# Beräkningsmatematik

Stig Larsson

4 oktober 2007

# Innehåll

# Förord

Detta är föreläsningsanteckningar som kompletterar läroboken

R. A. Adams, *Calculus: A Complete Course*, Sixth Edition, Addison Wesley, 2006.

De bygger på mina föreläsningar i kurserna TMV035 1999–2006 och TMV155 2007. Jag har inspirerats av

K. Eriksson, D. Estep, and C. Johnson, *Applied Mathematics - Body and Soul*, Springer, 2003.

Tidigare anteckningar skrevs på engelska och jag ber läsaren om ursäkt för att jag inte har hunnit översätta alla till svenska ännu.

2007–09–18 `/stig`

# Kapitel 1

# The number systems

We introduce the number systems. See also Adams: P1.

## 1.1 The natural numbers

The natural numbers are

$$\mathbf{N} = \{1, 2, 3, \dots\}$$

These are the numbers that we use for counting how many elements that are contained in a set. We have two arithmetic operations ("räkneoperationer"): addition and multiplication. The sum $m + n$ is the number of elements of the set which is the union of a set with $m$ elements and a set with $n$ elements. The product $m \cdot n$ is repeated addition:

$$m \cdot n = n + n + \cdots + n \quad (m \text{ times})$$

It is easy to prove the following rules:

$$\begin{array}{lll} m + n = n + m, & m \cdot n = n \cdot m, & \text{commutative laws} \\ m + (n + p) = (m + n) + p, & m \cdot (n \cdot p) = (m \cdot n) \cdot p, & \text{associative laws} \\ m \cdot (n + p) = m \cdot n + m \cdot p, & & \text{the distributive law} \end{array} \tag{1.1}$$

The associative laws mean that we may skip the parentheses and write $m + n + p$ and $m \cdot n \cdot p$. We usually skip the $\cdot$ and write $mn$ instead of $m \cdot n$.

We also define the power ("potens") by repeated multiplication:

$$n^m = n \cdot n \cdots n \quad (m \text{ times}) \tag{1.2}$$

It useful to represent the natural numbers by marking them on the number line.

There is also a natural *order relation* ("ordningsrelation") between the natural numbers: we know what it means to say that $m$ is less than $n$, $m < n$. We may then introduce the related notation $m > n$, $m \leq n$, $m \geq n$.

There is a concept of *subtraction* for $m \geq n$, namely, $m - n$ is the number of elements that remain if we remove a subset of $n$ elements from a set of $m$ elements, with zero being the number of elements of the empty set $\emptyset$, i.e., $0 = m - m$.

Note the special roles played by the numbers 0 and 1:

$$m + 0 = m, \quad m \cdot 1 = m. \tag{1.3}$$

## 1.2 The integers

In order to solve equations of the form $m + x = n$ (with solution $x = n - m$) for arbitrary natural numbers $m, n$ we need to introduce negative numbers.

The integers ("de hela talen") are

$$\mathbf{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$$

Here we have invented new numbers as follows: 0 (zero) and for each $n \in \mathbf{N}$ a negative number denoted $-n$.

We extend the addition and the multiplication to these new numbers as follows: (here $m, n \in \mathbf{N}$)

$$m + 0 = m, \quad 0 + 0 = 0, \quad m + (-n) = \begin{cases} m - n & \text{if } m \geq n, \\ -(n - m) & \text{if } m < n, \end{cases}$$
$$m \cdot 0 = 0, \quad 0 \cdot 0 = 0, \quad m \cdot (-n) = -(m \cdot n), \quad -(m) \cdot (-n) = m \cdot n.$$

Here we relate operations involving negative numbers and zero to the corresponding operations for positive numbers. In this way all the arithmetic rules in (1.1) hold also for the integers, i.e, for $m, n, p \in \mathbf{Z}$.

The order relation, $m < n$, is also extended to all integers $m, n \in \mathbf{Z}$ as follows:

$$-n < -m \quad \text{if} \quad m > n, \ m, n \in \mathbf{N}.$$

It useful to represent these numbers by marking them on the number line.

