything. But as long as the points $x^{(k)}$ stay inside $\tilde{D}$ we have $\delta_{k+1} \le q\delta_k$ and $D_{k+1} \subset D_k$. So we expect that for some k the condition $D_k \subset \tilde{D}$ will be satisfied (if $x^{(k)}$converges to a limit in the interior of $\tilde{D}$ the loop must terminate with $D_k \subset \tilde{D}$; but in general it is possible that the loop never terminates). If the loop does terminate with $D_k \subset \tilde{D}$ for $k = K$ we have the following result:

Theorem 3. *Let $\tilde{D} \subset \mathbb{R}^n$ and assume that the function $g\colon \tilde{D} \to \mathbb{R}^n$ satisfies for $q < 1$*

$$\forall x, y \in \tilde{D}: \qquad \|g(x) - g(y)\| \le q\,\|x - y\|$$

Let $x^{(0)} \in \tilde{D}$ and define for $k = 0, 1, 2, \ldots$

$$x^{(k+1)} := g(x^{(k)}), \qquad \delta_k := \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\|, \qquad D_k := \{x \mid \|x - x^{(k)}\| \le \delta_k\} \tag{13}$$

If for some K we have $x^{(K-1)} \in \tilde{D}$ and $D_K \subset \tilde{D}$ there holds

- *the equation $g(x) = x$ has a unique solution x^* in $\tilde{D}$*
- *this solution satisfies $x^* \in D_k$ for all $k \ge K$*

Proof. Let $x \in D_K$. We want to show that $g(x) \in D_K$: As $D_K \subset \tilde{D}$ the contraction property gives using the definition of D_k and δ_k

$$\|g(x) - x^{(K)}\| \le q\|x - x^{(K-1)}\| \le q\|x - x^{(K)}\| + q\|x^{(K)} - x^{(K-1)}\| \le q\delta_K + (1-q)\delta_K = \delta_K$$

As D_K is closed and $D_K \subset \tilde{D}$ the set $D := D_K$ satisfies all three assumptions of the fixed point theorem Theorem 1. Hence there is a unique solution $x^* \in D$. The a-posteriori estimate (3) states that $x^* \in D_k$ for all iterates $x^{(k)}$ with $k \ge K$. Assume that there is another fixed point $y^* \in \tilde{D}$ with $g(y^*) = y^*$. Then

$$\|y^* - x^*\| = \|g(y^*) - g(x^*)\| \le q\,\|y^* - x^*\|$$

As $q < 1$ we must have $\|y^* - x^*\| = 0$. □

Summary:

- Find a convex set $\tilde{D}$ for which you suspect $x^* \in \tilde{D}$ and where you can show $\|g'(x)\| \le q < 1$
- Pick $x^{(0)} \in \tilde{D}$ and perform the fixed point iteration:
  for each iteration:
  - find $x^{(k)}$ and D_k
  - if $x^{(k)} \notin \tilde{D}$: stop (we can't conclude anything)
  - if $D_k \subset \tilde{D}$: success: there is a unique solution $x^* \in \tilde{D}$, and there holds $x^* \in D_k$ for this and all following iterations

Example: Let $g(x) := \frac{1}{3}\begin{bmatrix} x_1 - x_1x_2 + 1 \\ x_2 + x_1x_2^2 + 1 \end{bmatrix}$. Then the Jacobian is $g'(x) = \frac{1}{3}\begin{bmatrix} 1 - x_2 & -x_1 \\ x_2^2 & 1 + 2x_1x_2 \end{bmatrix}$.
Let us try to use $\tilde{D} = [0, a] \times [0.a]$ with $a \le 1$ and the ∞-norm. We then obtain for $x \in \tilde{D}$ that

$$\|g'(x)\|_\infty \le \tfrac{1}{3}\max\{1 + a, a^2 + 1 + 2a^2\}$$

For $a = 1$ we get $\|g'(x)\|_\infty \le \frac{4}{3}$ which is too large. So we try $a = 0.6$ which gives $\|g'(x)\|_\infty \le \frac{2.08}{3} =: q < 1$. Therefore g is a contraction on $\tilde{D} = [0, .6] \times [0, .6]$. Note that $g(\begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}) = \begin{bmatrix} 0.41333 \\ 0.60533 \end{bmatrix} \notin \tilde{D}$, so $\tilde{D}$ does *not* satisfy all three assumptions of Theorem 1.

