Numerical methods

Vasile Gradinaru, Ralf Hiptmair and Arnulf Jentzen

February 25, 2013

Contents

Preface

These course materials are a modified version of the lecture slides written by Ralf Hiptmair and Vasile Gradinaru. They have been prepared for the course "401-0654-00L Numerische Methoden" in the spring semester 2013. These course materials are also inspired by the handwritten notes of Lars Kielhorn for the course "401-0654-00L Numerische Methoden" in the spring semester 2012. Special thanks are due to Markus Sprecher for proofreading the course materials.

Zürich, February 2013
Arnulf Jentzen

Reading instructions

These course materials are neither a textbook nor lecture notes. They are ment to be supplemented by explanations given in class.

Some pieces of advice:

- these course materials are not designed to be self-contained but to supplement explanations in class,
- this document is not meant for mere reading but for working with,
- study the relevant section of the course when doing homework problems.

Part I

Systems of Equations

Chapter 1

Iterative Methods for Non-Linear Systems of Equations

Goal 1.0.1 (Solving non-linear systems of equations)**.** *Given:* *natural number* $n \in \mathbb{N}$*, non-empty set* $D \subset \mathbb{R}^n$ *and function*

$$F\colon D \subset \mathbb{R}^n \to \mathbb{R}^n. \tag{1.1}$$

Aim: *Find* $x \in D$ *(approximatively) such that*

$$F(x) = 0. \tag{1.2}$$

Comments:

- Possible meaning: Availability of a procedure `function y=F(x)` evaluating F
- In general no existence & uniqueness
- Note: $F\colon D \subset \mathbb{R}^n \to \mathbb{R}^n$, i.e., "same number of equations and unknowns"
- Example: Consider $a \in (0, \infty)$, $n = 1$, $D = (0, \infty)$ and $F\colon (0, \infty) \to \mathbb{R}$ satisfying $F(x) = x^2 - a$ for all $x \in (0, \infty)$
- A nonlinear system of equations is a concept almost *too abstract to be useful*, because it covers an extremely wide variety of problems. Nevertheless, in this chapter we will mainly look at "generic" methods for such systems. This means that every method discussed may take a good deal of fine-tuning before it will really perform satisfactorily for a given non-linear system of equations.

1.1 Iterative methods

Remark 1.1.1 (Necessity of iterative methods)**.** *Gaussian elimination provides an algorithm that, if carried out in exact arithmetic, computes the solution of a linear system of equations with a finite number of elementary operations. However, linear systems of equations represent an exceptional case, because it is hardly ever possible to solve general systems of nonlinear equations using only finitely many elementary operations. Certainly, this is the case whenever irrational numbers are involved.*

Definition 1.1.1 (Iterative (m-point) methods)**.** *Let n, $m \in \mathbb{N}$ and $U \subset (\mathbb{R}^n)^m = \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ be a set. Then a function $\Phi\colon U \to \mathbb{R}^n$ is called* **(m-point) iteration function***. If $x^{(0)}, \ldots, x^{(m-1)} \in \mathbb{R}^n$ (***initial guess(es)***[1]), then we call $x^{(k)} \in \mathbb{R}^n \cup \{\infty\}$, $k \in \mathbb{N}_0$, given by*

$$x^{(k)} = \begin{cases} \Phi\big(x^{(k-m)}, \ldots x^{(k-1)}\big) & : (x^{(k-m)}, \ldots x^{(k-1)}) \in U \\ \infty & : \text{else} \end{cases} \tag{1.3}$$

for all $k \in \{m, m+1, \ldots\}$ the (sequence of) **iterates** *associated to Φ and $(x^{(0)}, \ldots, x^{(m-1)})$. The algorithm described by (1.3) is called* **iterative (m-point) method (with iteration function Φ)** *(sometimes also* **(m-point) iteration** *(with iteration function Φ) and in the case $m = 1$ sometimes also* **fixed point iteration** *(with iteration function Φ)).*

Example 1.1.1. *Consider $a \in (0, \infty)$, $n = m = 1$, $U = (0, \infty)$ and $\Phi\colon (0, \infty) \to \mathbb{R}$ given by*

$$\Phi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right) \tag{1.4}$$

for all $x \in (0, \infty)$. The iterates associated to Φ and $x^{(0)} \in (0, \infty)$

[1] german: Anfangsnäherung(en)

satisfy

$$x^{(k+1)} = \Phi(x^{(k)}) = \frac{1}{2}\left(x^{(k)} + \frac{a}{x^{(k)}}\right) \tag{1.5}$$

for all $k \in \mathbb{N}_0$.

Definition 1.1.2 (Convergence with given initial guess(es))**.** *An iterative method with iteration function* $\Phi\colon U \subset (\mathbb{R}^n)^m \to \mathbb{R}^n$ *and initial guess(es)* $(x^{(0)}, \dots, x^{(m-1)}) \in U$ **converges** *to* $x^* \in \mathbb{R}^n$ *if*

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

where $(x^{(k)})_{k\in\mathbb{N}_0}$ *are the iterates associated to* Φ *and* $(x^{(0)}, \dots, x^{(m-1)})$.

If an iterative method with a **continuous** iteration function $\Phi\colon U \subset (\mathbb{R}^n)^m \to \mathbb{R}^n$ and initial guess(es) $(x^{(0)}, \dots, x^{(m-1)}) \in U$ converges to $x^* \in \mathbb{R}^n$ and if $(x^*, \dots, x^*) \in U$, then

$$\begin{aligned} x^* &= \lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} \Phi\big(x^{(k-m)}, \dots, x^{(k-1)}\big) \\ &= \Phi(x^*, \dots, x^*) \end{aligned} \tag{1.6}$$

where $(x^{(k)})_{k\in\mathbb{N}_0}$ are the iterates associated to Φ and $(x^{(0)}, \dots, x^{(m-1)})$. In particular, if in addition $m = 1$, then (1.6) reduces to

$$x^* = \Phi(x^*) \tag{1.7}$$

and x^* is a **fixed point** of Φ.

Definition 1.1.3 (Consistency)**.** *An iterative method with iterative function* $\Phi\colon U \subset (\mathbb{R}^n)^m \to \mathbb{R}^n$ *is* <u>*consistent*</u> *with* (1.2) *if it holds for every* $x \in \mathbb{R}^n$ *that*

$$\begin{pmatrix} (x, \dots, x) \in U \\ \wedge\ \Phi(x, \dots, x) = x \end{pmatrix} \iff \begin{pmatrix} x \in D \\ \wedge\ F(x) = 0 \end{pmatrix}.$$

Definition 1.1.4 (Global convergence)**.** *An iterative method with iteration function* $\Phi\colon U \subset (\mathbb{R}^n)^m \to \mathbb{R}^n$ **converges globally** *to* $x^* \in \mathbb{R}^n$ *if it holds for every* $(x^{(0)}, \dots, x^{(m-1)}) \in U$ *that*

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

where $(x^{(k)})_{k\in\mathbb{N}_0}$ *are the iterates associated to* Φ *and* $(x^{(0)}, \dots, x^{(m-1)})$.

Definition 1.1.5 (Local convergence)**.** *An iterative method with iteration function* $\Phi\colon U \subset (\mathbb{R}^n)^m \to \mathbb{R}^n$ **converges locally** *to* $x^* \in \mathbb{R}^n$ *if there exists a neighborhood* $\mathcal{O} \subset \mathbb{R}^n$ *of* x^* *such that* $\mathcal{O}^m \cap U \neq 0$ *and such that for every* $(x^{(0)}, \ldots, x^{(m-1)}) \in \mathcal{O}^m \cap U$ *it holds that*

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

where $(x^{(k)})_{k\in\mathbb{N}_0}$ *are the iterates associated to* Φ *and* $(x^{(0)}, \ldots, x^{(m-1)})$.

Goal: Find iterative methods that converge (locally) to a solution of (1.2). Questions:

- How to measure the speed of convergence?
- When to terminate the iteration?

1.1.1 Vector norms

Norms provide tools for measuring errors. Recall the next definition. In the following $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$.

Definition 1.1.6 (Norm)**.** *Let* V *be a* $\mathbb{K}$*-vector space. A map* $\|\cdot\| : V \to [0, \infty)$ *is a* **norm** *on* V *if*

(i) $\forall\, v \in V\colon \quad v = 0 \Leftrightarrow \|v\| = 0 \qquad$ *(positive definite)*

(ii) $\forall\, v \in V, \lambda \in \mathbb{K}\colon \quad \|\lambda v\| = |\lambda|\,\|v\| \qquad$ *(homogeneous)*

(iii) $\forall\, v, w \in V\colon \quad \|v + w\| = \|v\| + \|w\| \qquad$ *(triangle inequality).*

Let $n \in \mathbb{N}$ and consider examples of norms on $\mathbb{K}^n$:

name	definition	MATLAB function
Euclidean norm	$\|x\|_2 := \sqrt{\lvert x_1\rvert^2 + \ldots + \lvert x_n\rvert^2}$	*norm(x)*
1-norm	$\|x\|_1 := \lvert x_1\rvert + \ldots + \lvert x_n\rvert$	*norm(x,1)*
∞-norm, max norm	$\|x\|_\infty := \max\{\lvert x_1\rvert, \ldots, \lvert x_n\rvert\}$	*norm(x,inf)*

for all $x = (x_1, \ldots, x_n) \in \mathbb{K}^n$.