We can now define subtraction for all integers:

$$m - n = m + (-n)$$

and we can solve the equation $m + x = n$ as follows:

$$m + x = n \Rightarrow m + x + (-m) = n + (-m) \Rightarrow x + m + (-m) = n + (-m)$$
$$\Rightarrow x + 0 = n + (-m) \Rightarrow x = n + (-m) = n - m.$$

## 1.3 The rational numbers

In order to solve equations of the form $m \cdot x = n$ (with solution $x = n/m$) for arbitrary integers $m, n$, $m \neq 0$, we need to introduce rational numbers. Since we have not yet defined the fraction $p/q$, we first define the rational numbers as the set of all pairs $x = (p, q)$ with $p, q \in \mathbf{Z}$, $q \neq 0$, where $p$ and $q$ are supposed to represent the numerator and denominator, respectively.

The rational numbers ("de rationella talen") are

$$\mathbf{Q} = \Big\{ x = (p, q) : p, q \in \mathbf{Z}, \ q \neq 0 \Big\}$$

Two rational numbers are considered to be equal if the numerator and denominator have a common factor:

$$(mp, mq) = (p, q), \quad m \in \mathbf{Z}.$$

The integers are identified with the rational numbers that have the denominator = 1:

$$p = (p, 1), \quad p \in \mathbf{Z}.$$

In particular, $(p, p) = (1, 1) = 1$.

We define addition and multiplication, for $x = (p, q)$, $y = (r, s)$, as follows:

$$x + y = (s \cdot p + r \cdot q, q \cdot s), \quad x \cdot y = (p \cdot r, q \cdot s)$$

which are suggested by the expected formulas

$$x+y=\frac{p}{q}+\frac{r}{s}=\frac{s\cdot p+r\cdot q}{q\cdot s},\quad x\cdot y=\frac{p}{q}\cdot\frac{r}{s}=\frac{p\cdot r}{q\cdot s}$$

In this way all the arithmetic rules in (1.1) hold also for the rational numbers.

We also define the inverse of $x$:

$$x^{-1}=(p,q)^{-1}=(q,p)\quad\text{for } x\neq 0$$

Note that

$$x^{-1}\cdot x=(p,q)^{-1}\cdot(p,q)=(qp,qp)=(1,1)=1.$$

We can now define division:

$$\frac{y}{x}=y\cdot x^{-1}=(r\cdot q,s\cdot p)\quad\text{for } x\neq 0$$

and we write the rational numbers in fractional form:

$$x=(p,q)=\frac{p}{q}$$

We can now solve the equation $a\cdot x=b$ for $a,b\in\mathbf{Z}$, $a\neq 0$:

$$a\cdot x=b\Rightarrow a^{-1}\cdot a\cdot x=a^{-1}\cdot b\Rightarrow 1\cdot x=a^{-1}\cdot b\Rightarrow x=a^{-1}\cdot b=\frac{b}{a}$$

The order relation, $x<y$, can also be extended to rational numbers. We note (without proof) the important implication (where $a,b,c\in\mathbf{Z}$)

$$a<b\Rightarrow\begin{cases}ca<cb & \text{if } c>0\\ ca>cb & \text{if } c<0\end{cases}\tag{1.4}$$

We define intervals of rational numbers:

$$\begin{aligned}(m,n)&=\{x\in\mathbf{Z}:m<x<n\}\\ [m,n]&=\{x\in\mathbf{Z}:m\leq x\leq n\}\\ (m,\infty)&=\{x\in\mathbf{Z}:m<x\}\\ (-\infty,n)&=\{x\in\mathbf{Z}:x<n\}\end{aligned}\tag{1.5}$$

Note that $\{x\in\mathbf{Z}:m<x<n\}$ reads "the set of all $x$ that belong to $\mathbf{Z}$ such that $x$ is between $m$ and $n$".

In order to measure the size of a rational number, irrespective of its sign, we define *absolute value* ("absolutbelopp")

$$|x|=\begin{cases}x & \text{if } x\geq 0\\ -x & \text{if } x<0\end{cases}$$

Note that $|x|$ is the distance of $x$ from zero, and $|x-y|$ is the distance from $x$ to $y$ measured along the number line.