For $x^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ we obtain

$$\begin{aligned}
x^{(1)} &= (.33333,\ .33333)^\top \in \tilde{D}, & D_1 &= [-0.42029,\ 1.08696] \times [-0.42029,\ 1.08696] \not\subset \tilde{D} \\
x^{(2)} &= (.40741,\ .45679)^\top \in \tilde{D}, & D_2 &= [0.12829,\ 0.68653] \times [0.17767,\ 0.73591] \not\subset \tilde{D} \\
x^{(3)} &= (.40710,\ .51393)^\top \in \tilde{D}, & D_3 &= [0.27791,\ 0.53629] \times [0.38474,\ 0.64313] \not\subset \tilde{D} \\
x^{(4)} &= (.39929,\ .54049)^\top \in \tilde{D}, & D_4 &= [0.33926,\ 0.45933] \times [0.48045,\ 0.60052] \not\subset \tilde{D} \\
x^{(5)} &= (.39449,\ .55238)^\top \in \tilde{D}, & D_5 &= [0.36761,\ 0.42138] \times [0.52549,\ 0.57926] \subset \tilde{D}
\end{aligned}$$

Therefore we can conclude from Theorem 3 that there exists a unique solution $x^* \in \tilde{D} = [0,0.6] \times [0,0.6]$. This solution x^* is located in the smaller square D_5. For $k = 5,6,7,\ldots$ we obtain $x^* \in D_k$ where D_k is a square with side length $2\delta_k$. As $\delta_k \le q^{k-5}\delta_5 \le \left(\frac{2.08}{3}\right)^{k-5} 0.027$ we can obtain arbitrarily small squares containing the solution if we choose k sufficiently large.

1.7 How to rewrite a nonlinear system in fixed point form: simplified Newton method

For a given nonlinear system $f(x) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$ there are several ways to rewrite it in the form $g(x) = x$. How can we do this so that the function g is a contraction?

First consider the **Newton method**: For a current guess $x^{(k)}$ we solve the linear system

$$f'(x^{(k)})d = -f(x^{(k)}) \tag{14}$$

and let $x^{(k+1)} := x^{(k)} + d = x^{(k)} - f'(x^{(k)})^{-1} f(x^{(k)})$. Therefore we have the iteration function $\boxed{g(x) = x - f'(x)^{-1} f(x)}$.

The Newton method is expensive: For each step we have to evaluate $f(x^{(k)})$ and $f'(x^{(k)})$ which are $n+n^2$ function evaluations. Then we have to solve an $n \times n$ linear system which costs $\frac{n^3}{3} + O(n^2)$ operations. In particular for large n this may be too much work per step.

We can "cheat" as follows: instead of solving (14) with the new Jacobian matrix $f'(x^{(k)})$ we keep reusing the first Jacobian $\boxed{A := f'(x^{(0)})}$ and solve

$$Ad = -f(x^{(k)}) \tag{15}$$

and let $x^{(k+1)} := x^{(k)} + d = x^{(k)} - A^{-1} f(x^{(k)})$. Therefore we have now the iteration function $\boxed{g(x) = x - A^{-1} f(x)}$. This "**simplified Newton method**" is much cheaper: We only evaluate the initial Jacobian matrix $A := f'(x^{(0)})$ and compute its LU decomposition. For each step we can solve the linear system (15) using forward and back substitution. Hence we have for each step only n function evaluations and n^2 operations to solve the linear system.

The simplified Newton method is still locally convergent. But it only converges with order 1 (whereas the Newton method converges with order 2). This is shown in the following theorem:

Theorem 4. *Assume we have $f(x^*) = 0$ where $f'(x^*)$ is nonsingular and the partial derivatives $\frac{\partial f_i}{\partial x_j}$ and $\frac{\partial^2 f_i}{\partial x_j \partial x_k}$ are continuous near x^*. Then the simplified Newton method will converge if $x^{(0)}$ is sufficiently close to a solution x^*.*