Recall: equivalence of norms on finite dimensional vector spaces $\mathbb{K}^n$.

Definition 1.1.7 (Equivalence of norms)**.** *Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and let V be a $\mathbb{K}$-vector space. Two norms $\|\cdot\|_A$ and $\|\cdot\|_B$ on V are* **equivalent** *if*

$$\exists\, \underline{C}, \overline{C} \in (0,\infty)\colon \quad \forall\, v \in V\colon \quad \underline{C}\,\|v\|_A \le \|v\|_B \le \overline{C}\,\|v\|_A. \tag{1.8}$$

Theorem 1.1.1. *If V is a finite dimensional $\mathbb{K}$-vector space, i.e., $\dim(V) < \infty$, then all norms on V are equivalent.*

Some explicit norm equivalences:

- $\|x\|_2 \le \|x\|_1 \le \sqrt{n}\,\|x\|_2$,
- $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\,\|x\|_\infty$,
- $\|x\|_\infty \le \|x\|_1 \le n\,\|x\|_\infty$

for all $x \in \mathbb{K}^n$ and all $n \in \mathbb{N}$.

1.1.2 Speed of convergence

In the following $n \in \mathbb{N}$ is a natural number and $\|\cdot\| : \mathbb{R}^n \to [0,\infty)$ is a norm.

Definition 1.1.8 (Linear convergence)**.** *A sequence $x^{(k)} \in \mathbb{R}^n$, $k \in \mathbb{N}_0$,* **converges linearly** *to $x^* \in \mathbb{R}^n$ if*

$$\exists\, L \in [0,1)\colon \forall\, k \in \mathbb{N}_0\colon \left\|x^{(k+1)} - x^*\right\| \le L \left\|x^{(k)} - x^*\right\|. \tag{1.9}$$

In that case, the smallest admissible value for $L \in [0,1)$ in (1.9) is referred as **rate of convergence**.

Comment on the impact of choice of norm:

- Fact of convergence of a sequence is **independent** of choice of norm.
- Fact of linear convergence **depends** of choice of norm.

- Rate of linear convergence **depends** of choice of norm.

Remark 1.1.2 (Seeing linear convergence)**.** *If* $(x^{(k)})_{k\in\mathbb{N}_0}$ *converges linearly to* $x^* \in \mathbb{R}^n$ *with rate* $L \in [0,1)$*, then*

$$\begin{aligned}&\left\|x^{(k)} - x^*\right\| \le L \left\|x^{(k-1)} - x^*\right\| \le L^2 \left\|x^{(k-2)} - x^*\right\| \\ &\le \cdots \le L^k \left\|x^{(0)} - x^*\right\| \\ &\Rightarrow \log\Big(\underbrace{\left\|x^{(k)} - x^*\right\|}_{=:\varepsilon_k}\Big) \le k\log(L) + \log\big(\left\|x^{(0)} - x^*\right\|\big)\end{aligned}$$

for all $k \in \mathbb{N}_0$*. Moreover, if in addition* $\forall\, k \in \mathbb{N}_0 \colon \varepsilon_{k+1} \approx L\varepsilon_k$ *and* $L \in (0,1)$*, then*

$$\forall\, k \in \mathbb{N}_0 \colon \quad \log\big(\varepsilon_k\big) \approx k\log(L) + \log\big(\varepsilon_0\big). \tag{1.10}$$

In that case $\log(L)$ *describes the slope of the error graph in lin-log chart.*

Example: Consider $\Phi \colon (0,2\pi)\backslash\{\pi\} \to \mathbb{R}$ satisfying

$$\Phi(x) = x + \frac{\cos(x)+1}{\sin(x)} \tag{1.11}$$

for all $x \in (0,2\pi)\backslash\{\pi\}$ and $x^{(0)} \in \{0.4, 0.6, 1\}$. The iterates $x^{(k)} \in (0,2\pi)\backslash\{\pi\}$, $k \in \mathbb{N}_0$, associated to Φ satisfy

$$x^{(k+1)} = x^{(k)} + \frac{\cos(x^{(k)})+1}{\sin(x^{(k)})} \tag{1.12}$$

for all $k \in \mathbb{N}_0$.

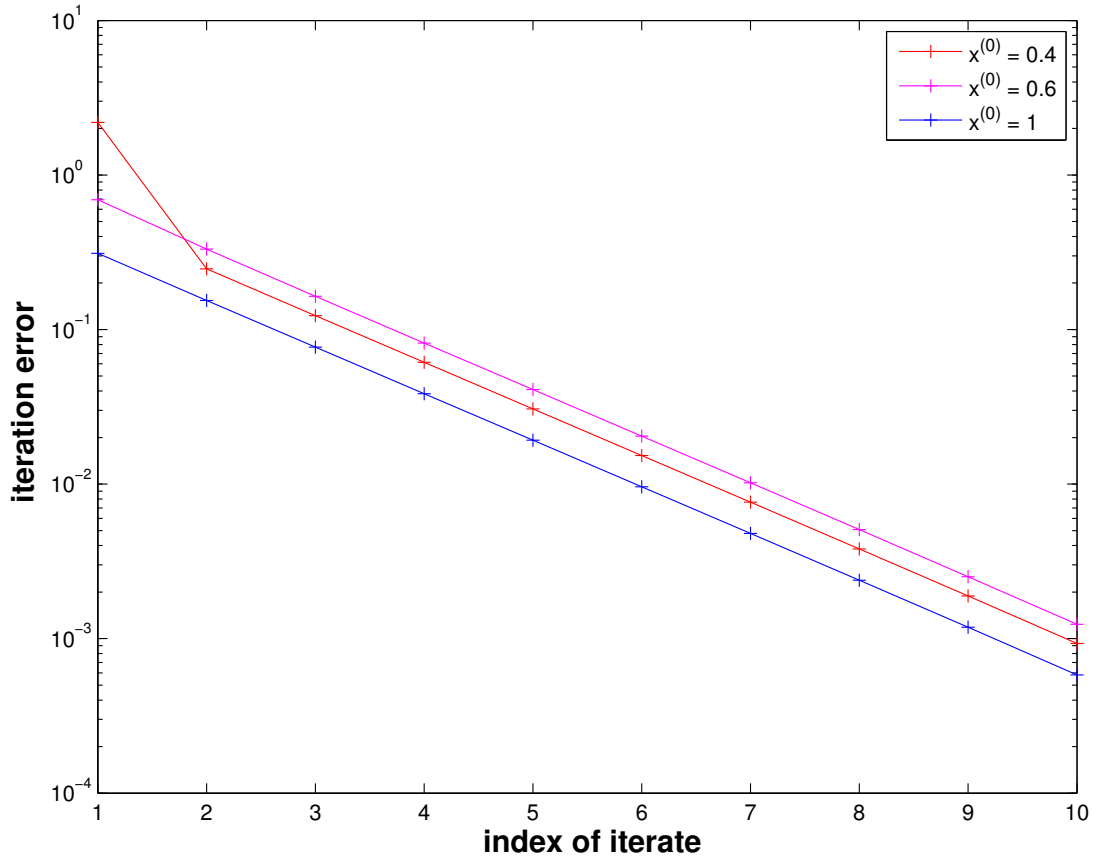
Listing 1.1: —MATLAB code for approximation of errors and rate of convergence for example (1.12)

```
y = [ ];
for i = 1:15
   x = x + (cos(x)+1)/sin(x);
   y = [y,x];
end
err = y - x;
rate = err(2:15)./err(1:14);
```

Note: $x^{(15)}$ replaces the limit $\lim_{k\to\infty} x^{(k)}$ in the computation of the rate of convergence.

k	$x^{(0)} = 0.4$		$x^{(0)} = 0.6$		$x^{(0)} = 1$	
	$x^{(k)}$	$\frac{\lvert x^{(k)}-x^{(15)}\rvert}{\lvert x^{(k-1)}-x^{(15)}\rvert}$	$x^{(k)}$	$\frac{\lvert x^{(k)}-x^{(15)}\rvert}{\lvert x^{(k-1)}-x^{(15)}\rvert}$	$x^{(k)}$	$\frac{\lvert x^{(k)}-x^{(15)}\rvert}{\lvert x^{(k-1)}-x^{(15)}\rvert}$
2	3.3887	0.1128	3.4727	0.4791	2.9873	0.4959
3	3.2645	0.4974	3.3056	0.4953	3.0646	0.4989
4	3.2030	0.4992	3.2234	0.4988	3.1031	0.4996
5	3.1723	0.4996	3.1825	0.4995	3.1224	0.4997
6	3.1569	0.4995	3.1620	0.4994	3.1320	0.4995
7	3.1493	0.4990	3.1518	0.4990	3.1368	0.4990
8	3.1454	0.4980	3.1467	0.4980	3.1392	0.4980

Numerical computations indicate: **rate of convergence** ≈ 0.5.