Note the following:

$$|-x|=|x|\tag{1.6}$$

$$|xy|=|x||y|\tag{1.7}$$

$$|x|^2=x^2\tag{1.8}$$

$$x\leq|x|\tag{1.9}$$

Prove them!

The following inequality is very important.

**Sats.** *(The triangle inequality)*

$$|a+b| \leq |a| + |b|, \quad a, b \in \mathbf{Z}. \tag{1.10}$$

*Bevis.* It is easier to compute with the square instead of the absolute value, so we use (1.8) and then (1.7) and (1.9) to get

$$|a+b|^2 = (a+b)^2 = a^2 + 2ab + b^2 \leq a^2 + |2ab| + b^2 = |a|^2 + 2|a||b| + |b|^2 = (|a| + |b|)^2$$

It follows that $|a+b| \leq |a| + |b|$ if we take the square root of both sides and use the next theorem with $x = |a+b|$ and $y = |a| + |b|$. □

**Sats.** *If $x, y > 0$ then*

$$x^2 \leq y^2 \Rightarrow x \leq y. \tag{1.11}$$

*Bevis.* Let $x, y > 0$ and $x^2 \leq y^2$. Assume that the conlusion is false, i.e., assume that $y < x$. Then multiply this inequality by the positive numbers $y$ and $x$ and use (1.4) to get

$$y^2 < yx \quad \text{and} \quad xy < x^2.$$

It follows that $y^2 < x^2$, which is a contradiction ("motsägelse") to our assumption that $x^2 \leq y^2$. Hence the assumption $y < x$ leads to a contradiction and it must be false. We conclude that $x \leq y$. This kind of proof is called "proof by contradiction" ("motsägelsebevis"). □

So far we have discussed the basic properties of the integers and rational numbers. This should be familiar to you: you already know very well how to compute with these numbers.

You also know the real numbers. We need some preparations before we can introduce them. For example, we need decimal expansions.

## 1.4 Periodic decimal expansion of rational numbers

If we perform a long division ("liggande stolen") of a rational number, then two things can happen: (i) the division stops after a finite number of decimals have been generated; or (ii) the division does not stop but the decimals repeat themselves. Loosely speaking, this is because there are only finitely many possible numbers for the remainder that occurs in each step and so after a finite number of steps it we get the same remainder and the calculation repeats itself. Try this! For example:

$$\frac{3}{4} = 0.75$$
$$\frac{1}{3} = 0.33333333333\ldots$$
$$\frac{16}{7} = 2.\underbrace{285714}\underbrace{285714}\underbrace{285714}\underbrace{285714}\ldots$$

In the first case we have a finite ("ändlig") decimal expansion and the number can be expressed exactly in terms of powers of 10, e.g., $\frac{3}{4} = 7 \cdot 10^{-1} + 5 \cdot 10^{-2}$. In the other cases we have an infinite, periodic, decimal expansion and the number cannot be expressed exactly with powers of 10.

(Note, by the way, that also a finite decimal expansion can be considered to be periodic with trailing zeros repeated: $\frac{3}{4} = 0.75000\ldots$)

Suppose on the other hand that we have a periodic decimal expansion. Does it represent a rational number? If so: which number is it? Take, for example,

$$0.1818181818181818\ldots$$

Let $p_m = 0.1818\ldots18_m$ be the number that we get if we truncate it after $m$ periods:

$$\begin{aligned} p_m &= 0.181818\ldots18_m \quad (m \text{ times}) = 18\cdot 10^{-2} + 18\cdot 10^{-4} + 18\cdot 10^{-6} + \cdots + 18\cdot 10^{-2m} \\ &= 18\cdot 10^{-2}(1 + 10^{-2} + 10^{-4} + \cdots + 10^{-2m+2}) \\ &= 18\cdot 10^{-2}(1 + 10^{-2} + (10^{-2})^2 + \cdots + (10^{-2})^{m-1}) \\ &= 18\cdot 10^{-2}\frac{1-(10^{-2})^m}{1-10^{-2}} = \frac{18}{10^2-1}\big(1-(10^{-2})^m\big) = \frac{18}{99}\big(1-(10^{-2})^m\big) = \frac{2}{11}\big(1-(10^{-2})^m\big). \end{aligned}$$

