1 Nonlinear Equations in One Unknown

1.1 Exploring Equations—A Short Tour

Let us start this course with some standard methods for solving equations in one unknown. Such equations are called *linear* if they can be expressed in the form

$$kx = d, \quad k, d \text{ given}, \quad x \text{ unknown}.$$

If $k \neq 0$, a unique solution exists. Not much else is to say in this simple case, and so, for the time being, we only care about *nonlinear* equations. (Linear equations will concern us more intensively when they appear in masses, as systems with several unknowns).

Analytical versus numerical solutions If algebraic transformations make it possible to write the solution of an equation explicitly in the form $x = \ldots$ (in the above example: $x = d/k$, in general some mathematical expression that uses a finite number of standard operations), one speaks of an *analytical solution.* However, in the vast majority of real-world problems, only *numerical solutions* are possible.

Analytically solvable are, for example, *quadratic* equations, i.e. those that can be expressed as

$$x^2 + px + q = 0 \quad p, q \text{ given}, \quad x \text{ unknown}.$$

You certainly know the quadratic formula

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$$

But it is not enough to write down a solution formula; you must also be sure it will deliver reliable results. In this example, the seemingly trivial solution of a quadratic equation calculated by the above formula can become entirely inaccurate. Let your calculator use it to find the smaller solution of the quadratic equation

$$x^2 - 12345678x + 9 = 0.$$

The exact value, up to sixteen digits, is $x_1 = 7{,}290\,000\,597\,780\,479 \times 10^{-7}$. Although common calculators work with ten to fourteen digits of precision, they return only the first few digits correctly. The numerically more accurate method first calculates the solution *larger in absolute value*, x_1, using the classical formula and then finds the second solution x_2 with the alternative solution formula

$$x_2 = \frac{q}{x_1}\,.$$

Algebraic versus transzendental equations Linear, quadratic and cubic equations are the simplest examples of *polynomial* equations. A polynomial in a variable x is a sum of powers of x, multiplied by coefficients, i.e. an expression of the form

$$a_n x^n + \cdots + a_2 x^2 + a_1 x + a_0.$$

The highest power occurring is the *order* of the polynomial or the equation. Both cubic and fourth-order equations are, in principle, analytically solvable, but the formulas (Cardano[1]-Tartaglia[2] formula, Ferrari's[3] solution) are so unwieldy that they are hardly used in practice. Numerical methods for such equations are computationally more straightforward and more elegant. They provide approximations that stepwise, with ever-increasing accuracy, approximate the solutions. For equations of order five or higher, no algebraic solution formulas exist anyway.

> In 1826, the young Norwegian mathematician Niels Henrik Abel proved the impossibility of solving the general quintic equation by some algebraic formula. Thus, from the fifth degree on, equations cannot (in general) be solved by a finite number of elementary arithmetic operations (i.e., addition, subtraction, multiplication, division, taking roots).

Let us conclude the introduction of the different types: Equations involving also fractions, roots or rational exponents can be transformed (but maybe only by cumbersome manipulations) into systems of polynomial equations. An equation containing functions that cannot be formulated through a finite number of elementary arithmetic operations is something that exceeds the powers of algebra; "quod vires algebrae transcendit", said Leibnitz) and is therefore called *transcendental*. For example, the trigonometric functions, the exponential function and the corresponding inverse functions are transcendental functions. If an equation involves algebraic and transzendental terms, normally only numerical methods can solve it.

> Explicit solution formulas exist for low-order polynomial and very simple transzendental equations only. In all other cases, only numerical methods can find solutions.

1.2 Definitions, problems, solutions

Types of problems covered here:

$$\begin{aligned} g(x) &= h(x), && \text{finding a } \textit{solution} \text{ of an equation} \\ f(x) &= 0, && \text{finding a } \textit{zero} \text{ of the function } f \\ x &= \phi(x), && \text{finding a } \textit{fixed point} \text{ of the function } \phi \end{aligned}$$

> A *zero* of the function f is a solution of the equation $f(x) = 0$.
> A *fixed point* of the function ϕ is a solution of the equation $x = \phi(x)$.

> Of course, any equation in fixed-point form $x = \phi(x)$ can be transormed into $\phi(x) - x = 0$. Thus, any fixed point of ϕ is also a zero of $f(x) = \phi(x) - x$.
> Moreover, the names f and ϕ are not reserved for problems involving zeros or fixed points, respectively. These notes, however, usually write $x = \phi(x)$ for an equation resulting from some transformation of $f(x) = 0$.

An *analytical solution*, also called a *closed-form expression*, is an explicit expression involving only well-known quantities and functions. In contrast, a *numerical solution* repeatedly uses a set of calculations to improve a known approximation step by step.

> Which functions are assumed to be "well-known" is not exactly defined. Trigonometric functions like sine or cosine definitely count as well-known, but even these can be evaluated by numerical methods only. (You just don't notice, because your calculator does this work for you.)

[1] Girolamo Cardano is also known for the Cardan shaft and the gimbal suspension, which he did not invent either.
[2] Niccolò Fontana Tartaglia, revealed the solution to Cardano under the promise to keep it secret; was extremely upset when Cardano published the formula anyway.
[3] Lodovico Ferrari was mainly responsible for the solution of quartic equations that Cardano published.

Multiple zeros: A function f has at x a root of multiplicit n, if $f(x) = 0, f'(x) = 0, f''(x) = 0, \ldots, f^{(n-1)}(x) = 0$ and $f^{(n)}(x) \neq 0$ (assuming continuous derivatives up to order n exist).

In this lecture, the functions $f, g, \ldots$ and variables $x, y, \ldots$ denote *real* quantities. The *complex* numbers, however, are the natural environment for polynomials and functions (among other things because there polynomials of degree n always have exactly n zeros; fundamental theorem of algebra). Most definitions and methods can be easily generalized for complex variables and complex-valued functions. Nevertheless, we restrict ourselves (apart from occasional hints) to computational procedures in the real numbers.

Checklist for solving nonlinear equations

Serves also as table of contents and review for the following sections.

- Preliminary work
  - Examine the shape of your functions (table of values, graphical representation).
  - Domain of the functions? Where may a solution lie? How many solutions can exist?
  - Can you find suitable transformations?
- Basic methods using computer or pocket calculator
  - Systematically compute a table of values
  - Plot the function and zoom in
- Standard methods
  - Interval bisection
  - Secant method and Regula Falsi
  - Newton's method (also known as the Newton-Raphson method)
  - Fixed point iteration

1.3 Warm-Up Examples

The exercises and the lecture discuss examples of the following type. Also the next Sections 1.5 and 1.6 present some further explanations.

From financial mathematics

A loan of € 100 000 will be repaid in 180 monthly installments of € 900 each. What is the interest rate on these terms?

The annuity formula for payment in arrears yields for the (monthly) compounding factor q the equation

$$900 = 100\,000 \frac{q - 1}{1 - q^{-180}}. \tag{1}$$

Equation of state of a real gas

What is the molar volume of nitrogen at 20 C and 1 bar $= 1 \times 10^5$ Pa according to the Van der Waals equation?

The equation of state

$$\left(p + \frac{a}{V_{mol}^2}\right)(V_{mol} - b) = RT$$

describes the relationship between pressure p, molar volume V_{mol} and Temperature T. For nitrogen, the constants a and b are

$$a = 0{,}129\,\mathrm{Pa\,m^6/mol^2}, \quad b = 38{,}6 \times 10^{-6}\,\mathrm{m^3/mol}.$$

The molar gas constant is $R = 8{,}3145\,\mathrm{J/molK}$. Inserting all numerical values leaves an equation for V_{mol},

$$\left(100\,000 + \frac{0{,}129}{V_{mol}^2}\right)(V_{mol} - 0{,}000\,038\,6) = 2437{,}4 \tag{2}$$

Friction losses in pipe flow

The friction factor f depends on the Reynolds number Re. For laminar flow, the simple rule is $f = 64/\mathrm{Re}$. In the turbulent range, from about $\mathrm{Re} > 2000$, technical manuals list different, partly empirical formulas for f. For a smooth pipe, Prandtl found the relation

$$f = \frac{1}{(2\log_{10}(\mathrm{Re}\sqrt{f}) - 0{,}8)^2}\ , \tag{3}$$

which agrees with experiments up to $\mathrm{Re} = 3{,}4 \times 10^6$. What is the value of f at $\mathrm{Re} = 1 \times 10^6$?