Definition 1.1.9 (Order of convergence)**.** *A sequence* $x^{(k)} \in \mathbb{R}^n$, $k \in \mathbb{N}_0$, **converges with order** $p \in [1,\infty)$ *to* x^* *if* $x^* = \lim_{k\to\infty} x^{(k)}$ *and*

$$\exists\, C \in [0,\infty)\colon \forall\, k \in \mathbb{N}_0\colon \left\|x^{(k+1)} - x^*\right\| \leq C \left\|x^{(k)} - x^*\right\|^p.$$

Comments:

- Convergence with order $p \in [1, \infty)$ is independent of choice of norm.
- Convergence with order 2 is sometimes referred as **quadratic convergence** and convergence with order 3 is sometimes referred as **cubic convergence**

Remark 1.1.3 (Seeing higher order convergence)**.** *Assume that* $x^{(k)} \in \mathbb{R}^n$, $k \in \mathbb{N}_0$, *converges with order* $p \in (1, \infty)$ *to* $x^* \in \mathbb{R}^n$. *Then*

$$\begin{aligned}
&\underbrace{\left\|x^{(k)} - x^*\right\|}_{=:\varepsilon_k} \leq C\left\|x^{(k-1)} - x^*\right\|^p \leq C^{(p^0+p^1)}\left\|x^{(k-2)} - x^*\right\|^{(p^2)} \\
&\leq \cdots \leq C^{(p^0+p^1+\ldots+p^{(k-1)})}\left\|x^{(0)} - x^*\right\|^{(p^k)} \\
&= C^{\frac{(p^k-1)}{(p-1)}}\left\|x^{(0)} - x^*\right\|^{(p^k)} \\
&\Rightarrow \log\left(\varepsilon_k\right) \leq -\frac{\log(C)}{(p-1)} + \left(\frac{\log(C)}{(p-1)} + \log\left(\varepsilon_0\right)\right) p^k
\end{aligned}$$

for all $k \in \mathbb{N}_0$. *Moreover, if* $\forall\, k \in \mathbb{N}_0 \colon \varepsilon_{k+1} \approx C\,(\varepsilon_k)^p$ *and* $C > 0$, *then*

$$\forall\, k \in \mathbb{N}_0 \colon \log\left(\varepsilon_k\right) \approx -\frac{\log(C)}{(p-1)} + \left(\frac{\log(C)}{(p-1)} + \log\left(\varepsilon_0\right)\right) p^k. \quad (1.13)$$

In that case, the error graph is a concave power curve in a lin-log chart provided that ε_0 *is sufficiently small. Furthermore, if* $C > 0$ *then for every* $k \in \mathbb{N}$:

$$\begin{pmatrix} \varepsilon_{k+1} \approx C\,(\varepsilon_k)^p \\ \varepsilon_k \approx C\,(\varepsilon_{k-1})^p \\ 0 < \varepsilon_{k+1} < \varepsilon_k < \varepsilon_{k-1} \end{pmatrix} \quad \Rightarrow \quad \frac{\log\left(\varepsilon_{k+1}\right) - \log\left(\varepsilon_k\right)}{\log\left(\varepsilon_k\right) - \log\left(\varepsilon_{k-1}\right)} \approx p. \quad (1.14)$$

Example 1.1.2. *Consider* $a \in (0, \infty)$, $D = (0, \infty)$, $F(x) = x^2 - a$ *for all* $x \in (0, \infty)$, $x^{(0)} \in D$ *and* $\Phi \colon (0, \infty) \to \mathbb{R}$ *given by*

$$\Phi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right) \quad (1.15)$$

for all $x \in (0,\infty)$ *(cf. Example 1.1.1). The iterates associated to* Φ *and* $x^{(0)}$ *satisfies*

$$x^{(k+1)} = \frac{1}{2}\left(x^{(k)} + \frac{a}{x^{(k)}}\right) \tag{1.16}$$

for all $k \in \mathbb{N}_0$. *Hence*

$$\begin{aligned}\left|x^{(k+1)} - \sqrt{a}\right| &= \frac{1}{2x^{(k)}}\left|\left(\left(x^{(k)}\right)^2 + a\right) - 2x^{(k)}\sqrt{a}\right| \\ &= \frac{1}{2x^{(k)}}\left|x^{(k)} - \sqrt{a}\right|^2\end{aligned} \tag{1.17}$$

for all $k \in \mathbb{N}_0$. *Arithmetic-geometric mean inequality (AGM), i.e.,* $\forall a,b \in [0,\infty)\colon \sqrt{ab} \le \frac{1}{2}(a+b)$, *implies* $\forall k \in \mathbb{N}\colon x^{(k)} \ge \sqrt{a}$ *and hence*

$$\forall\, k \in \mathbb{N}\colon x^{(k)} \ge x^{(k+1)} \ge \sqrt{a} \qquad \textit{and} \tag{1.18}$$

$$\forall\, k \in \mathbb{N}_0\colon \left|x^{(k+1)} - \sqrt{a}\right| \le \frac{1}{2\min\left(\sqrt{a}, x^{(0)}\right)}\left|x^{(k)} - \sqrt{a}\right|^2 \tag{1.19}$$

This shows that $\left(x^{(k)}\right)_{k\in\mathbb{N}_0}$ *converges with order* 2 *(***quadratically***) to* $\sqrt{a}$. **Numerical experiment:** *Iterates for* $a = 2$*:*

k	$x^{(k)}$	$e^{(k)} := x^{(k)} - \sqrt{2}$	$\frac{\log(\lvert e^{(k)}\rvert/\lvert e^{(k-1)}\rvert)}{\log(\lvert e^{(k-1)}\rvert/\lvert e^{(k-2)}\rvert)}$
0	2.0000000000000000	0.5857864376269485	
1	1.5000000000000000	0.0857864376269485	
2	1.4166666666666652	0.0024531042935713 7	1.850
3	1.4142156862745096 6	0.0000021239014145 2	1.984
4	1.4142135623746898 7	0.0000000000015947 2	2.000
5	1.4142135623730949 2	0.0000000000000002 2	0.630[2]

Note the doubling of the numbers of significant digits in each step.

Remark 1.1.4 (Number of significant digits)**.** *Assume that* $x^{(k)} \in \mathbb{R}$, $k \in \mathbb{N}_0$, *converges with order* $p \in (1,\infty)$ *to* $x^* \in \mathbb{R}\setminus\{0\}$. *Define*

$$\delta_k := \frac{x^{(k)} - x^*}{x^*} \qquad \textit{(number of significant digits)} \tag{1.20}$$

[2] impact of roundoff errors

for all $k \in \mathbb{N}_0$. *Then*

$$x^{(k)} - x^* = \delta_k x^* \qquad \text{and} \qquad x^{(k)} = x^*\,(1 + \delta_k) \tag{1.21}$$

for all $k \in \mathbb{N}_0$. *Note that* $\delta_k \approx 10^{-l}$ *means that* $x^{(k)}$ *has* $l \in \mathbb{N}$ *significant digits. Next observe that*

$$\left|x^* \delta_{k+1}\right| = \left|x^{(k+1)} - x^*\right| \leq C \left|x^{(k)} - x^*\right|^p = C \left|x^* \delta_k\right|^p \tag{1.22}$$

for all $k \in \mathbb{N}_0$ *and hence*

$$|\delta_{k+1}| \leq \left(C\,|x^*|^{(p-1)}\right) |\delta_k|^p \tag{1.23}$$

for all $k \in \mathbb{N}_0$. *Hence, if* $C\,|x^*|^{(p-1)} \approx 1$ *and* $\delta_k \approx 10^{-l}$, *then* $\delta_{k+1} \approx 10^{-lp}$.

1.1.3 Termination criteria

Let $(x^{(k)})_{k \in \mathbb{N}_0}$ be a sequence of iterates associated to an iteration function $\Phi \colon U \subset (\mathbb{R}^n)^m \to \mathbb{R}^n$ and initial guess(es) $(x^{(0)}, \ldots, x^{(m-1)}) \in U$ which converges to a solution $x^* \in D$ of (1.2). Typically there exist no $K \in \mathbb{N}$ such that

$$\forall\, k \in \{K, K+1, \ldots\} \colon \qquad x^{(k)} = x^*. \tag{1.24}$$

Thus we can typically only hope to compute an **approximative** solution of (1.2) by accepting $x^{(K)}$ as result for some suitable $K \in \mathbb{N}$. **Termination criteria** (german: **Abbruchbedingungen**) are used to determine a suitable value for K.