Here we used the formula for a geometric sum:

$$1 + a + a^2 + \cdots + a^{m-1} = \frac{1-a^m}{1-a}, \quad a \neq 1,$$

with $a = 10^{-2}$. We find that

$$\Big|\frac{2}{11} - p_m\Big| = \frac{2}{11}\cdot 10^{-2m} < 10^{-2m}.$$

This means that the distance between the rational numbers $p_m$ and $2/11$ is less than $10^{-2m}$. In other words: $p_m$ is an approximation of $2/11$ with $2m$ correct decimals. By taking $m$ large enough we can compute a decimal approximation of $2/11$ which is correct to any number of decimals. This is what we mean when we write

$$\frac{2}{11} = 0.181818\ldots$$

More generally, let

$$0.\underbrace{q_1q_2\ldots q_n}\,\underbrace{q_1q_2\ldots q_n}\,\underbrace{q_1q_2\ldots q_n}\cdots$$

be a periodic decimal expansion and let $p_m$ be the number that we get if we truncate it after $m$ periods. A similar calculation gives

$$\Big|\frac{q_1q_2\ldots q_n}{10^n-1} - p_m\Big| < 10^{-nm}$$

and we conclude that $p_m$ approximates the rational number

$$p = \frac{q_1q_2\ldots q_n}{10^n-1}$$

to $nm$ correct decimals. We write

$$\frac{q_1q_2\ldots q_n}{10^n-1} = 0.\underbrace{q_1q_2\ldots q_n}\,\underbrace{q_1q_2\ldots q_n}\,\underbrace{q_1q_2\ldots q_n}\cdots$$

## 1.5 The real numbers

We now define the set of *real numbers* $\mathbf{R}$ as *the set of all decimal expansions*, finite, periodic, or non-periodic. This set includes the integers and the rational numbers but also many new numbers. For example,

$$\pi = 3.141592\ldots$$
$$\sqrt{2} = 1.41421356\ldots$$

It is known that these decimal expansions are not periodic, and hence that these numbers are not rational, therefore they are called irrational numbers. By the way, the word rational refers to ratio ("kvot, bråk").

When we do numerical computations on the computer, we actually compute decimal expansions. A typical numerical algorithm can compute a certain number to any desired accuracy, in other words, it can produce as many decimals from the decimal expansion as we wish. For example, the task may be to compute a certain number to a certain accuracy, for example, six correct decimals. Another time we may need ten decimals. Then we have run the algorithm again. However, unless the decimal expansion is periodic (rational number) we can not compute the whole expansion.

We will discuss the real numbers more later in the course.

## 1.6 Functions

Adams P4. We say that we have a function $f$ if for *each* element $x$ of one set $D_f$ we can find *exactly one* element $y = f(x)$ in some other set $B$. A function $f$ therefore consists of three things:

1. a rule: $x \mapsto f(x)$
2. a domain of definition ("definitionsmängd"):
$$D(f) = \{x : f(x) \text{ is defined}\}$$
3. a target set ("målmängd") $B$ where the values of the function are found.

We then write

$$f : D(f) \to B$$

and

$$f : x \mapsto y = f(x)$$

Note the different kinds of arrows for sets ($\to$) and elements ($\mapsto$).

We also define the range of $f$ ("värdemängden"):

$$R(f) = \{y \in B : y = f(x) \text{ for some } x \in D_f\}$$

It is often very difficult (and often not important) to determine exactly what $R(f)$ is. We can always find a target set, it only specifies which kind of objects the values $f(x)$ are, for example, integers or rational numbers.

In mathematics the sets $D(f)$ and $B$ are usually sets of numbers but they could be any kind of sets.