No deeper meaning

It is good if the examples so far have given you the impression of certain practicality. However, the technical background of the equations and the related difficulties in understanding them may obscure the view of the mathematical contents. These notes do not intend to teach physics but numerical methods, which are easier to illustrate with simple examples.

Therefore: Find the solutions to the equation

$$3\cos x = \log x \tag{4}$$

Important note: here log, of course, means the natural logarithm[4]. Arguments in trigonometric functions are always in radians!

[4] There are hardly any arguments in favor of the decadic logarithm, except for the evolutionary coincidence that humans have ten fingers. For people who cannot count to three, the base $e = 2{,}718\,281\,8\ldots$ is more natural anyway.

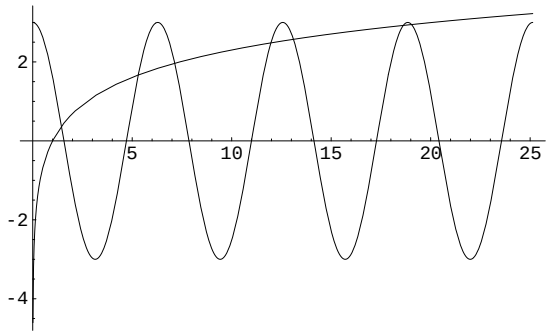

Abbildung 1: Diagram for the equation $3\cos x = \log x$. The x-values at the intersections of the graphs correspond to the solutions of the equation.

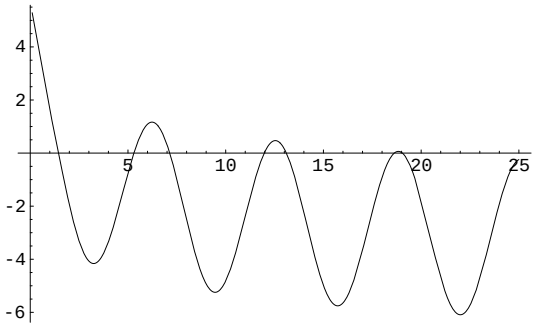

Abbildung 2: Graph of the function $f(x) = 3\cos x - \log x$. The zeros of f correspond to the x-values at the intersections in Figure 1

1.4 Graphical solution: A picture says more than a thousand formulas

According to the checklist from chapter 1.2, we use the example of Equation 4 to get a first overview. Looking at the equation, it is not immediately evident if, where and how many solutions may exist. Since cosine and logarithm are well-known functions, a graphical representation is helpful. (Figure 1). The graph immediately shows the number and approximate position of the solutions. Computational environments can easily calculate a table of values or zoom into the function graph. This way, they quickly provide suitable values. (the checklist calls these procedures “basic methods using computer or pocket calculator”).

1.5 Suitable transformations; zeroes and fixed points

The solutions of the equation $3\cos x = \log x$ are exactly the zeros of the function $f(x) = 3\cos x - \log x$. A comparison of Figure 1 with Figure 2 clarifies this fact and shows, for example, that in the vicinity of $x = 5$, at any rate in the range $4 < x < 6$, one of the zeros of f must lie.

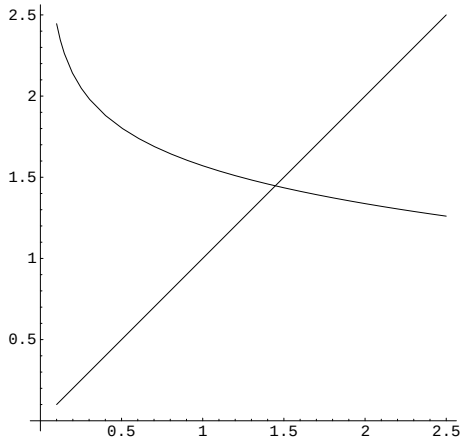

Abbildung 3: The function $\phi(x) = \arccos((\log x)/3)$ with one fixed point. It corresponds to the zero of f near 1,4. Additional fixed points of ϕ do not exist. Due to the reformulation, all other solutions to the original equation have been lost!

Which form of graphical representation is best chosen depends on the given equation. In this example, the well-known functions cos and log can be quickly sketched. Therefore, the visualization of the solution by the (x-values of) intersections is clear. On the other hand, the representation of $f(x) = 3\cos x - \log x$ lets recognize the immediately. The classical methods for finding zeros from Chapter 1.7 require such a transformation of the equation anyway.

It is, for example, also possible to formulate the equation $3\cos x = \log x$ as

$$x = \arccos \frac{\log x}{3} \quad . \tag{5}$$

In this form, it is a fixed-point problem $x = \phi(x)$, with $\phi(x) = \arccos((\log x)/3)$.

Fixed-point iteration

What happens if you insert a value for x on the right-hand side of Equation 5, evaluate the term, and substitute the result back into the right-hand side? Starting, for example, with $x = 1$, this procedure yields the sequence

$$1; \quad 1{,}5708; \quad 1{,}419\,69; \quad 1{,}453\,72; \quad 1{,}445\,76; \quad 1{,}447\,61; \quad 1{,}447\,18\ldots$$

The sequence converges to $\xi = 1{,}447\,258\,6$, which is the smallest solution of the given equation and at the same time, the only fixed point of the function

$$\phi(x) = \arccos \frac{\log x}{3}.$$

You can see here an example of a *fixed-point iteration*.

Fixed-point iteration
Given an equation $x = \phi(x)$.

Start with initial value
Insert value on right-hand side and evaluate
Repeat inserting and evaluating until values no longer change

More examples of fixed-point iterations:

- Enter a number into the calculator and press the root key repeatedly. The results converge to 1 (fixed point of $f(x) = \sqrt{x}$).

- Enter a number < 20 into the calculator and press the exp and $1/x$ keys alternately several times. The results (after the $1/x$ step) converge toward 0,567 14 (fixed point of $f(x) = 1/\exp x$).

- Calculation of square roots was an important problem already in Greek antiquity and (for rational numbers) solved. The nonlinear equation defining the square root of a is $x^2 = a$. For $x \neq 0$, this is equivalent to

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right) \ .$$

  Already the Babylonians are said to have used the iteration (often called Heron's method)

$$x^{(0)} = a; \quad x^{(k+1)} = \frac{1}{2}\left(x^{(k)} + \frac{a}{x^{(k)}}\right) \text{ for } k = 0, 1, 2, \ldots$$

- Equation 3 is a fixed-point equation. With the initial value of 0,05 few fixed-point iterations provide an accurate solution.

But it doesn't always work: another possible fixed-point form of Equation 4 is

$$x = \exp(3 \cos x) \ .$$

If you substitute here $x = 1$ on the right, evaluate and iterate, you get the sequence

$$1; \quad 5{,}057\,68; \quad 2{,}760\,46; \quad 0{,}061\,745\,5; \quad 19{,}971; \quad 3{,}6805 \ldots$$

These values change irregularly and do not converge.

Summary

Not every fixed-point iteration converges. Suitable transformations are not always easy to find. On the other hand, many numerical methods are of the fixed-point iteration type. This justifies a detailed theoretical investigation of such methods in Chapter 1.12.

1.6 Discussion of the Examples: Important and Unimportant Terms

Here we discuss in detail the examples presented in Chapter 1.1

1.6.1 A Nearly Linear Equation

The equation mentioned at the beginning of Chapter 1.1,

$$x^2 - 12345678x + 9 = 0 \ ,$$

is, when it comes to the smaller of the two solutions, actually not a quadratic equation! Reason: The sought solution is of order 10^{-6} bis 10^{-7}; the term x^2 in the equation is smaller than the linear term $12345678x$ by more than ten orders of magnitude. For all practical purposes, such an equation is just a linear one with a small quadratic correction term. Therefore, solve for the linear term:

$$x = \frac{1}{12345678}(x^2 + 9) \ .$$

The starting value$x^{(0)} = 0$ already gives, even on the cheapest calculators without a root key, a better approximation $x^{(1)} = 7{,}290\,000\,597\,78 \times 10^{-7}$ than most calculators can achieve by using the standard solution formula.