For the sake of efficiency: stop iteration at $K \in \mathbb{N}$ when norm of iteration error

$$\left\|x^{(K)} - x^*\right\| \tag{1.25}$$

is just "**small enough**".

"**Small enough**" depends on concrete setting: Let $\tau \in (0, \infty)$ be given (**prescribed tolerance**). Then the usual goal is to calculate $x^{(K)}$ such that

$$\left\|x^{(K)} - x^*\right\| \leq \tau. \tag{1.26}$$

The **optimal termination index** is

$$K = \min\Big\{k \in \mathbb{N}_0 \colon \big\|x^{(k)} - x^*\big\| \leq \tau\Big\}. \tag{1.27}$$

Termination criteria:

- **A priori termination criteria:**

  1.) stop iteration after fixed number of steps; possibly depending on $x^{(0)}$ but independent of $(x^{(k)})_{k\in\{1,2,\ldots,K\}}$

  Drawback: hardly ever possible to ensure that (1.26) is fulfilled!

- **A posteriori termination criteria** (use already computed iterates $(x^{(k)})_{k\in\{0,1,\ldots,K\}}$ to decide when to stop):

  2.) Reliable termination: stop iteration when (1.26) is fulfilled.

  Drawback: x^* is not know.

  However: **invoking additional properties** of the **nonlinear equation** (1.2) or the **iteration** is sometimes possible to tell for sure that

  $$\big\|x^{(K)} - x^*\big\| \leq \tau \tag{1.28}$$

  is fulfilled for some $K \in \mathbb{N}$. Though this K may be (significantly) larger than the optimal termination index in (1.27).

  3.) Let $\mathbb{M}$ be the set of **machine numbers**. Termination criteria based on finiteness of $\mathbb{M}$:

  Wait until iteration becomes stationary in $\mathbb{M}$.

  Drawback: Possibly grossly inefficient! (Always computes "**up to machine precision**")

Listing 1.2: stationary iteration in $\mathbb{M}$ (cf. Example 1.1.2)

```
function x = sqrtit(a)
x_old = -1; x = a;
while (x_old ~= x)
  x_old = x;
  x = 0.5*(x+a/x);
end
```

4.) Residual based termination: stop iteration when

$$\left\|F(x^{(K)})\right\| \leq \tau \tag{1.29}$$

is fulfilled for some $K \in \mathbb{N}$.

Drawback: no guaranteed accuracy

Remark 1.1.5 (An a posteriori termination criterion for linearly convergent iterations)**.** *Assume that $(x^{(k)})_{k\in\mathbb{N}_0}$ converges linearly to x^* with rate $L \in [0,1)$ and let $\tilde{L} \in [L,1)$. Then*

$$\begin{aligned}\left\|x^{(k-1)}-x^*\right\| &\leq \left\|x^{(k)}-x^{(k-1)}\right\| + \left\|x^{(k)}-x^*\right\| \\ &\leq \left\|x^{(k)}-x^{(k-1)}\right\| + \tilde{L}\left\|x^{(k-1)}-x^*\right\|\end{aligned} \tag{1.30}$$

and hence

$$\left\|x^{(k-1)}-x^*\right\| \leq \frac{1}{(1-\tilde{L})}\left\|x^{(k)}-x^{(k-1)}\right\| \tag{1.31}$$

and therefore

$$\left\|x^{(k)}-x^*\right\| \leq \frac{\tilde{L}}{(1-\tilde{L})}\left\|x^{(k)}-x^{(k-1)}\right\| \tag{1.32}$$

for all $k \in \mathbb{N}$. This suggests that we take the right hand side of (1.32) *as an* **a posteriori error bound** *for a* **reliable termination criterion**.

1.2 1-point iteration methods

General construction of fixed point iterations that are **consistent** with (1.2): rewrite

$$F(x) = 0 \tag{1.33}$$

for $x \in D$ to

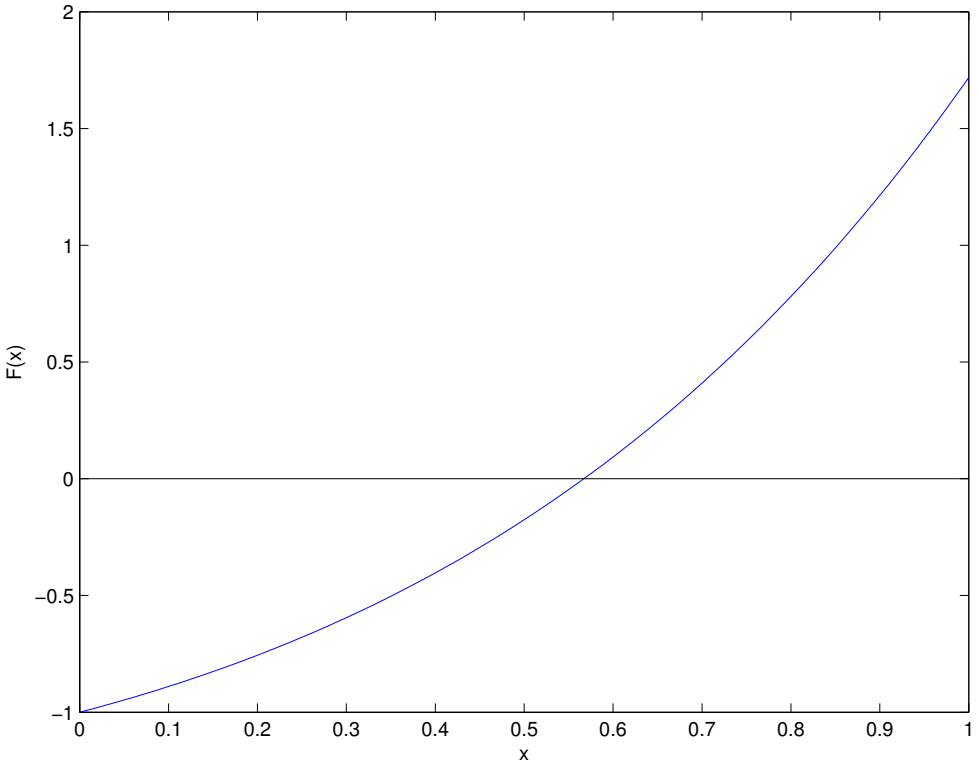
$$\Phi(x) = x \tag{1.34}$$

for $x \in D$ where $\Phi\colon D \to \mathbb{R}^n$ is an appropriate function. Note: there are *many ways* to transform (1.33) to (1.34).

Example: Consider $n = 1$, $D = [0, 1]$ and $F\colon D \to \mathbb{R}$ given by

$$F(x) = x\, e^x - 1 \tag{1.35}$$

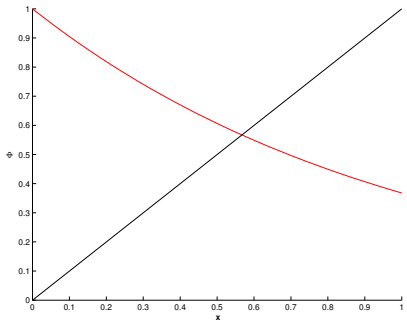
for all $x \in [0, 1]$.



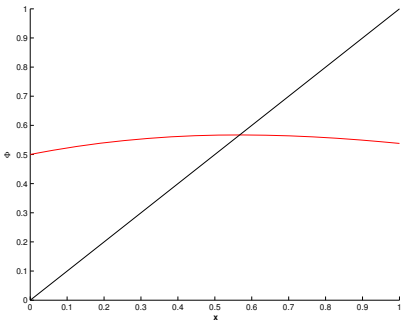
Moreover, consider $\Phi_1, \Phi_2, \Phi_3\colon [0, 1] \to \mathbb{R}$ given by

$$\Phi_1(x) = e^{-x}, \quad \Phi_2(x) = \frac{1 + x}{1 + e^x}, \quad \Phi_3(x) = x + 1 - x\, e^x \tag{1.36}$$

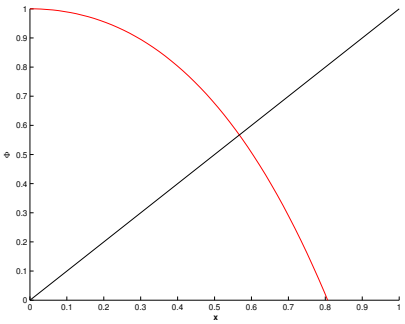
for all $x \in [0, 1]$.