Proof. Assume we have a solution x^*, and for $\left\|x - x^{(0)}\right\|_\infty \le \varepsilon$ we have bounds $\left|\frac{\partial^2 f_i(x)}{\partial x_j \partial x_k}\right| \le M_{ijk}$ and $\left\|f'(x)^{-1}\right\|_\infty \le c_1$. Let $c_2 := \max_i \sum_{j,k} M_{ijk}$. Then $\|f'(x) - f'(y)\|_\infty \le c_2 \|x - y\|_\infty$. Now assume $\|x - x^*\| \le \delta$ and $\left\|x^{(0)} - x^*\right\| \le \delta$ (we will define δ in a moment). Then

$$\begin{aligned}
g'(x) &= I - A^{-1} f'(x) = A^{-1}\left(f'(x^{(0)}) - f'(x)\right) \\
\left\|g'(x)\right\|_\infty &\le \left\|f'(x^{(0)})\right\| \left\|f'(x^{(0)}) - f'(x)\right\| \le c_1 c_2 \underbrace{\left\|x^{(0)} - x\right\|}_{\left\|x^{(0)} - x^*\right\| + \|x^* - x\|} \le c_1 c_2 2\delta
\end{aligned}$$

Now we can define $\delta > 0$ such that $\delta \le \varepsilon$ and $c_1 c_2 2\delta \le q < 1$. Hence for $D := \{x \mid \|x - x^*\| \le \delta\}$ we have that D is closed and that g is a contraction on D with $q < 1$. We also have for $x \in D$ that $\|g(x) - x^*\| = \|g(x) - g(x^*)\| \le q\|x - x^*\| \le q\delta \le \delta$, hence $g(x) \in D$. □

We can now use the contraction mapping theorem to show that there exists a unique solution in a region around our initial guess $x^{(0)}$.

We assume: Near our initial guess $x^{(0)}$ we have bounds β_{ijk} for the 2nd partial derivatives:

$$\left|\frac{\partial^2 f_i(x)}{\partial x_j \partial x_k}\right| \le \beta_{ijk} \qquad \text{for } \left\|x - x^{(0)}\right\|_\infty \le R$$

Then we have for x with $\left\|x - x^{(0)}\right\|_\infty \le R$

$$\left|\frac{\partial f_i}{\partial x_j}(x^{(0)}) - \frac{\partial f_i}{\partial x_j}(x)\right| \le \underbrace{\left(\sum_{k=1}^{n} \beta_{ijk}\right)}_{=: B_{ij}} \left\|x^{(0)} - x\right\|_\infty$$

$$g'(x) = A^{-1}\left(f'(x^{(0)}) - f'(x)\right)$$

$$\left\|g'(x)\right\|_\infty \le \underbrace{\left\|\mathrm{abs}(A^{-1})B\right\|_\infty}_{=: M} \cdot \left\|x^{(0)} - x\right\|_\infty \tag{16}$$

where $\mathrm{abs}(A^{-1}) \in \mathbb{R}^{n\times n}$ has entries which are the absolute values of the entries of A^{-1} and $B \in \mathbb{R}^{n\times n}$ has entries B_{ij}. Here

We perform one Newton step: (the first step of the simplified Newton method is an actual Newton step)

$$d := -A_0^{-1} f(x^{(0)}), \qquad x^{(1)} := x^{(0)} + d, \qquad \delta := \|d\|_\infty$$

We want to use Theorem 3 on the region $\tilde{D}_r := \left\{x \mid \left\|x - x^{(0)}\right\|_\infty \le r\right\}$ with a suitable $r \le R$.

We need to choose $r \in (0, R]$ so that the assumptions of the theorem are satisfied:

- g' is a contraction for $x \in \tilde{D}_r$: Because of (16) we need $q := Mr \overset{!}{<} 1$
- the "a-posteriori region" $D_1 = \left\{x \mid \left\|x - x^{(1)}\right\|_\infty \le \frac{q}{1-q}\left\|x^{(1)} - x^{(0)}\right\|_\infty\right\}$ is a subset of $\tilde{D}_r$. This is satisfied if

$$\left\|x - x^{(0)}\right\|_\infty \le \left\|x - x^{(1)}\right\|_\infty + \left\|x^{(1)} - x^{(0)}\right\|_\infty \le \frac{q}{1-q}\delta + \delta \overset{!}{\le} r$$