> Loosely speaking, many equations contain terms in which the unknown has little influence compared to other terms. If an equation becomes more manageable this way, you may neglect those terms in a first approximation. In the following steps, you correct the result, using the approximate values for the unimportant terms.

1.6.2 Van der Waals Equation

You may transform the Equation (2) to a cubic equation,

$$-4.9794 \cdot 10^{-6} + 0.129V_{mol} - 2441.3V_{mol}^2 + 100000V_{mol}^3 = 0 \quad , \tag{6}$$

which would be, in principle, analytically solvable. Please don't do it! A little insight into the physical background of this equation suggests a different procedure: At room temperature, nitrogen is almost an ideal gas, which obeys the equation

$$pV_{mol} = RT \ .$$

In the Van der Waals equation

$$\left(p + \frac{a}{V_{mol}^2}\right)(V_{mol} - b) = RT \ , \tag{7}$$

the term a/V_{mol}^2 is a correction of the ideal gas equation and is, compared to p, negligibly small for the given data. Even if you don't see it in the polynomial (6): the original Equation (7) does not stand for a "real" cubic equation but a linear equation in V_{mol} plus a small correction term a/V_{mol}^2. You can solve this equation if you move the "unimportant" terms to the right side. Here we reshape to

$$V_{mol} = \frac{RT}{p + a/V_{mol}^2} + b = \frac{2437{,}4}{100000 + 0{,}129/V_{mol}^2} + 0{,}000\,038\,6 \ .$$

Let's ignore, for the moment, the correction term a/V_{mol}^2. We get a zero approximation for the molar volume,

$$V_0 = \frac{2437{,}4}{100000} + 0{,}000\,038\,6 = 0{,}024\,413 \ .$$

Now, the trick is to insert this approximation for V_{mol} on the right-hand side of the equation. It produces an improved approximation

$$V_1 = \frac{2437{,}4}{100000 + 0{,}129/0{,}024\,413^2} + 0{,}000\,038\,6 = 0{,}024\,360\ .$$

Repeated insertion does not yield any further improvement:

$$V_2 = \frac{2437{,}4}{100000 + 0{,}129/0{,}024\,360^2} + 0{,}000\,038\,6 = 0{,}024\,360\ .$$

Thus we have calculated (at least to five decimal places) the value $V_{mol} = 0{,}024\,360\,\mathrm{m}^3$.

> Penitential exercise for Lent: Look up the Cardan formulae in Wikipedia and solve the problem this way. Compare the time required with the method above.

1.6.3 Financial Mathematics

In equation 1, we expect the compounding factor q to be just above 1. The term q^{-180} in the denominator is likely to be $\ll 1$ and unimportant. Thus, we solve the equation for the q in the numerator.

$$q = 1 + \frac{900}{100000}(1 - q^{-180})$$

If we ignore q^{-180} on the right side, then the zero approximation is

$$q_0 = 1 + \frac{900}{100000} = 1{,}009$$

Again, the trick works to insert q_0 on the right-hand side and gets us an improved approximation

$$q_1 = 1 + \frac{900}{100000}(1 - 1{,}009^{-180}) = 1{,}007\,206\ .$$

Repeated insertion yields

$$q_2 = 1{,}006\,529 \quad q_3 = 1{,}006\,210 \quad q_4 = 1{,}006\,047\ldots$$

However, it takes a total of 14 iterations here for the values to stabilize at 1,005 851.

Concluding Remarks

If an equation is given in the form $f(x) = g(x)$ (example: Equation 4), it is not immediately recognizable which terms are “important” or “unimportant”. Rule: ignore the unknown on that side of the equation with the *less steep* function graph at the intersection.

Suitable transformations for fixed point iterations often require a deeper understanding of the individual terms in an equation. Fortunately, black-box type solution methods exist. The next chapter presents one of them.

1.7 Bisection

Do you know the story of the two possibilities? It begins with the intermediate value theorem.

Intermediate Value Theorem

A function f that is continuous on a closed interval $[a, b]$ takes on any give value between $f(a)$ and $f(b)$ somewhere inside the interval.

In particular, if f is negative for $x = a$ and positive for $x = b$ (or vice versa), then the intermediate value theorem guarantees that f has at least one zero in this interval.

There are always two possibilities...

Suppose we are looking for a zero of a function continuous in the range $a \leq x \leq b$. We can immediately check whether $f(a)$ and $f(b)$ have different signs. If so, then the intermediate value theorem guarantees the existence of a zero in the domain $a \leq x \leq b$, but we do not know where it lies. Now there are two possibilities: Either $b - a$ is already small, in which case it is good: we can take both a and b as approximations for a zero of f. Otherwise, we calculate the midpoint c of the interval, $c = (a + b)/2$. Now there are again two possibilities. If $f(c) = 0$, it is good: We have found a zero there. Otherwise, f has different signs at the ends of one of the subintervals $a \leq x \leq c$ or $c \leq x \leq b$ (got it? That's the point!). So there must be a zero in one of the two intervals. Let's consider this interval and, for simplicity, call its boundaries a and b again.

Now there are two possibilities: Either $b - a$ is small, in which case it is good: we can take both a and b as approximations for a zero of f. Otherwise we form $c = (a + b)/2$. Now there are again two possibilities...

You can now continue the story yourself. But note that the interval length gets halved each time you take the story one step further. For any arbitrarily small given precision $\epsilon > 0$, you reach an interval with length $b - a < \epsilon$ after a finite number of steps. So, the story ends just as in real life: There may always be two choices, but each decision restricts the freedom for further actions. At some point, the alternatives are exhausted.

Written down in formalized form, this procedure is the

Bisection Method
Given a function f, two values a and b with $f(a) \cdot f(b) < 0$, and an error tolerance $\epsilon > 0$. If f is continuous in the interval $a \leq x \leq b$, then this algorithm finds the approximation c to a zero ξ of f with error $|c\xi| < \epsilon$.

```
Repeat
  set c ← (a + b)/2
  if f(a) · f(c) < 0
    set b ← c
  else
    set a ← c
until |b − a| < ε or f(c) = 0
```

Linear convergence

The best estimate for the zero is the midpoint of the interval. In this case, the error $\epsilon_0 \leq |b-a|/2$ cannot be larger than half the interval width. Interval bisection reduces this error bound by a factor of $1/2$ per step or, since

$$\left(\frac{1}{2}\right)^{3,3} \approx \frac{1}{10} \quad ,$$

by a factor of 1/10 per (average) 3,3 steps. One can say: interval bisection produces one correct decimal per 3,3 iterations. The maximum error after the i-th step, ϵ_i, is at most half as large as the previous maximum error ϵ_{i-1} . It is thus holds

$$\epsilon_i \leq C\epsilon_{i-1} \quad \text{with } C = \frac{1}{2} \ .$$

In general: If in a procedure the error bounds of successive iteration steps fulfil

$$\epsilon_i \leq C\epsilon_{i-1} \quad \text{mit } C < 1 \ .$$

this behaviour is called *linear* convergence.

Advantages and Disadvantages

Advantages of interval bisection: easy to understand and simple to program. If the assumptions are met, it converges with certainty. It is an *inclusion method*, which means that it not only provides an approximate value but also bounds the solution from both sides.

Disadvantages: One needs initial values—but that is a problem for any numerical method. Interval bisection is slow, it converges only linearly—but that is for sure.

1.8 *Regula Falsi* (false position method)

Functions running smoothly in the vicinity of a zero can be approximated there by a straight line. Instead of choosing, as with interval bisection, the value c exactly in the middle between a and b, we take c as the zero of the straight line through $(a, f(a))$ and $(b, f(b))$, see Figure 4.

$$c = a - f(a)\frac{a-b}{f(a)-f(b)} = \frac{af(b) - bf(a)}{f(b)-f(a)}$$

***Regula Falsi* (false position method)**
Given a function f, two values a and b with $f(a) \cdot f(b) < 0$ and an error tolerance $\epsilon > 0$. If $f(x)$ is continuous in the interval $a \leq x \leq b$, then this algorithm[a] finds an approximation c to a zero ξ of f with error $|c - \xi| < \epsilon$.