function Φ_1



function Φ_2



function Φ_3

k	$x_1^{(k+1)} := \Phi_1(x_1^{(k)})$	$x_2^{(k+1)} := \Phi_2(x_2^{(k)})$	$x_3^{(k+1)} := \Phi_3(x_3^{(k)})$
0	0.500000000000000	0.500000000000000	0.500000000000000
1	0.606530659712633	0.566311003197218	0.675639364649936
2	0.545239211892605	0.567143165034862	0.347812678511202
3	0.579703094878068	0.567143290409781	0.855321409174107
4	0.560064627938902	0.567143290409784	-0.156505955383169
5	0.571172148977215	0.567143290409784	0.977326422747719
6	0.564862946980323	0.567143290409784	-0.619764251895580
7	0.568438047570066	0.567143290409784	0.713713087416146
8	0.566409452746921	0.567143290409784	0.256626649129847
9	0.567559634262242	0.567143290409784	0.924920676910549
10	0.566907212935471	0.567143290409784	-0.407422405542253

k	$\lvert x_1^{(k+1)} - x^*\rvert$	$\lvert x_2^{(k+1)} - x^*\rvert$	$\lvert x_3^{(k+1)} - x^*\rvert$
0	0.067143290409784	0.067143290409784	0.067143290409784
1	0.039387369302849	0.000832287212566	0.108496074240152
2	0.021904078517179	0.000000125374922	0.219330611898582
3	0.012559804468284	0.000000000000003	0.288178118764323
4	0.007078662470882	0.000000000000000	0.723649245792953
5	0.004028858567431	0.000000000000000	0.410183132337935
6	0.002280343429460	0.000000000000000	1.186907542305364
7	0.001294757160282	0.000000000000000	0.146569797006362
8	0.000733837662863	0.000000000000000	0.310516641279937
9	0.000416343852458	0.000000000000000	0.357777386500765
10	0.000236077474313	0.000000000000000	0.974565695952037

Numerical simluations indicate:

- linear convergence of $(x_1^{(k)})_{k\in\mathbb{N}_0}$,
- quadratic convergence of $(x_2^{(k)})_{k\in\mathbb{N}_0}$,
- no convergence (erratic behavior) of $(x_3^{(k)})_{k\in\mathbb{N}_0}$.

Question: Can we explain/forecast the behavior of a 1-point iteration?

1.2.1 Convergence analysis

As above, $n \in \mathbb{N}$ is a natural number and $\|\cdot\| : \mathbb{R}^n \to [0,\infty)$ is a norm in the following .

Definition 1.2.1 (**Contractive mapping**). *A function* $\Phi\colon U \subset \mathbb{R}^n \to \mathbb{R}^n$ *is* **contractive** *(a* **contraction**[3]*) (w.r.t. norm* $\|\cdot\|$*) if*

$$\exists\, L \in [0,1)\colon \quad \forall\, x,y \in U\colon \quad \|\Phi(x)-\Phi(y)\| \le L\,\|x-y\|\,. \tag{1.37}$$

The real number $L \in [0,1)$ *is referred as Lipschitz constant.*

Remark 1.2.1 (**Properties of contractions**). *Let* $\Phi\colon U \subset \mathbb{R}^n \to \mathbb{R}^n$ *be a contraction with Lipschitz constant* $L \in [0,1)$ *and let* $x^* \in U$ *be a fixed point of* Φ *(i.e.,* $\Phi(x^*) = x^*$*). If* $y^* \in U$ *is a further fixed points of* Φ*, then*

$$\|x^*-y^*\| = \|\Phi(x^*)-\Phi(y^*)\| \le L\,\|x^*-y^*\| \tag{1.38}$$

and therefore

$$(1-L)\,\|x^*-y^*\| \le 0 \tag{1.39}$$

and hence

$$x^* = y^*. \qquad (\textbf{Uniqueness} \text{ of fixed points for contractive maps})$$

Moreover, if $x^{(0)} \in U$ *and if* $x^{(k)} \in U$*,* $k \in \mathbb{N}_0$*, are iterates associated to* Φ *and* $x^{(0)}$*, then*

$$\left\|x^{(k+1)}-x^*\right\| = \left\|\Phi(x^{(k)})-\Phi(x^*)\right\| \le L\left\|x^{(k)}-x^*\right\| \tag{1.40}$$

for all $k \in \mathbb{N}_0$ *and* $(x^{(k)})_{k\in\mathbb{N}_0}$ *thus converges linearly to* x^**.*

[3] german: kontraktiv bzw. Kontraktion

Theorem 1.2.1 (**Banach fixed point theorem**). *Let $U \subset \mathbb{K}^n$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$) be a non-empty closed set and let $\Phi\colon U \to \mathbb{K}^n$ be a contraction with $\Phi(U) \subset U$. Then there* **exists** *a unique fixed point $x^* \in U$ of Φ and the iterative method with iteration function Φ converges globally to x^*. Moreover, for every $x^{(0)}$ it holds that the iterates associated to Φ and $x^{(0)}$ converge linearly to x^*.*

Proof. Uniqueness is already proved in Remark 1.2.1. Existence proof based on 1-point iteration: Let $L \in [0,1)$ be a Lipschitz constant for Φ, let $x^{(0)} \in U$ and let $x^{(k)} \in U$, $k \in \mathbb{N}_0$, be iterates associated to Φ and $x^{(0)}$. Then

$$\begin{aligned}\left\|x^{(k+N)} - x^{(k)}\right\| &\leq \sum_{j=k}^{k+N-1} \left\|x^{(j+1)} - x^{(j)}\right\| \leq \sum_{j=k}^{k+N-1} L^j \left\|x^{(1)} - x^{(0)}\right\| \\ &\leq \left(\sum_{j=k}^{\infty} L^j\right) \left\|x^{(1)} - x^{(0)}\right\| = \frac{L^k}{(1-L)} \left\|x^{(1)} - x^{(0)}\right\|\end{aligned} \tag{1.41}$$

for all $k, N \in \mathbb{N}_0$. Hence, $(x^{(k)})_{k\in\mathbb{N}_0}$ is Cauchy sequence and therefore convergent to $x^* \in U$. This together with (1.6)–(1.7) completes the proof. □

Definition 1.2.2 (**Induced matrix norm**). *Let $n, m \in \mathbb{N}$, let $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and let $\|\cdot\|_A : \mathbb{K}^n \to [0,\infty)$ and $\|\cdot\|_B : \mathbb{K}^m \to [0,\infty)$ be norms. Then the norm*

$$\|\cdot\|_{L(\|\cdot\|_A,\|\cdot\|_B)} : \mathbb{K}^{n,m} \to [0,\infty) \tag{1.42}$$

given by

$$\|M\|_{L(\|\cdot\|_A,\|\cdot\|_B)} := \sup_{v\in\mathbb{K}^m\setminus\{0\}} \frac{\|Mv\|_A}{\|v\|_B} \tag{1.43}$$

for all $M \in \mathbb{K}^{n,m}$ is called the **matrix norm induced by $\|\cdot\|_A$ and $\|\cdot\|_B$**. *For simplicity we sometimes also write*

- $\|\cdot\|_2$ *instead of* $\|\cdot\|_{L(\|\cdot\|_2,\|\cdot\|_2)}$,

- $\|\cdot\|_\infty$ *instead of* $\|\cdot\|_{L(\|\cdot\|_\infty,\|\cdot\|_\infty)}$,
- $\|\cdot\|_1$ *instead of* $\|\cdot\|_{L(\|\cdot\|_1,\|\cdot\|_1)}$ *and*
- $\|\cdot\|_{L(\|\cdot\|_A)}$ *instead of* $\|\cdot\|_{L(\|\cdot\|_A,\|\cdot\|_A)}$.

Remark 1.2.2 (**A simple criterion for a continuously differentiable function to be contractive**). *Let $U \subset \mathbb{R}^n$ be a convex open set and let $\Phi\colon U \subset \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable. Then*

$$\begin{aligned}\big\|\Phi(x)-\Phi(y)\big\| &= \left\|\int_0^1 \Phi'\big(y+r(x-y)\big)(x-y)\,dr\right\| \\ &\le \left[\int_0^1 \big\|\Phi'\big(y+r(x-y)\big)\big\|_{L(\|\cdot\|)}\,dr\right]\big\|x-y\big\| \\ &\le \left[\sup_{r\in[0,1]} \big\|\Phi'\big(y+r(x-y)\big)\big\|_{L(\|\cdot\|)}\,dr\right]\big\|x-y\big\|\end{aligned} \tag{1.44}$$

for all $x, y \in U$. Hence, if

$$\sup_{x\in U}\|\Phi'(x)\|_{L(\|\cdot\|)} < 1, \tag{1.45}$$

then Φ is a **contraction** *with Lipschitz constant*

$$L := \sup_{x\in U}\|\Phi'(x)\|_{L(\|\cdot\|)} \tag{1.46}$$

and Remark 1.2.1 **applies***! This yields the following corollary.*

Corollary 1.2.1. *Let $U \subset \mathbb{R}^n$ be a convex open set, let $\Phi\colon U \subset \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable with $\Phi(U) \subset U$ and*

$$\tilde{L} := \sup_{x\in U}\|\Phi'(x)\|_{L(\|\cdot\|)} < 1, \tag{1.47}$$

let $x^* \in U$ be a fixed point of Φ and let $x^{(0)} \in U$. Then the iterates associated to Φ and $x^{(0)}$ **converge linearly** *to x^* with* **rate of convergence** *$L \le \tilde{L}$. In particular, the iterative method with iteration function Φ* **converges globally** *to x^*.*

Note: Corollary 1.2.1 provides

$$\tilde{L} := \sup_{x \in U} \|\Phi'(x)\|_{L(\|\cdot\|)} < 1 \tag{1.48}$$

for Remark 1.1.5.