$$\text{i.e.,} \quad \frac{\delta}{1-q} \overset{!}{\le} r \tag{17}$$

With $q := Mr < 1$ the condition (17) can be written as

$$\delta \overset{!}{\le} r(1 - Mr) =: F(r)$$

Note that the quadratic function $F(r) = r(1-Mr)$ is zero at $r = 0$ and $r = M^{-1}$ and has at $\frac{1}{2}M^{-1}$ the maximum $F(\frac{1}{2}M^{-1}) = \frac{1}{4}M^{-1}$. Hence we need $\delta \le \frac{1}{4}M^{-1}$. The smallest r with $F(r) \ge \delta$ is $r_1 := \frac{1}{2}M^{-1}\left(1 - \sqrt{1 - 4M\delta}\right)$. We have shown:

Theorem 5. *Assume that for* $\left\|x - x^{(0)}\right\|_\infty \le R$

- *the functions* $f(x)$, $\frac{\partial f_i}{\partial x_j}(x)$, $\frac{\partial^2 f_i}{\partial x_j \partial x_k}$ *are continuous*
- *we have bounds* $\left|\frac{\partial^2 f_i}{\partial x_j \partial x_k}(x)\right| \le \beta_{ijk}$

Let $A := f'(x^{(0)})$ *and let*

$$\delta := \left\| A^{-1} f(x^{(0)}) \right\|_\infty, \qquad B_{ij} := \sum_{k=1}^{n} \beta_{ijk}, \qquad M := \left\| \mathrm{abs}(A^{-1}) B \right\|_\infty, \qquad r_{1,2} := \frac{1 \mp \sqrt{1-4M\delta}}{2M}$$

If $\boxed{\delta \le \frac{1}{4} M^{-1}}$ *and* $R \ge r_1$

- *the problem* $f(x^*) = \vec{0}$ *has a unique solution* x^* *in the region* $\left\| x - x^{(0)} \right\|_\infty \le \min\{r_2, R\}$
- *for* $x^{(1)} := x^{(0)} - A^{-1} f(x^{(0)})$ *we have* $\left\| x^* - x^{(1)} \right\|_\infty \le \dfrac{\delta}{(Mr_1)^{-1} - 1} = \dfrac{4M\delta^2}{(1-4M\delta) + \sqrt{1-4M\delta}}$
- *the simplified Newton method starting with* $x^{(0)}$ *converges to* x^*.

Note that we do not assume that the problem $f(x) = \vec{0}$ has a solution.

Example 1: Find a zero of $f(x) = \begin{bmatrix} x_1 + \sin(x_1 + x_2) - \frac{1}{2} \\ x_2 + \cos(x_1 - x_2) - 1 \end{bmatrix}$. We have $f'(x) = \begin{bmatrix} 1 + \cos(x_1 + x_2), & \cos(x_1 + x_2) \\ -\sin(x_1 - x_2), & 1 + \sin(x_1 - x_2) \end{bmatrix}$. Note that all second derivatives $\dfrac{\partial^2 f_i}{\partial x_j \partial x_k}(x)$ are of the form $\pm\sin(\cdots)$ or $\pm\cos(\cdots)$, hence $\beta_{ijk} = 1$ for all i, j, k where R is arbitrarily large. We have $B_{ij} = \sum_{k=1}^{2} \beta_{ijk} = 2$ for all i, j.

First try $x^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$**:** Then $f(x^{(0)}) = \begin{bmatrix} -.5 \\ 0 \end{bmatrix}$, $A := f'(x^{(0)}) = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$, $A^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ 0 & 1 \end{bmatrix}$ and $d = -A^{-1} f(x^{(0)}) = \begin{bmatrix} \frac{1}{4} \\ 0 \end{bmatrix}$, $\delta = \|d\|_\infty = \frac{1}{4}$. Hence

$$\mathrm{abs}(A^{-1}) B = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}, \qquad M := \left\| \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \right\|_\infty = 4$$

But here $\boxed{\delta = \frac{1}{4}}$ is greater than $\boxed{\frac{1}{4} M^{-1} = \frac{1}{16}}$, so we cannot use the theorem. We need an initial guess closer to the solution. The Newton step starting at $x^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ gives $x^{(1)} = \begin{bmatrix} \frac{1}{4} \\ 0 \end{bmatrix}$.