Repeat
  set $c \leftarrow a - f(a)\dfrac{a-b}{f(a)-f(b)}$
  if $f(b) \cdot f(c) < 0$
    set $a \leftarrow b$
  else
    (standard version) do nothing
    (Illinois version) reduce $f(a)$ to $\frac{1}{2}f(a)$
    (Pegasus version) reduce $f(a)$ to $\dfrac{f(a)f(b)}{f(b)+f(c)}$
  set $b \leftarrow c$
until $|b - a| < \epsilon$ or $f(c) = 0$

[a] however, for the stopping criterion given here, only the non-standard versions

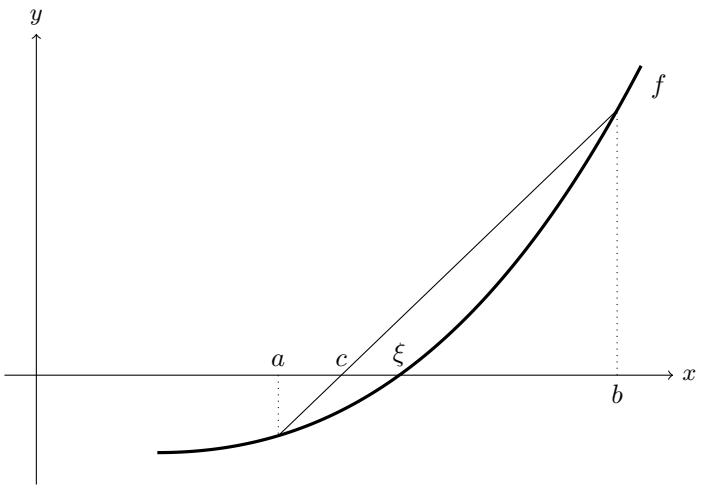

Abbildung 4: *Regula falsi* computes c, the zero of the connecting staight line, as an approximation to ξ, the zero of f.

However, compared to simple bisection, the standard version of *regula falsi* will not significantly improve the convergence behavior. Typically, aafter the first few iterations the interval boundary a will remain fixed. The other bound b will converge to the zero, but the stopping criterion $|b-a| < \epsilon$ will never be reached. Therefore, careful programmers would add an emergency exit in the algorithm above: count the number of iterations and abort if they exceed a maximum number.

The Illinois or Pegasus variants improve the convergence behavior compared to bisection; brave programmers would, in this case, dispense with the query for a maximum number of iterations.

Interval bisection and the various versions of *regula falsi* have in common that they *bracket* the zero from both sides - they are inclusion methods, which is good. The disadvantage is that at the beginning of the method, you need two approximations, one on each side of the zero. Moreover, these methods can only find zeros where the function changes sign. They will not work for multiple zeros of even order.

> What is "false" in the *regula falsi*? Not the rule itself, of course, just the assumed starting values a and b. From these two "false solutions," the rule calculates a better approximate solution.
> The method is ancient. Babylonians, Egyptians, Indians, and Chinese used it centuries before Christ to solve linear problems. From Arabic sources, Leonardo of Pisa, known as Fibonacci, brought it to Europe around 1200. He describes among several variants the *regula duarum falsarum positionum*, the "method of the two false positions." This is what it should be called, but it has been sloppily shortened to *regula falsi*.
> Fibonacci solved only linear problems with it; there, the rule calculates the correct solution from two wrong starting values immediately. The application as an iterative method for zeros of non-linear functions is not so old. Small but significant modifications (as in the Illinois or Pegasus variants) have been found around the middle of the last century. Even as recent as 2020, Oliveira und Takahashi proposed a new variant (`https://en.wikipedia.org/wiki/ITP_method`)

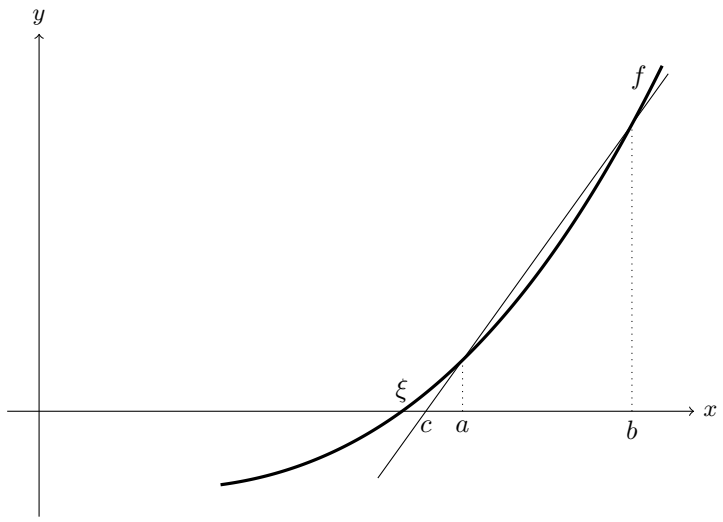

Abbildung 5: The secant method computes the next approximation c from a straight line (secant) cutting through two points in the graph of f. The values a and b do not necessarily bracket the root ξ.

1.9 Secant Method

The secant method computes a new approximation by interpolation in the same way as the Regula Falsi, but does not request that the values a and b bracket the zero, see Figure 5.

The formal description of the algorithm here denotes the initial values a and b by $x^{(0)}$ und $x^{(1)}$ and writes $x^{(k)}, x^{(k+1)}, \ldots$ for the iteratively computet approximations.

Secant Method
Given a function f, two values $x^{(0)}$ and $x^{(1)}$, an error tolerance $\epsilon > 0$, and a maximum number of iterations k_{max}. For sufficiently good initial values $x^{(0)}$ and $x^{(1)}$ this algorithm finds the approximation $x^{(k)}$ to a zero ξ of f with accuracy $|x^{(k)} - \xi| \approx \epsilon$ or terminates after a maximum number of k_{max} steps.

set $k = 1$
repeat
  set $x^{(k+1)} = x^{(k)} - f(x^{(k)}) \dfrac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}$
  increase $k = k + 1$
until $|x^{(k+1)} - x^{(k)}| < \epsilon$ or $k \geq k_{max}$

Superlinear Convergence

The secant method shows *superlinear* convergence. (Necessary technical details: f twice continuously differentiable, no multiple zeros). That is, for the error bounds $|x^{(k+1)} - \xi|$ and $|x^{(k)} - \xi|$ of successive steps, provided that $|x^{(k)} - \xi|$ is already sufficiently is sufficiently small:

$$|x^{(k+1)} - \xi| \leq C|x^{(k)} - \xi|^p \quad \text{with } p > 1 \ .$$

Thus, the error is not only reduced by a factor C but additionally by the power p. For the secant method, it can be shown that

$$p = \frac{1+\sqrt{5}}{2} \approx 1{,}618 \ .$$

Assume that $|x^{(k)} - \xi| = 0{,}01$. Consider which reduces the error more: Multiplying by a factor $C = 1/2$, or exponentiating by $p = 1{,}6$!

1.10 Newton's method

> Also known as the Newton–Raphson method. However, it was Thomas Simpson who, some decades after Isaac Newton and Joseph Raphson, formulated the method as we know it today.

We are looking for a zero of the function f. Let $x^{(0)}$ be a starting value in the vicinity of the zero. Then, the Newton method tries, similar to the secant method, to approximate the function f by a linear function and uses the tangent to f at the point $(x^{(0)}, f(x^{(0)}))$. The point of intersection of the tangent with the x-axis is the next approximation, see Figure 6.

Derivation from the Taylor expansion of f around the point $x^{(0)}$. If f is sufficiently differentiable,

$$f(x) = f(x^{(0)}) + (x - x^{(0)})f'(x^{(0)}) + \frac{(x-x^{(0)})^2}{2!}f''(x^{(0)}) + \dots$$

We want $f(x) = 0$. Neglecting higher-order terms in the expansion results in the equation

$$0 = f(x^{(0)}) + (x - x^{(0)})f'(x^{(0)}) \ ,$$

which we can solve for x,

$$x = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})} \ .$$

Newton's method
Given a differentiable function f and an initial value $x^{(0)}$.
Wanted: a zero of f.