Lemma 1.2.1. *Let $\Phi\colon U \subset \mathbb{R}^n \to \mathbb{R}^n$ be a function which is differentiable in an interior point $x^* \in U$ of U and assume that $\Phi(x^*) = x^*$ and*

$$\|\Phi'(x^*)\|_{L(\|\cdot\|)} < 1. \tag{1.49}$$

Then there exists a $\delta \in (0, \infty)$ such that for every

$$x^{(0)} \in \mathcal{O} := \{x \in \mathbb{R}^n \colon \|x - x^*\| \leq \delta\} \tag{1.50}$$

it holds that the iterates associated to Φ and $x^{(0)}$ lie in $\mathcal{O}$ and converge linearly to x^*. In particular, the iterative method with iteration function Φ **converges locally** *to x^*.*

Proof. Let $\delta \in (0, \infty)$ be a real number such that

$$\{y \in \mathbb{R}^n \colon \|x^* - y\| \leq \delta\} \subset U \tag{1.51}$$

and such that

$$\frac{\|\Phi(x) - \Phi(x^*) - \Phi'(x^*)(x - x^*)\|}{\|x - x^*\|} \leq \frac{1 - \|\Phi'(x^*)\|_{L(\|\cdot\|)}}{2} \tag{1.52}$$

for all $x \in \mathbb{R}^n \backslash \{x^*\}$ with $\|x - x^*\| \leq \delta$. The inverse triangle inequality then shows for every $x \in \mathbb{R}^n \backslash \{x^*\}$ with $\|x - x^*\| \leq \delta$ that

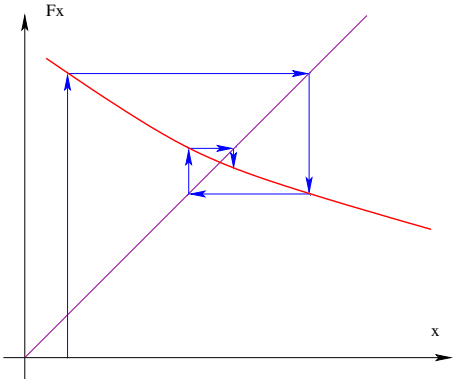
$$\frac{\|\Phi(x) - \Phi(x^*)\|}{\|x - x^*\|} - \frac{\|\Phi'(x^*)(x - x^*)\|}{\|x - x^*\|} \leq \frac{1 - \|\Phi'(x^*)\|_{L(\|\cdot\|)}}{2} \tag{1.53}$$

and therefore

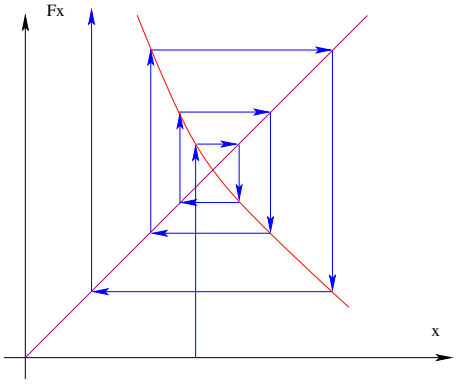
$$\begin{aligned} \|\Phi(x) - \Phi(x^*)\| &\leq \|\Phi'(x^*)(x - x^*)\| + \frac{\left(1 - \|\Phi'(x^*)\|_{L(\|\cdot\|)}\right) \|x - x^*\|}{2} \\ &\leq \underbrace{\left[\frac{1 + \|\Phi'(x^*)\|_{L(\|\cdot\|)}}{2}\right]}_{=:\tilde{L} \in [\frac{1}{2}, 1)} \|x - x^*\|. \end{aligned} \tag{1.54}$$

This completes the proof. □

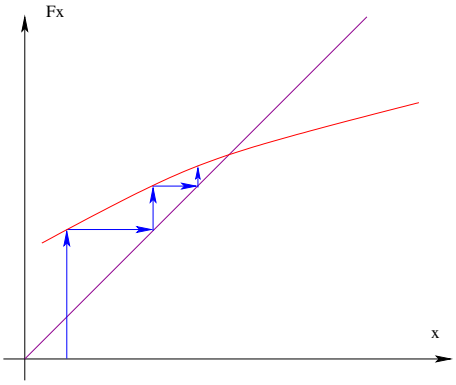
"Visualization" of the statement of Lemma 1.2.1: The iteration converges *locally* if Φ is flat in a neighborhood of x^*.



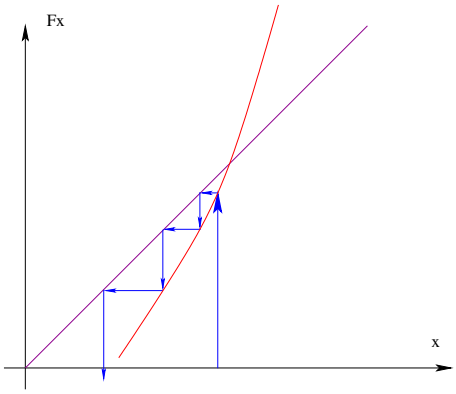
$-1 < \Phi'(x^*) \leq 0$; convergence



$\Phi'(x^*) < -1$; divergence



$0 \leq \Phi'(x^*) < 1$; convergence



$1 < \Phi'(x^*)$; divergence

Theorem 1.2.2 (Higher order local convergence of fixed point iterations). *Let $U \subset \mathbb{R}^n$ be an open set, let $k \in \mathbb{N}$, let $\Phi\colon U \to \mathbb{R}^n$ be $(k+1)$-times continuously differentiable and let $x^* \in U$ be a* **fixed point** *of Φ with*

$$\Phi^{(l)}(x^*) = 0 \tag{1.55}$$

for all $l \in \{1, 2, \ldots, k\}$. Then there exists a $\delta \in (0, \infty)$ such that for every

$$x^{(0)} \in \mathcal{O} := \{x \in \mathbb{R}^n\colon \|x - x^*\| \leq \delta\} \tag{1.56}$$

it holds that the iterates associated to Φ and $x^{(0)}$ lie in $\mathcal{O}$ and **converge with order** *$k+1$ to x^*.*

Proof. Lemma 1.2.1 proves that there exists a $\delta \in (0, \infty)$ such that such that for every

$$x^{(0)} \in \mathcal{O} := \{x \in \mathbb{R}^n \colon \|x - x^*\| \leq \delta\} \tag{1.57}$$

it holds that the iterates associated to Φ and $x^{(0)}$ lie in $\mathcal{O}$ and converge linearly to x^*. Let $x^{(0)} \in \mathcal{O}$ and let $(x^{(k)})_{k \in \mathbb{N}_0}$ be iterates associated to Φ and $x^{(0)}$. Next **Talyor's formula** proves

$$\begin{aligned}
&\Phi(x) \\
&= \Phi(x^*) + \sum_{l=1}^{k} \frac{1}{l!} \Phi^{(l)}(x^*)\,(x - x^*, \ldots, x - x^*) \\
&+ \int_0^1 \Phi^{(k+1)}(x^* + r(x - x^*))\,(x - x^*, \ldots, x - x^*)\,\frac{(1-r)^k}{k!}\,dr
\end{aligned} \tag{1.58}$$

and hence

$$\begin{aligned}
&\|\Phi(x) - \Phi(x^*)\| \\
&\leq \int_0^1 \left\| \Phi^{(k+1)}(x^* + r(x - x^*))\,(x - x^*, \ldots, x - x^*)\,\frac{(1-r)^k}{k!} \right\| dr \\
&\leq \left[\sup_{r \in [0,1]} \left\| \Phi^{(k+1)}(x^* + r(x - x^*)) \right\|_{L^{(k+1)}(\|\cdot\|)} \right] \|x - x^*\|^{(k+1)} \\
&\leq \underbrace{\left[\sup_{y \in \mathcal{O}} \left\| \Phi^{(l+1)}(y) \right\|_{L^{(k+1)}(\|\cdot\|)} \right]}_{=:C} \|x - x^*\|^{(k+1)}
\end{aligned} \tag{1.59}$$

for all $x \in \mathcal{O}$. Therefore

$$\left\| x^{(k+1)} - x^* \right\| = \left\| \Phi(x^{(k)}) - \Phi(x^*) \right\| \leq C \left\| x^{(k)} - x^* \right\|^{(k+1)} \tag{1.60}$$

for all $k \in \mathbb{N}_0$. □

1.2.2 Model function methods

Model function methods are a class of iterative m-point methods for finding zeroes of F where $m \in \mathbb{N}$.