Second try $x^{(0)} = \begin{bmatrix} \frac{1}{4} \\ 0 \end{bmatrix}$**:** Then $f(x^{(0)}) = \begin{bmatrix} -.002596 \\ -.031088 \end{bmatrix}$, $A := f'(x^{(0)}) = \begin{bmatrix} 1.9689 & .96891 \\ -.2474 & 1.2474 \end{bmatrix}$, $A^{-1} = \begin{bmatrix} .46273 & -.35942 \\ .091776 & .73038 \end{bmatrix}$ and $d = A^{-1} f(x^{(0)}) = \begin{bmatrix} -.0099723 \\ .022944 \end{bmatrix}$, $\delta = \|d\|_\infty = .022944$. Hence

$$\mathrm{abs}(A^{-1}) B = \begin{bmatrix} .46273 & .35942 \\ .091776 & .73038 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1.6443 & 1.6443 \\ 1.6443 & 1.6443 \end{bmatrix}, \qquad M := \left\| \mathrm{abs}(A^{-1}) B \right\|_\infty = 3.2886$$

Now $\boxed{\delta = .022944}$ is less than $\boxed{\frac{1}{4} M^{-1} = 0.07602}$, yielding $r_1 = .024999$, $r_2 = .27908$. Therefore the theorem states that there is a unique solution in the square $[-.0290799, .52908] \times [-.27908, .27908]$ of size $2r_2$, centered in $x^{(0)}$. This solution is actually located in the smaller square $[.237972, 0.242083] \times [0.0208887, 0.0249992]$ of size $2\frac{\delta}{(Mr_1)^{-1}-1}$, centered in $x^{(1)}$.

```
f = @(x) [ x(1)+sin(x(1)+x(2))-.5 ; x(2)+cos(x(1)-x(2))-1 ];
fp = @(x)[ 1+cos(x(1)+x(2)) , cos(x(1)+x(2)) ; -sin(x(1)-x(2)) , 1+sin(x(1)-x(2))];
B = [2 2;2 2];                          % from bounds beta_{ijk} for 2nd derivatives

x = [0;0];                              % initial guess
for i=1:5
  b = f(x); A = fp(x); Ai = inv(A);
  d = -Ai*b; delta = norm(d,Inf);
  x = x + d;
```

```
  fprintf('x = [%g,%g]\n',x);
  M = norm(abs(Ai)*B,Inf);
  if delta <= 1/(4*M)                    % condition for delta in theorem
    r1 = (1-sqrt(1-4*M*delta))/(2*M);    % condition for r1 is true since bounds hold everywhere
    e = delta/(1/(M*r1)-1);              % a-posteriori error bound
    fprintf('     inf-norm error <= %g\n',e)
  end
end
```

This prints

```
x = [0.25,0]
x = [0.240028,0.022944]
     inf-norm error <= 0.00205526
x = [0.239842,0.0233441]
     inf-norm error <= 5.39279e-07
x = [0.239842,0.0233442]
     inf-norm error <= 5.45924e-14
x = [0.239842,0.0233442]
     inf-norm error <= 5.60471e-28
```

Note that the actual errors $\left\|x^{(k)}-x^*\right\|_\infty$ for $k=2,3,4,5$ are $4\cdot 10^{-4}$, $1.3\cdot 10^{-7}$, $1.3\cdot 10^{-14}$, $8.3\cdot 10^{-17}$. The error bound $5.6\cdot 10^{-28}$ in the last case is for the exact $x^{(5)}$. The computed $x^{(5)}$ is affected by roundoff error of about 10^{-16}.