Iteration

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \qquad \text{for } k = 0, 1, 2 \dots$$

Quadratische Konvergenz

Das Newton-Verfahren zeigt *quadratische* Konvergenz. Das heißt, für die Fehlerschranken $\epsilon_{k+1} = |x^{(k+1)} - x|$ und $\epsilon_k = |x^{(k)} - x|$ aufeinanderfolgender Schritte gilt, sofern ϵ_k schon hinreichend klein ist:

$$\epsilon_{k+1} \leq C\epsilon_k^2$$

Der neue Fehler ist also um einen Faktor C kleiner als das *Quadrat* des alten Fehlers. Der genaue Wert von C ist dabei nicht so wichtig.

Angenommen, es ist $\epsilon_k = 1 \times 10^{-4}$. Das heißt, der Fehler beträgt eine Einheit in der vierten Nachkommastelle. Dann gilt bei quadratischer Konvergenz $\epsilon_{k+1} = C \cdot 1 \times 10^{-8}$. Der Fehler beträgt also C Einheiten in der achten Nachkommastelle. Wenn C größenordnungsmäßig im Bereich 1 ist, hat sich die Anzahl der korrekten Stellen ungefähr verdoppelt.

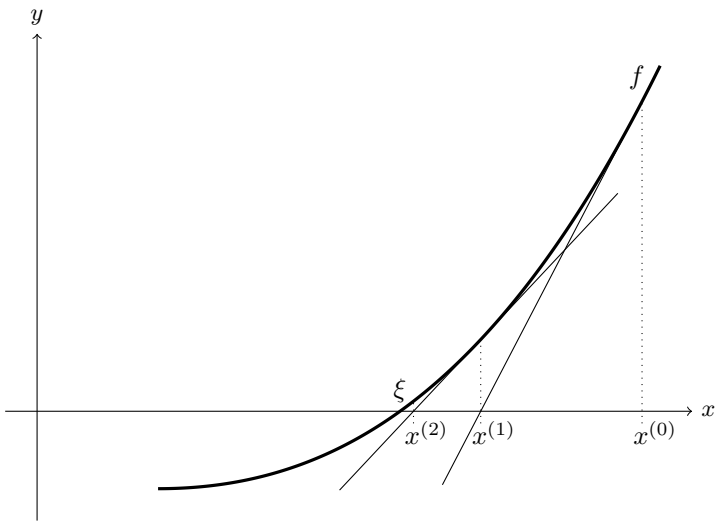

Abbildung 6: Visualization of Newton's method: The tangent to f in point $(x^{(0)}, f(x^{(0)}))$ intersects the x-axis in $x^{(1)}$. The value $x^{(2)}$ from the next step is alresdy close to the zero ξ.

Quadratische Konvergenz: Neuer Fehler $\sim$ Quadrat des alten Fehlers.

Faustregel: Sofern schon einige signifikante Stellen exakt sind, sind im nächsten Näherungswert etwa doppelt so viele signifikante Stellen korrekt.

1.11 Stopping criteria

Computers use only a fixed number of binary digits to store floating point numbers. It is possible that $f(x)$ does not reach the exact value zero for any floating point argument x. If the zero ξ is in the neighborhood of 1, you can easily find an approximation x with absolute error $|x - \xi| < 10^{-6}$. If the zero lies around $\xi \approx 10^{22}$, you will not be able to achieve an absolute error of this quality. A usual choice of the error tolerance ϵ is$\varepsilon_m(|a| + |b|)/2$ if ε_m is the machine precision and a, b are the original interval boundaries. If a, b, and the zero ξ itself are close to zero, you should apply this formula with some caution only. In any case, the termination bound must not be smaller than the smallest positive machine number (typically around 10^{-38} for 4-byte data types, 10^{-308} for 8-byte data types).

Machine precision

The machine precision ε_m is the smallest positive floating point number which, when added to the floating point number 1,0, results in a sum different from 1,0 (typically around $10^{-7}$for 4-byte data types, 10^{-16} for 8-byte data types).

1.12 Fixpunkt-Iteration

In section 1.5, we have already determined fixed points of functions by repeated insertion. Many numerical methods are just special cases of a fixed-point iteration. Therefore statements about the convergence behavior of fixed-point iterations are of general importance.

Fixed-point iteration
Given a function ϕ and an initial value $x^{(0)}$.
Wanted: a fixed point ξ von ϕ.

Iteration
$x^{(k+1)} = \phi(x^{(k)})$ for $k = 0, 1, 2 \ldots$

Fixed-point iterations converge for contraction mappings
Let ϕ be a function with fixed point ξ. Let I be an open interval $(\xi - r, \xi + r)$ around the fixed point ξ. If ϕ acts in I as a *contraction mapping*, i. e.,

$$|\phi(x) - \phi(y)| \leq C|x - y| \,, \quad C < 1 \text{ for all } x, y \in I \,,$$

then the fixed-point iteration $x^{(k+1)} = \phi(x^{(k)})$ converges for all $x^{(0)} \in I$ at least linearly to ξ.

Proof: First, we show by induction $x^{(k)} \in I$ for all $k = 0, 1, 2 \ldots$ The statement is true for $k = 0$.

If, by induction, already $x^{(k)} \in I$, this means $|x^{(k)} - \xi| < r$. We apply the contraction property to $x^{(k)}$ and the fixed point ξ and get

$$|x^{(k+1)} - \xi| = |\phi(x^{(k)}) - \phi(\xi)| \leq C|x^{(k)} - \xi| < Cr \,.$$

Since $C < 1$, it also holds that

$$|x^{(k+1)} - \xi| < r \quad \text{and thus,} \quad x^{(k+1)} \in I$$

From this argument, it also follows for the errors $\epsilon^{(k)} = |x^{(k)} - \xi|$ and $\epsilon^{(k+1)} = |x^{(k+1)} - \xi|$:

$$\epsilon^{(k+1)} \leq C\epsilon^{(k)} \leq C^k \epsilon_0, \text{ and thus, } \epsilon^{(k+1)} \to 0 \text{ for } k \to \infty \,.$$

As formulated here, the theorem already assumes the existence of a fixed point. This condition makes the convergence proof quick and easy. However, a more general formulation and a technically more elaborate argument can prove the existence and uniqueness of a fixed point from the contraction property alone. This version is the famous Banach fixed-point theorem.

Relationship between slope and contraction

The property $|\phi(x) - \phi(y)| \leq C|x - y|$ means for $C < 1$: function values differ less than input values. In the limit for small changes, the function's slope determines how much function values change in relation to input values.

If ϕ is continuously differentiable in a neighborhood of ξ and $|\phi'(\xi)| < 1$, the contraction property is satisfied in some neighborhood of ξ: Because of the continuity of ϕ' there is an open interval I around ξ in which $|\phi'| \leq C < 1$. For $x, y \in I$, according to the mean value theorem of calculus,

$$\phi(x) - \phi(y) = (x - y)\phi'(\eta) \quad \text{for some } \eta \in I \,.$$

Thus also

$$|\phi(x) - \phi(y)| \leq C|x - y| \,, \quad C < 1$$

A short version of this statement:

The fixed point method converges locally if $|\phi'(\xi)| < 1$.

Figure 7 illustrates the convergence behavior of fixed-point iteration for different functions ϕ.