Idea: Given: approximate zeroes $x^{(k-m+1)}, x^{(k-m+2)}, \ldots, x^{(k)} \in D$; Calculate approximate zero $x^{(k+1)}$ in two steps:

1. approximate F by **model function** $\tilde{F}$
2. $x^{(k+1)} :=$ zero of $\tilde{F}$ (has to be readily available, e.g., though an analytic formula)

Examples:

- **Newton's method** (iterative 1-point method; see Section 1.2.3 below); assume that F is differentiable and approximate F by the **linear** model function

$$F(x) \approx \tilde{F}(x) := \underbrace{F(x^{(k)}) + F'(x^{(k)})\big(x - x^{(k)}\big)}_{\text{linearization of } F \text{ at } x^{(k)}} \tag{1.61}$$

  for $x \in D$ with $x \approx x^{(k)}$.

- **Secant method** (iterative 2-point method; see Section 1.3.1 below); assume that $n = 1$ and approximate F by the **linear** model function

$$F(x) \approx \tilde{F}(x) := F(x^{(k)}) + \frac{F(x^{(k)}) - F(x^{(k-1)})}{\big(x^{(k)} - x^{(k-1)}\big)}\big(x - x^{(k)}\big) \tag{1.62}$$

  for $x \in D$ with $x \approx x^{(k)}$.

- ...

1.2.3 Newton's method

Definition 1.2.3. *Let $D \subset \mathbb{R}^n$ be an open set and let $F\colon D \to \mathbb{R}^n$ be differentiable. Then the* 1**-point iteration method** *with the*

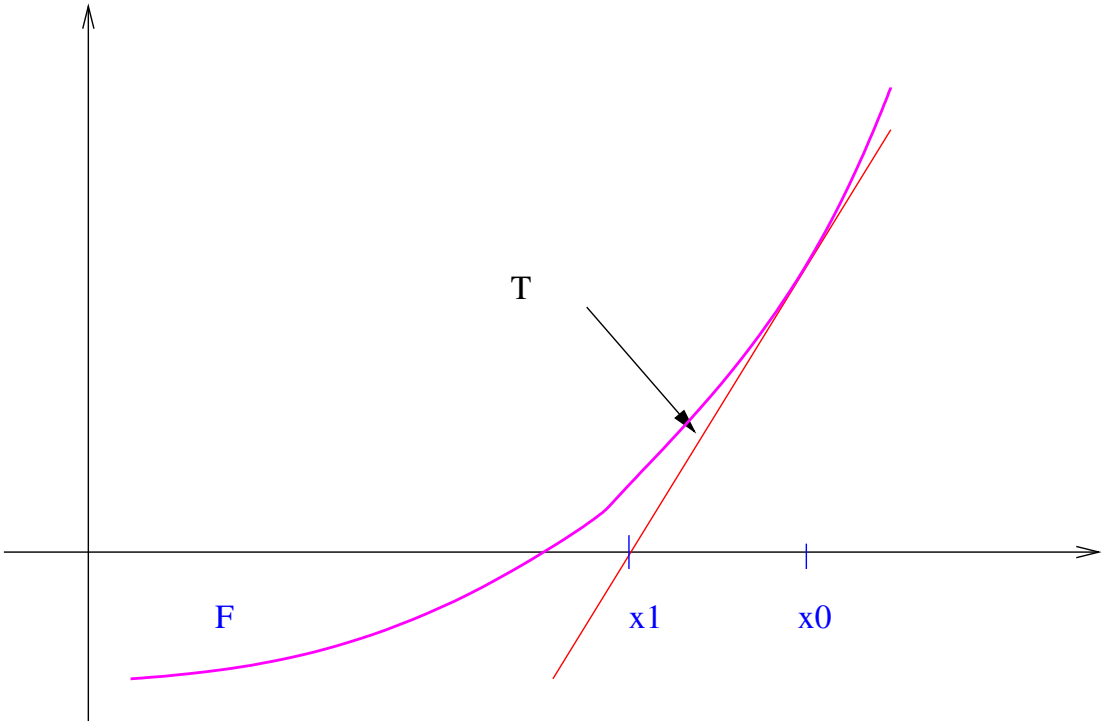
iteration function

$$\Phi\colon \{x \in D\colon F'(x) \text{ invertible}\} \to \mathbb{R}^n \tag{1.63}$$

given by

$$\Phi(x) = x - \big(F'(x)\big)^{-1} F(x) \tag{1.64}$$

for all $x \in D$ *with* $det(F'(x)) \neq 0$ *is called* **Newton's method** *for* F.



Theorem 1.2.3 (Second order convergence for Newton's method). *Let* $D \subset \mathbb{R}^n$ *be an open set, let* $F\colon D \to \mathbb{R}^n$ *be* **twice** *continuously differentiable and let* $x^* \in D$ *with*

$$F(x^*) = 0 \qquad \text{and} \qquad det(F'(x^*)) \neq 0. \tag{1.65}$$

Then there exists a $\delta \in (0, \infty)$ *such that for every*

$$x^{(0)} \in \mathcal{O} := \{x \in O\colon \|x - x^*\| \leq \delta\} \tag{1.66}$$

it holds that the iterates associated to the iteration function of **Newton's method** *for* F *and* $x^{(0)}$ *lie in* $\mathcal{O}$ *and* **converge quadratically** *to* x^* *(local quadratic convergence of Newton's method).*

Proof. Let Φ be the iteration function of Newton's method for F. Then note that

$$\Phi(x^*) = x^* - \big(F'(x^*)\big)^{-1} F(x^*) = x^* \tag{1.67}$$

and

$$\begin{aligned} \Phi'(x)v &= v - \big(F'(x)\big)^{-1} F'(x)\, v - \left[\left[\tfrac{\partial}{\partial x} \big(F'(x)\big)^{-1} \right] v \right] F(x) \\ &= - \left[\left[\tfrac{\partial}{\partial x} \big(F'(x)\big)^{-1} \right] v \right] F(x) \end{aligned} \tag{1.68}$$

for all $x \in D$ with $\det(F'(x)) \neq 0$. Hence

$$\Phi'(x^*) = 0. \tag{1.69}$$

An application of Theorem 1.2.2 thus completes the proof. □

1.2.4 Modified Newton method

Definition 1.2.4 (**Modified Newton method).** *Let $n = 1$, let $D \subset \mathbb{R}$, $U \subset \mathbb{R}$ be open intervals, let $F\colon D \to \mathbb{R}$ be a twice differentiable function and let $G\colon U \to \mathbb{R}$ be a function. Then the* 1**-point iteration method** *with iteration function*

$$\Phi\colon \left\{ x \in D\colon F'(x) \neq 0 \text{ and } \tfrac{F(x)F''(x)}{(F'(x))^2} \in U \right\} \subset \mathbb{R} \to \mathbb{R} \tag{1.70}$$

given by

$$\Phi(x) = x - \frac{F(x)}{F'(x)} G\left(\frac{F(x)F''(x)}{(F'(x))^2} \right) \tag{1.71}$$

*for all $x \in D$ with $F'(x) \neq 0$ and $\frac{F(x)F''(x)}{(F'(x))^2} \in U$ is called G-***modified Newton method** *for F.*

Remark: Under suitable assumptions, the iterates of a G-modified Newton method **converge cubically**.

1.2.5 Bisection method

Definition 1.2.5 (**Bisection method**[4]). *Let $I \subset \mathbb{R}$ be a non-empty intervall and let $G\colon I \to \mathbb{R}$ be a function. Then the iteration method with the iteration function*

$$\Phi_G\colon \left\{(a,b) \in I^2 \colon G(a)\cdot G(b) \leq 0\right\} \subset \mathbb{R}^2 \to \mathbb{R}^2 \tag{1.72}$$

given by

$$\Phi_G(a,b) = \begin{cases} \left(a, \frac{a+b}{2}\right) : & G(a)\cdot G(\frac{a+b}{2}) \leq 0 \\ \left(\frac{a+b}{2}, b\right) : & else \end{cases} \tag{1.73}$$

for all $a, b \in \mathbb{R}$ with $G(a)\cdot G(b) \leq 0$ is called **bisection method** *for G.*

Let $I \subset \mathbb{R}$ be an intervall, let $G\colon I \to \mathbb{R}$ be a function and let $a, b \in I$ with

$$G(a)\cdot G(b) \leq 0. \tag{1.74}$$

Moreover, let $x^{(0)} = (a,b)$, let Φ_G be the iteration function of the bisection method for G and let $x^{(k)} = (x_1^{(k)}, x_2^{(k)}) \in \mathbb{R}^2$, $k \in \mathbb{N}_0$, be the iterates associated to Φ_G and $x^{(0)}$. Then note that $x_1^{(k)}, x_2^{(k)} \in [a,b]$ and

$$\left[\min(x_1^{(k+1)}, x_2^{(k+1)}), \max(x_1^{(k+1)}, x_2^{(k+1)})\right] \subseteq \left[\min(x_1^{(k)}, x_2^{(k)}), \max(x_1^{(k)}, x_2^{(k)})\right],$$

$$G(x_1^{(k)})\cdot G(x_2^{(k)}) \leq 0, \qquad \left|x_1^{(k)} - x_2^{(k)}\right| = \frac{|a-b|}{2^k} \tag{1.75}$$

for all $k \in \mathbb{N}_0$. Hence, $(x_1^{(k)})_{k\in\mathbb{N}_0}$ and $(x_2^{(k)})_{k\in\mathbb{N}_0}$ converge and let

$$x^* := \lim_{k\to\infty} x_1^{(k)} = \lim_{k\to\infty} x_2^{(k)} \in [a,b]. \tag{1.76}$$