1.8 Newton-Kantorovich theorem

In Theorem 5 we used the convergence of order 1 for the simplified Newton method. One can prove a sharper result using the convergence of order 2 for the Newton method:

Theorem 6 (Newton-Kantorovich)**.** *Assume that for* $\left\|x-x^{(0)}\right\|_\infty \le R$

- *the functions* $f(x)$, $\frac{\partial f_i}{\partial x_j}(x)$, $\frac{\partial^2 f_i}{\partial x_j \partial x_k}$ *are continuous*
- *we have bounds* $\left|\frac{\partial^2 f_i}{\partial x_j \partial x_k}(x)\right| \le \beta_{ijk}$

Let $A := f'(x^{(0)})$ *and let*

$$\delta := \left\|A^{-1}f(x^{(0)})\right\|_\infty, \qquad B_{ij} := \sum_{k=1}^{n} \beta_{ijk}, \qquad M := \left\|\mathrm{abs}(A^{-1})B\right\|_\infty, \qquad r := \frac{1-\sqrt{1-2\delta M}}{M} = \frac{2\delta}{1+\sqrt{1-2\delta M}} \tag{18}$$

If $\boxed{\delta \le \tfrac{1}{2}M^{-1}}$ *and* $R \ge r$

- *the problem* $f(x^*)=\vec{0}$ *has a unique solution* x^* *in the region* $\left\|x-x^{(0)}\right\|_\infty \le r$
- *for* $x^{(1)} := x^{(0)} - A^{-1}f(x^{(0)})$ *we have* $\left\|x^*-x^{(1)}\right\|_\infty \le \dfrac{M\delta^2}{(1-\delta M)+\sqrt{1-2\delta M}}$
- *the Newton method starting with* $x^{(0)}$ *converges to* x^*

Example 1: For $f(x)=\begin{bmatrix} x_1+\sin(x_1+x_2)-\frac{1}{2} \\ x_2+\cos(x_1-x_2)-1 \end{bmatrix}$ with $x^{(0)}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ we get $\delta=\frac{1}{4}$ and $\frac{1}{2}M^{-1}=\frac{1}{8}$, so we cannot use Theorem 6. But with $x^{(0)}=\begin{bmatrix} \frac{1}{4} \\ 0 \end{bmatrix}$ we have $\delta \le \frac{1}{2}M^{-1}$ and we can use the theorem. If we modify the program from above to use the a-posteriori bound from the Newton-Kantorovich theorem we get for $x^{(k)}$, $k=2,\ldots,5$ the bounds $9.4\cdot 10^{-4}$, $2.7\cdot 10^{-7}$,

$2.7\cdot 10^{-14}$, $2.8\cdot 10^{-28}$. The actual errors were $4\cdot 10^{-4}$, $1.3\cdot 10^{-7}$, $1.3\cdot 10^{-14}$, $8.3\cdot 10^{-17}$ (this last value affected by roundoff error).

Example 2: Find a zero of $f(x)=\begin{bmatrix} 2x_1-x_2-x_1x_2-\frac{1}{2} \\ 2x_2+x_1(1-x_2^2)\end{bmatrix}$ in the square $[-\frac{1}{2},\frac{1}{2}]\times[-\frac{1}{2},\frac{1}{2}]$. We find the Jacobian $J(x)$, its partial derivatives $\left[\frac{\partial}{\partial x_1}J_{ij},\frac{\partial}{\partial x_2}J_{ij}\right]$ and bounds $\left|\frac{\partial}{\partial x_1}J_{ij}\right|+\left|\frac{\partial}{\partial x_2}J_{ij}\right|\le B_{ij}$

$$J(x)=\begin{bmatrix} 2-x_2, & -1-x_1 \\ 1-x_2^2, & 2-2x_1x_2\end{bmatrix},\qquad \begin{bmatrix} [0,-1] & [-1,0] \\ [0,-2x_2] & [-2x_2,-2x_1]\end{bmatrix},\qquad B=\begin{bmatrix} 1 & 1\\ 1 & 2\end{bmatrix}$$

We try $x^{(0)}=\begin{bmatrix}0\\0\end{bmatrix}$**:** We have $R=\frac{1}{2}$ and obtain $f(x^{(0)})=\begin{bmatrix}-.5\\0\end{bmatrix}$, $A:=f'(x^{(0)})=\begin{bmatrix}2&-1\\1&2\end{bmatrix}$, $A^{-1}=\begin{bmatrix}.4&.2\\-.2&.4\end{bmatrix}$ and $d=-A^{-1}f(x^{(0)})=\begin{bmatrix}.2\\-.1\end{bmatrix}$, $\delta=\|d\|_\infty=.2$. Hence

$$\mathrm{abs}(A^{-1})B=\begin{bmatrix}.4&.2\\.2&.4\end{bmatrix}\begin{bmatrix}1&1\\1&2\end{bmatrix}=\begin{bmatrix}.6&.8\\.6&1\end{bmatrix},\qquad M:=\left\|\begin{bmatrix}.6&.8\\.6&1\end{bmatrix}\right\|_\infty=1.6$$