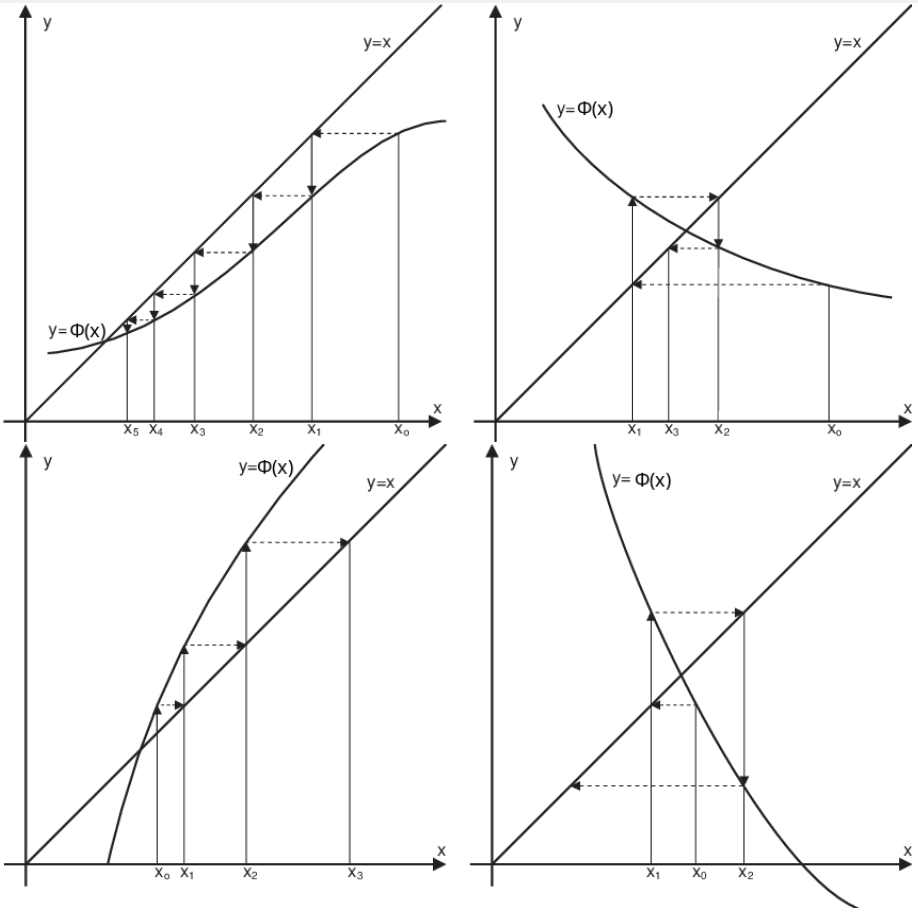

Abbildung 7: Fixed-point iterations illustrated for different functions ϕ. Possible cases: one-sided approach to the fixed point if $0 < \phi' < 1$ in a neighborhood around the fixed point; alternating convergence if $-1 < \phi' < 0$, divergence if $\phi' > 1$ or $\phi' < -1$.

1.13 Order of Convergence

We have already mentioned linear, superlinear and quadratic convergence. Here, we make the definition more precise.

Order of convergence
Let ξ be a fixed point of ϕ, and for all initial values from an interval around ξ and the corresponding sequence $\{x^{(k)}\}$ generated by the rule $x^{(k+1)} = \phi(x^{(k)}), k = 0, 1, 2 \ldots$ holds

$$|x^{(k+1)} - \xi| \leq C|x^{(k)} - \xi|^p$$

with $p \geq 1$ and $C < 1$ if $p = 1$.
Then the iteration is said to have an order of convergence of (at least) p.

For the local convergence behavior of a fixed-point iteration, the value of the first derivative at the fixed point is decisive. For $|\phi'(\xi)| < 1$, linear convergence is ensured; the smaller the magnitude of the derivative, the faster the method converges. Moreover, $C \approx |\phi'(\xi)|$. However, when $|\phi'(\xi)| = 0$, the convergence behavior becomes superlinear.

Taylor expansion can show : If $\phi(x)$ in a neighborhood of ξ is sufficiently often differentiable and

$$\phi'(\xi) = 0,\ \phi''(\xi) = 0, \ldots, \phi^{(p-1)}(\xi) = 0, \text{ and } \phi^{(p)}(\xi) \neq 0\ ,$$

then for $p = 2, 3, \ldots$ the method is of order p. If $p = 1$, first-order convergence requires the additional condition $|\phi'(\xi)| < 1$.

1.14 Convergence of Newton's Method

Newton's method, applied to the function f, corresponds to a fixed point method for the function ϕ,

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

Now,

$$\phi'(x) = \frac{f''(x)f(x)}{(f'(x))^2}\ ,$$

and since at a simple zero $f(x) = 0, f'(x) \neq 0$, the value $\phi'(x)$ vanishes there. It is easy to check that $\phi''(x) \neq 0$, provided that $f''(x) \neq 0$. From this follows the quadratic convergence of Newton's method for single zeros. For multiple zeros, linear convergence can be proved.

2 Systems of Non-Linear Equations

Section 1.2 defines the terms *solution, zero*, and *fixed point* for scalar functions $\mathbb{R} \to \mathbb{R}$. These notations can be easily generalized to vector-valued functions $\mathbb{R}^n \to \mathbb{R}^n$. As in the scalar case, there are different ways to formulate equations.

Notation for vectors and vector-valued functions: Bold

Real-valued functions, scalars: $f : \mathbb{R} \to \mathbb{R}$, $y = f(x)$
Vector-valued functions, vectors: $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{y} = \mathbf{f}(\mathbf{x})$

Components of a vector $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{or } \mathbf{x}^T = [x_1, x_2, \ldots, x_n]$$

Normally, $\mathbf{x}$ denotes a column vector while $\mathbf{x}^T$ denotes a row vector.

To count iterations, we set indices (to distinguish them from indices for vector components) as superscripts enclosen in brackets), e. g., $\mathbf{x}^{(k)}, k = 0, 1, 2 \ldots$

2.1 Solution, Zero, Fixed Point: the Multi-Dimensional Case

Types of problems for equations in $\mathbb{R}^n$

Let $\mathbf{f}, \mathbf{g}, \mathbf{h}, \mathbf{\Phi}$ be functions $\mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^n$
Find an $\mathbf{x}$ that fulfils...

$\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x})$,	(Find a *solution* for a system of equations)
$\mathbf{f}(\mathbf{x}) = 0$,	(Find a *zero* of the function $\mathbf{f}$)
$\mathbf{x} = \mathbf{\Phi}(\mathbf{x})$,	(Find a *fixed point* of the function $\mathbf{\Phi}$)

Compared to the definitions of Section 1.2, almost nothing has changed except the typeface.

For example a *nonlinear system of equations* with two unknowns

$$\begin{aligned} 4x_1 - x_2 + x_1 x_2 &= 1 \\ -x_1 + 6x_2 &= 2 - \log(x_1 x_2) \end{aligned}$$

has the form $\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x})$ with

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} 4x_1 - x_2 + x_1 x_2 \\ -x_1 + 6x_2 \end{bmatrix}, \qquad \mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(x_1, x_2) \\ h_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 - \log(x_1 x_2) \end{bmatrix}$$

This System can be transformed to

$$\begin{aligned} 4x_1 - x_2 + x_1 x_2 - 1 &= 0 \\ -x_1 + 6x_2 + \log(x_1 x_2) - 2 &= 0 \end{aligned}$$

In this formulation the problem is to find *zeros of the vector-valued function* $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$, that is, solutions of $\mathbf{f}(\mathbf{x}) = 0$ with

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} 4x_1 - x_2 + x_1 x_2 - 1 \\ -x_1 + 6x_2 + \log(x_1 x_2) - 2 \end{bmatrix}$$

Another equivalent transformation would be

$$\begin{aligned} x_1 &= \frac{1}{4}(x_2 - x_1 x_2 + 1) \\ x_2 &= \frac{1}{6}(x_1 - \log(x_1 x_2) + 2) \end{aligned}$$

Now, *fixed points of the vector-valued function* $\boldsymbol{\Phi} : \mathbb{R}^2 \to \mathbb{R}^2$ are requested, that is, solutions of $\mathbf{x} = \boldsymbol{\Phi}(\mathbf{x})$ with

$$\boldsymbol{\Phi}(\mathbf{x}) = \begin{bmatrix} \phi_1(x_1, x_2) \\ \phi_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{4}(x_2 - x_1 x_2 + 1) \\ \frac{1}{6}(x_1 - \log(x_1 x_2) + 2) \end{bmatrix}$$

One more note on notation: When we have found a particular fixed point, we denote it by $\boldsymbol{\xi}$ in the following to distinguish it from other general $\mathbf{x}$ values.