Note that

$$\min(x_1^{(k)}, x_2^{(k)}) \leq x^* \leq \max(x_1^{(k)}, x_2^{(k)}) \tag{1.77}$$

[4]It is also possible to formulate the bisection method as a iterative 2-point method.

for all $k \in \mathbb{N}_0$. Hence

$$\left|x^* - x_i^{(k)}\right| \leq \frac{|a-b|}{2^k}, \qquad \left|x^* - \left(\frac{x_1^{(k)} + x_2^{(k)}}{2}\right)\right| \leq \frac{|a-b|}{2^{(k+1)}} \quad (1.78)$$

for all $i \in \{1, 2\}$ and all $k \in \mathbb{N}_0$. Moreover, if G is continuous, then

$$0 \leq (G(x^*))^2 = G(x^*) \cdot G(x^*) = \lim_{k\to\infty} \left(G(x_1^{(k)}) \cdot G(x_2^{(k)})\right) \leq 0 \quad (1.79)$$

and hence

$$G(x^*) = 0. \qquad \textbf{(Intermediate value theorem)}$$

Listing 1.3: —MATLAB code for bisection method

```
function x = bisect(G,a,b,tol)
% Searching zero by bisection
Ga = G(a); Gb = G(b);
if (Ga*Gb>0)
   error('G(a), G(b) same sign');
end;
if (Ga < 0), v=-1; else v = 1; end;
x = 0.5*(b+a);
tol2 = 2*tol;
while ( (b-a > tol2) & ((a<x) & (x<b)) )
   if (v*G(x)<=0), b=x; else a=x; end;
   x = 0.5*(a+b)
end
```

Advantages:

- Provides a reliable a priori termination criteria
- Requires only G evaluations

Drawbacks:

- The error does in general not decay faster than $\frac{|a-b|}{2^k}$, $k \in \mathbb{N}_0$, and $\frac{|a-b|}{2^{(k+1)}}$, $k \in \mathbb{N}_0$, respectively (see (1.78)). In general

$$\min\left(\left\lceil \log_2\left(\frac{|a-b|}{2\tau}\right)\right\rceil, 0\right) \tag{1.80}$$

evaluations of G are necessary where $\tau \in (0, \infty)$ is prescribed tolerance.

Remark 1.2.3. MATLAB *function fzero is based on bisection method.*

1.3 Multi-point iteration methods

1.3.1 Secant method

Definition 1.3.1 (Secant method)**.** *Let $D \subset \mathbb{R}$ be an interval and let $F\colon D \to \mathbb{R}$ be a function. Then the* 2**-point iteration method** *with the iteration function*

$$\Phi\colon \left\{(x,y) \in D^2 \colon F(x) \neq F(y)\right\} \subset \mathbb{R}^2 \to \mathbb{R} \tag{1.81}$$

given by

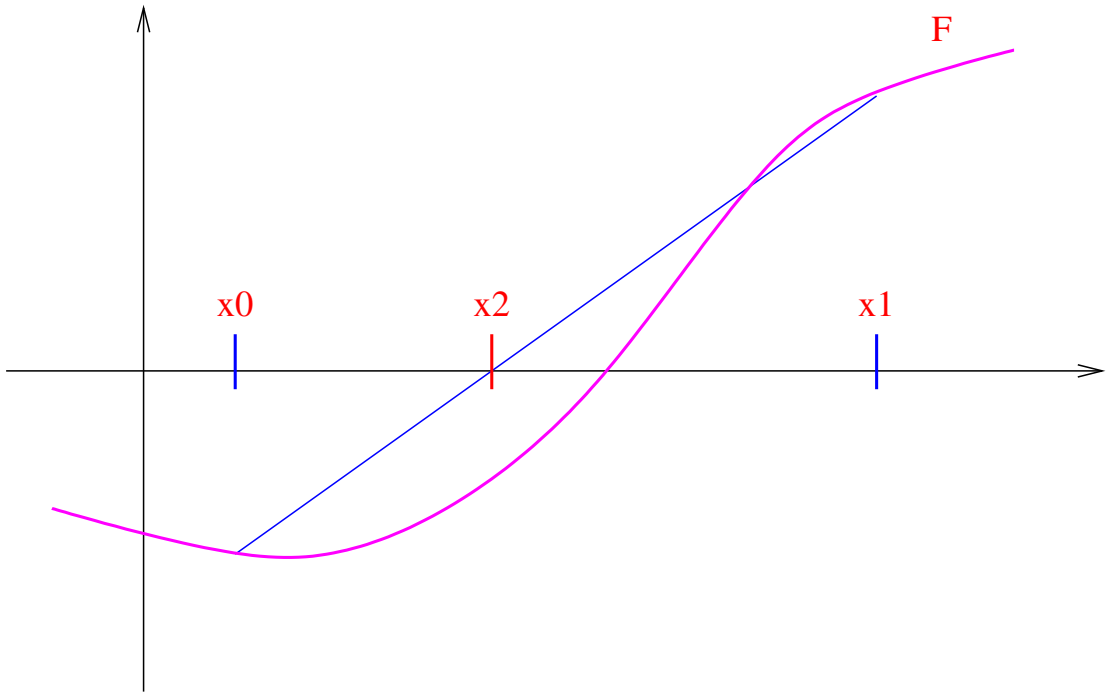
$$\Phi(x,y) = y - \frac{F(y)\,(y-x)}{(F(y)-F(x))} \tag{1.82}$$

for all $x, y \in D$ with $F(x) \neq F(y)$ is called **secant method** *for F.*

Let $n = 1$, let Φ be the iteration function of the secant method for F, let $x^{(0)}, x^{(1)} \in D$ and let $x^{(k)} \in D$, $k \in \mathbb{N}_0$, be iterates associated to Φ and $(x^{(0)}, x^{(1)})$. Then

$$x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})\left(x^{(k)} - x^{(k-1)}\right)}{\left(F(x^{(k)}) - F(x^{(k-1)})\right)} \tag{1.83}$$

for all $k \in \mathbb{N}$.



Listing 1.4: secant method

```
function x = secant(x0,x1,F,tol)
fo = F(x0);
for i=1:MAXIT
   fn = F(x1);
   s = fn*(x1-x0)/(fn-fo);
   x0 = x1; x1 = x1-s;
   if (abs(s) < tol), x = x1; return; end
   fo = fn;
end
```

Example 1.3.1 (Secant method). *Consider* $F\colon [0,\infty) \to \mathbb{R}$ *given by* $F(x) = x\,e^x - 1$ *for all* $x \in [0,\infty)$ *and* $x^{(0)} = 0$, $x^{(1)} = 5$.

k	$x^{(k)}$	$F(x^{(k)})$	$e^{(k)} := x^{(k)} - x^*$	$\frac{\log\lvert e^{(k+1)}\rvert - \log\lvert e^{(k)}\rvert}{\log\lvert e^{(k)}\rvert - \log\lvert e^{(k-1)}\rvert}$
2	0.00673794699909	-0.99321649977589	-0.56040534341070	
3	0.01342122983571	-0.98639742654892	-0.55372206057408	24.43308649757745
4	0.98017620833821	1.61209684919288	0.41303291792843	2.70802321457994
5	0.38040476787948	-0.44351476841567	-0.18673852253030	1.48753625853887
6	0.50981028847430	-0.15117846201565	-0.05733300193548	1.51452723840131
7	0.57673091089295	0.02670169957932	0.00958762048317	1.70075240166256
8	0.56668541543431	-0.00126473620459	-0.00045787497547	1.59458505614449
9	0.56713970649585	-0.00000990312376	-0.00000358391394	1.62641838319117
10	0.56714329175406	0.00000000371452	0.00000000134427	
11	0.56714329040978	-0.00000000000001	-0.00000000000000	

Advantages:

- requires only **one evalation** of F in each step; no derivatives of F required

  Remember: $F(x)$ may only be available as output of a (complicated) procedure. In this case it is difficult to find a procedure that evaluates $F'(x)$. Thus the significance of methods that do not involve evaluations of derivatives.

Drawbacks:

- converges typically with order $p = \frac{1}{2}\left(1+\sqrt{5}\right) \approx 1.62$ and not faster; is thus typically not as fast as **Newton's method**