Here $\boxed{\delta=.2\le\frac{1}{2}M^{-1}=.3125}$ and $R\ge r=.25$ so we can use the Newton-Kantorovich theorem (we cannot use Theorem 5 since $\delta=.2>\frac{1}{4}M^{-1}=.15625$). We obtain that there is a unique solution in the square $\left\|x-x^{(0)}\right\|_\infty\le r$, i.e., $[-.25,.25]\times[-.25,.25]$. For $x^{(1)}=x^{(0)}+d=\begin{bmatrix}.2\\-.1\end{bmatrix}$ we obtain the bound $\left\|x^{(1)}-x^*\right\|_\infty\le\frac{M\delta^2}{(1-\delta M)+\sqrt{1-2\delta M}}=.05$, i.e., the solution is actually in the square $[.15,.25]\times[-.15\times-.05]$.

```
f = @(x) [ 2*x(1)-x(2)-x(1)*x(2)-.5 ; 2*x(2)+x(1)*(1-x(2)^2) ];
fp = @(x)[ 2-x(2) , -1-x(1) ; 1-x(2)^2 , 2-2*x(1)*x(2) ];
B = [1 1;1 2];                          % from bounds beta_{ijk} for 2nd derivatives
xmin = [-.5;-.5]; xmax = [.5;.5];       % rectangle where bounds hold: xmin(j) <= x(j) <= xmax(j)

x = [0;0];                              % initial guess
for i=1:4
  b = f(x); A = fp(x); Ai = inv(A);
  d = -Ai*b; delta = norm(d,Inf);
  xnew = x + d;
  fprintf('x = [%g,%g]\n',xnew);
  M = norm(abs(Ai)*B,Inf);
  if delta <= 1/(2*M)                   % condition for delta in Newton-Kantorovich Theorem
    H = delta*M;
    r = 2*delta/(1+sqrt(1-2*H));
    if all(x-r>=xmin) && all(x+r<=xmax) % check condition for r in Theorem
      e = M*delta^2/(1-H+sqrt(1-2*H));  % a-posteriori error bound
      fprintf('     inf-norm error <= %g\n',e)
    end
  end
  x = xnew;
end
```

This prints

```
x = [0.2,-0.1]
     inf-norm error <= 0.05
x = [0.192982,-0.095614]
     inf-norm error <= 3.76594e-05
x = [0.192973,-0.0956046]
```

```
     inf-norm error <= 6.68672e-11
x = [0.192973,-0.0956046]
     inf-norm error <= 6.3505e-22
```

The actual errors $\left\|x^{(k)}-x^*\right\|_\infty$ for $k=1,\ldots,4$ are $7\cdot 10^{-3}$, $9.4\cdot 10^{-6}$, $2.8\cdot 10^{-11}$, $3\cdot 10^{-17}$ (the last value is affected by roundoff error)

Summary:

- find bounds $\left|\dfrac{\partial^2 f_i}{\partial x_j \partial x_k}(x)\right| \le \beta_{ijk}$ in a region D where you suspect a solution to be
- pick an initial guess $x^{(0)}$ in D
- perform one Newton step: $d := -f'(x^{(0)})^{-1} f(x^{(0)})$, $x^{(1)} := x^{(0)} + d$
- with M, r from (18): $\boxed{\text{if } \|d\|_\infty \le \frac{1}{2}M^{-1} \text{ and the points with } \left\|x-x^{(0)}\right\|_\infty \le r \text{ are in } D}$
  there is a **unique solution** x^* **in the region** $\left\|x-x^{(0)}\right\|_\infty \le r$
  we have an **a-posteriori bound** $\left\|x^*-x^{(1)}\right\|_\infty \le \varepsilon$
- if the conditions in the box do not hold: try it again for $x^{(1)}, x^{(2)}, \ldots$ in place of $x^{(0)}$.
  if $x^{(k)}$ does not seem to converge: you need a better initial guess $x^{(0)}$