2.2 Multidimensional Fixed-Point Iterations

Fixed-point iterations are also possible in the multidimensional case. A fixed point of a mapping $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathbb{R}^n$ is—entirely analogous to the one-dimensional definition— a value $\boldsymbol{\xi} \in \mathbb{R}^n$, for which holds

$$\boldsymbol{\xi} = \boldsymbol{\Phi}(\boldsymbol{\xi}) \, .$$

Just as in the one-dimensional case, fixed-point iteration (if it converges) finds a fixed point. Once again, we write here vectors from the $\mathbb{R}^n$ and vector-valued functions in boldface type ($\boldsymbol{\Phi}, \boldsymbol{\xi}, \mathbf{x} \ldots$), to distinguish them from variables and real-valued functions ($\phi, \xi, x, \ldots$). Otherwise, nothing changes in the scheme of fixed-point iteration.

Multidimensional fixed-point iteration
Given a mapping $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{x} \to \boldsymbol{\Phi}(\mathbf{x})$.
To find a fixed point $\boldsymbol{\xi}$ of $\boldsymbol{\Phi}$,

start with initial value $\mathbf{x}^{(0)}$.
iterate
$\mathbf{x}^{(k+1)} = \boldsymbol{\Phi}(\mathbf{x}^{(k)})$ for $k = 0, 1, 2 \ldots$

Check the convergence conditions (Section 2.4)!

Example: Fixed-point iteration for a system of two nonlinear equations

Let the following system of nonlinear equations be given (where naturally, log denotes the natural logarithm).

$$\begin{aligned} 4x - y + xy - 1 &= 0 \\ -x + 6y + \log(xy) - 2 &= 0 \end{aligned}$$

Start with the approximative solution $x_0 = y_0 = 1$ and use a suitable fixed-point iteration to determine better approximations.

In the first equation and for the given initial values, the term $4x$ makes the most substantial contribution. The second equation depends most strongly on $6y$. In this situation, you should make variables x and y explicit from these equations where they have the most decisive influence.

$$\begin{aligned} x &= \frac{1}{4}(y - xy + 1) \\ y &= \frac{1}{6}(x - \log(xy) + 2) \end{aligned}$$

Here, the function $\mathbf{\Phi}$ is a vector of two real-valued functions ϕ and ψ, the vector $\mathbf{x}$ has two components x and y.

$$\mathbf{\Phi}(\mathbf{x}) = \begin{bmatrix} \phi(x,y) \\ \psi(x,y) \end{bmatrix} = \begin{bmatrix} \frac{1}{4}(y - xy + 1) \\ \frac{1}{6}(x - \log(xy) + 2) \end{bmatrix}$$

Iteration provides the sequence $(1;1)$, $(1/4;1/2)$, $(0{,}343\,75;0{,}721\,574)$, $(0{,}368\,383;0{,}622\,985)$, …, which converges to the fixed point $(0{,}353\,443\,88;0{,}639\,968\,47)$.

2.3 Norms

The exact solution, the approximate solution, and the error in systems of equations are all vectors in $\mathbb{R}^n$. We need to measure the magnitude, or length, of error vectors and also the distance of an approximation from the exact solution. In the one-dimensional case, we calculate the "size" of x by the absolute value x, and the distance between two values x and y on the real axis by $|y - x|$.

But while there is only one reasonable definition for the absolute value in $\mathbb{R}$, several possibilities are open in $\mathbb{R}^n$. First, there is the usual definition for the length of a vector, also called Euclidean length or *2-Norm* . But often, it is easier to work with other norms. We will use here the *1-Norm* and the *∞-Norm* .

Norms in $\mathbb{R}^n$ for a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \quad , \quad \text{1-norm, taxicab norm, Manhattan norm}$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i)^2} \quad , \quad \text{euklidian norm, 2-norm}$$

$$\|\mathbf{x}\|_\infty = \max_i |x_i| \quad , \quad \text{infinity norm, maximum norm}$$

Do you remember the definition of a norm from Mathematics 2?

A norm in $\mathbb{R}^n$ is a function that assigns to each vector $\mathbf{x} \in \mathbb{R}^n$ a nonnegative real number $\|\mathbf{x}\| \in \mathbb{R}_0^+$, so that $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\forall \alpha \in \mathbb{R}$ must satisfy three conditions.

- Only the zero vector has norm 0

$$\|\mathbf{x}\| = 0 \quad \Rightarrow \quad \mathbf{x} = 0$$

- Absolute value of scalar α can be factored out

$$\|\alpha \cdot \mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$$

- The triangle inequality holds

$$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

Norm and distance

A norm can also measure the distance between two points x and y.

$$\text{dist}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$$

- Cabs in Manhattan measure distances in the 1-norm.

  Therefore the 1-norm is also called taxicab norm or Manhattan norm

- Distance as the crow flies corresponds to the 2-norm.

- Biggest difference in components: ∞-norm.

Matrix norms

- The fundamental destiny of matrices is to multiply vectors.

- The result of a matrix-vector multiplication is also a vector; it is usually rotated and longer or shorter than the original vector.

- A *matrix norm* measures how strong it acts on vectors.

- A given matrix cannot extend vectors arbitrarily. For each matrix, there is a "maximum lengthening factor."

The "maximum lengthening factor" is a matrix norm.

Some matrix norms

The 1-, 2- and ∞-norms are defined via the corresponding vector norms: They specify how much matrix-vector multiplication maximally can enlarge $\mathbf{y} = A \cdot \mathbf{x}$ compared to $\mathbf{x}$. It is easy to calculate the 1-norm or the ∞-norm of a matrix.

$\|A\|_1$	*1-norm*: maximum absolute column sum
$\|A\|_\infty$	*∞-Norm*: maximum absolute row sum

Unfortunately, for the frequently used matrix 2-norm no such simple calculation rule exists.

However, MATLAB can easily calculate all norms. $\|A\|_1$ = `norm(A,1)`, $\|A\|_2$ = `norm(A)`, $\|A\|_\infty$ = `norm(A,Inf)`.

Matrix norm, general definition

You can add Matrices and multiply them by scalars. In this sense, they act precisely like vectors of $\mathbb{R}^n$. We can interpret everything that behaves like a vector as a "vector": The $m \times n$-matrices form a *vector space*. Therefore, term "norm of a matrix" can be defined in the same way as the norm of vectors of $\mathbb{R}^n$. Compare the definition of a norm in Rn on page 22 and try to find the differences—there are hardly any!

A norm in $\mathbb{R}^m \times \mathbb{R}^n$ is a function that assigns to each $m \times n$ matrix A a nonnegative real number $\|A\| \in \mathbb{R}_0^+$, so that $\forall A, B \in \mathbb{R}^m \times \mathbb{R}^n,\ \forall \alpha \in \mathbb{R}$ must satisfy three conditions.

- Only the zero matrix has norm 0

$$\|A\| = 0 \quad \Rightarrow \quad A = 0$$

- Absolute value of scalar α can be factored out

$$\|\alpha \cdot A\| = |\alpha| \cdot \|A\|$$

- The triangle inequality holds

$$\|A + B\| \leq \|A\| + \|B\|$$

These three basic rules must apply to every norm. However, there are bonus features for the 1-, 2-, or ∞-norm. For these matrix norms, the following additional rules apply.

$$\|A \cdot B\| \leq \|A\| \cdot \|B\| \tag{8}$$

$$\|A \cdot \mathbf{x}\| \leq \|A\| \cdot \|\mathbf{x}\| \tag{9}$$

Compare with the absolute value $|a \cdot b| = |a| \cdot |b|$

Frobenius norm

The Frobenius-Norm $\|A\|_F$ is calculated like the 2-norm of a vector: *square root of sum of squares*

Frobenius-Norm: $\|A\|_F = \sqrt{\sum a_{ij}^2}$

It is easier to calculate the Frobenius norm instead of the 2-Norm, and you can use it as an upper bound.

$$\|A\|_2 \leq \|A\|_F$$

Moreover, $\|A\|_F$ also provides bonus features similar to those of the 1-, 2-, or ∞-norm,

$$\|A \cdot B\|_F \leq \|A\|_F \cdot \|B\|_F \quad , \quad \|A \cdot \mathbf{x}\|_2 \leq \|A\|_F \|\mathbf{x}\|_2$$

MATLAB: $\|A\|_F =$ `norm(A,'fro')`.

Matrix norms—the small print[5]

[5] Don't think much about it; incomplete this text would be without it.

The informal explanation *"matrix norm is maximum extension factor"* is mathematically correct for 1-, 2-, and ∞-norms when vector lengths are measured in the respective norms. However, the Frobenius norm usually overestimates the maximum extension factor when vector lengths are measured in the 2-norm. Nevertheless, it provides an upper bound for the lengthening factor.
There is also another rule, $\|A\| = \max_{i,j} |a_{ij}|$, which satisfies the three conditions of a norm, but is not always an upper bound for the lengthening factor.

2.4 Convergence

As in the one-dimensional case, convergence of an n-dimensional fixed-point iteration depends on a contraction property.

Fixed-point iterations in $\mathbb{R}^n$ converge for contraction mappings
Let $\mathbf{\Phi}(x)$ be a function with fixed popint $\boldsymbol{\xi}$: $\mathbf{\Phi}(\boldsymbol{\xi}) = \boldsymbol{\xi}$. Also, let B be an open neighborhood around $\boldsymbol{\xi}$ in the form $B = \{\mathbf{x} : \|\boldsymbol{\xi} - \mathbf{x}\| < r\}, \quad r > 0$. If $\mathbf{\Phi}$ acts in B as a *contraction mapping* in some norm $\|\cdot\|$, i. e

$$\|\mathbf{\Phi}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{y})\| \leq C\|\mathbf{x} - \mathbf{y}\| \,, \quad C < 1 \text{ for all } \mathbf{x}, \mathbf{y} \in B \,,$$

then the fixed-point iteration $\mathbf{x}^{(k+1)} = \mathbf{\Phi}(\mathbf{x}^{(k)})$ converges for all $\mathbf{x}^{(0)} \in B$ at least linearly to $\boldsymbol{\xi}$.

One can prove the convergence of the multidimensional fixed point iteration in the same way as in the one-dimensional case when a contraction property holds. Also, the concept of the order of convergence can be directly applied to the multidimensional case by using norms.

Contraction and Jacobian matrix

The convergence criterion $|\phi'(\xi)| < 1$ from the one-dimensional case (compare Page 17) can be generalized to several dimensions. For this end, one collects the partial derivatives of $\mathbf{\Phi}$ in a matrix D_ϕ, called the *Jacobani matrix*.

Jacobian matrix D_ϕ of a function $\mathbf{\Phi} : \mathbb{R}^n \to \mathbb{R}^n$

$$D_\phi = \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \frac{\partial \phi_1}{\partial x_2} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \frac{\partial \phi_2}{\partial x_1} & \frac{\partial \phi_2}{\partial x_2} & \cdots & \frac{\partial \phi_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \phi_n}{\partial x_1} & \frac{\partial \phi_n}{\partial x_2} & \cdots & \frac{\partial \phi_n}{\partial x_n} \end{bmatrix}$$

Then, similar to the one-dimensional case, one can state

Fixed-point iterations converge locally,
if in the 1-,2-, Frobenius- oder ∞-norm holds

$$\|D_\phi\| < 1$$

2.5 Newton's Method for Systems of Equations

Given a vector-valued function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$. Let us find a zero of $\mathbf{f}$. The zero is a vector $\mathbf{x} \in \mathbb{R}^n$ that solves

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

This is the general formulation of a system of n linear or nonlinear equations in n unknowns. And once again, let us note: we put vectors from $\mathbb{R}^n$ and vector-valued functions in bold type $(\mathbf{x}, \mathbf{f}(\mathbf{x}), \ldots)$, as distinguished from variables and real-valued functions $(x, f(x), \ldots)$.

Component-wise written out for

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{, the system is} \quad \begin{array}{l} f_1(x_1, x_2, \ \ldots \ , x_n) = 0 \\ f_2(x_1, x_2, \ \ldots \ , x_n) = 0 \\ \qquad\qquad \ldots \\ f_n(x_1, x_2, \ \ldots \ , x_n) = 0 \end{array} .$$

Newton's method for systems reduces the solution of a nonlinear system to the solution of a sequence of linear systems. Solving linear equations is comparatively simple compared to nonlinear systems. We will treat systems of linear equations in detail later, but for the time being, we will assume that you are familiar enough with them from school.

Assuming that the corresponding partial derivatives exist, we define the *Jacobian matrix* D_f of $\mathbf{f}$ as

$$D_f = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Let us assume that an initial value $\mathbf{x}^{(0)}$ is given in the vicinity of a zero. Then Taylor's theorem approximates $\mathbf{f}$ in the neighborhood of $\mathbf{x}^{(0)}$,

$$\mathbf{f}(\mathbf{x}^{(0)} + \Delta\mathbf{x}) = \mathbf{f}(\mathbf{x}^{(0)}) + D_f(\mathbf{x}^{(0)}) \cdot \Delta\mathbf{x} + \mathbf{R}$$

with a remainder term $\mathbf{R}$ that vanishes in the limit $\Delta\mathbf{x} \to 0$ with higher order. We drop this remainder term and require $\mathbf{f}(\mathbf{x}^{(0)} + \Delta\mathbf{x}) = 0$. From the resulting equation

$$0 = \mathbf{f}(\mathbf{x}^{(0)}) + D_f(\mathbf{x}^{(0)}) \cdot \Delta\mathbf{x}$$

it is easy to determine the correction vector $\Delta\mathbf{x}$ and thus an improved approximation $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \Delta\mathbf{x}$.

Newton's method for systems
Given a differentiable vector-valued function $\mathbf{f}$ and an initial value $\mathbf{x}^{(0)}$.
Wanted a zero of $\mathbf{f}$.

iterate

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

with $\Delta\mathbf{x}^{(k)}$ as solution of $D_f(\mathbf{x}^{(k)})\Delta\mathbf{x}^{(k)} = -\mathbf{f}(\mathbf{x}^{(k)})$

Actually, this method is a fixed-point iteration for the function

$$\mathbf{\Phi}(\mathbf{x}) = \mathbf{x} - D_f^{-1}(\mathbf{x})\mathbf{f}(\mathbf{x}).$$

Of course, D_f^{-1} must exist for the method to work.

One can show: If D_f^{-1} exists at the zero then Newton's method converges quadratically for sufficiently accurate initial values.

Since it is often very tedious to calculate all elements of Df for each iteration, one sometimes computes Df just for the initial value $\mathbf{x}^{(0)}$ and keeps this D_f for the next iterations. This procedure is called the simplified Newton method. Here $\mathbf{x}^{(0)}$ should already be a useful approximation. However, this simplified Newton method converges only linearly.

The Newton method for systems requires the solution of a linear system of equations in each step. Therefore, the next chapter brings the systematic treatment of linear systems.

Example: the nonlinear system from Section 2.2

The function $\mathbf{f}$ and its Jacobian D_f are in this example

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 4x - y + xy - 1 \\ -x + 6y + \log(xy) - 2 \end{bmatrix}, \qquad D_f = \begin{bmatrix} 4 + y & -1 + x \\ -1 + \frac{1}{x} & 6 + \frac{1}{y} \end{bmatrix}.$$

Inserting the initial value $(1; 1)$ gives

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \qquad D_f = \begin{bmatrix} 5 & 0 \\ 0 & 7 \end{bmatrix}.$$

Now, Newton's method requires the solution of the linear system

$$\begin{bmatrix} 5 & 0 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = - \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

Thus, we get the correction vector and the next approximation

$$\Delta\mathbf{x}^{(0)} = \begin{bmatrix} -0,6 \\ -0,428571 \end{bmatrix}, \quad \mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \Delta\mathbf{x}^{(0)} = \begin{bmatrix} 0,4 \\ 0,571429 \end{bmatrix}.$$

The next step evaluates $\mathbf{f}$ and D_f for the new values of $\mathbf{x}$, solves the linear system for the correction term $\Delta\mathbf{x}^{(1)}$, and calculates from this the improved approximation $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \Delta\mathbf{x}^{(1)}$. However, the new matrix D_f does not have such nice entries as the initial D_f. Therefore, the system of equations is not as directly solvable as in the first step. The simplified version of Newton's method would re-evaluate $\mathbf{f}$ but keep the simpler diagonal matrix D_f of the first step. The effect would be a more straightforward calculation for the cost of only linear instead of quadratic convergence.