**Exempel.** $f_1(x) = x^2$, $D(f_1) = \mathbf{Z}$, $B = \mathbf{Z}$. Alternatively, we could have taken $B = \mathbf{Z}^+$ the nonnegative integers. Then $f_1 : \mathbf{Z} \to \mathbf{Z}$ and $R(f_1) = \{0, 1, 4, 9, \dots\}$ is the set of all squares. But it is not easy to determine exactly which numbers are included in this set. For example, if we are given a large nonnegative integer, we cannot easily say if it is the square of some integer.

**Exempel.** $f_2(x) = x^2$, $D(f_2) = \mathbf{Q}$, $B = \mathbf{Q}$ or $B = \mathbf{Q}^+$. Then $f_2 : \mathbf{Q} \to \mathbf{Q}$ and $R(f_1) = \{y \in \mathbf{Q} : y = x^2\}$. It is not easy to determine exactly which numbers are included in this set.

Note that these are different functions although the rule is $y = x^2$ in both cases.

Often we only specify the rule $y = f(x)$ but not $D(f)$ or $B$. Then it is understood that $D(f)$ is the largest possible set for which $f$ is defined and $B$ is obvious.

The graph of a function $f$ is the set of pairs $(x, y)$ where $x \in D(f)$ and $y = f(x)$. If these are numbers then we can plot them in the $xy$-plane.

**Exempel.** $f_3(x) = x^2$, $D(f_3) = [0, 3] \subset \mathbf{Z}$, $B = \mathbf{Z}^+$. The graph is

$$(0, 0), (1, 1), (2, 4), (3, 9)$$

**Exempel.** $f_4(x) = x^2$, $D(f_4) = [0, 2] \subset \mathbf{R}$, $B = \mathbf{R}^+$. The graph now consists of infinitely many points and we cannot compute all of them. Then we choose a stepsize $h$ and compute the points $(nh, (nh)^2)$, $n = 0, 1, 2, \ldots$, as long as $nh \leq 2$. For example, with $h = .1$

$$(0, 0), (0.1, 0.01), (0.2, 0.04), \ldots, (2, 4)$$

This is easy to do with MATLAB:

```
>>x=0:0.1:2
>>y=x.^2
>>plot(x,y)
```

A function may be considered as a "mapping" ("avbildning") or an "operator". We often use these words as synonyms to the word "function".

**Exempel.** $f(x) = -x$, $f : \mathbf{R} \to \mathbf{R}$.

Mapping: this is reflection in the origin ("spegling i origo"). For example, the interval $(1, 2)$ is mapped (reflected) to the interval (-2,-1).

Operator: the operation "multiply by $-1$" is performed.

# Kapitel 2

# Lipschitz-kontinuitet

Vi börjar med att presentera den formella definitionen av gränsvärde och kontinuitet. Vi presenterar sedan en variant av kontinuitet som är lättare att använda och som ger ett kvantitativt mått på funktionens kontinuitet.

## 2.1 Den formella definitionen av gränsvärde

(Adams 1.5)

**Definition 1.** *(Gränsvärde) (Adams 1.5 Def 8) Vi säger att*

$$\lim_{x \to a} f(x) = L$$

*om* $\forall \epsilon > 0$ $\exists \delta > 0$ *sådant att*

$$0 < |x - a| < \delta \quad \Rightarrow \quad x \in D(f) \text{ och } |f(x) - L| < \epsilon.$$

Förkortningen $\forall \epsilon > 0$ $\exists \delta > 0$ skall utläsas "för alla positiva tal $\epsilon$ existerar ett positivt tal $\delta$".

Notera att definitionen kräver att $f$ är definierad i en punkterad omgivning till $a$, dvs i en omgivning utom punkten $a$ själv ($a - \delta < x < a \cup a < x < a + \delta$).

Definitionen handlar om **noggrannhet**: Hur noggrannt, $\delta$, måste vi ange $x$ för att få en viss noggrannhet, $\epsilon$, i $y = f(x)$?

**Exempel 1.** (Adams 1.5 Ex 1) Arean för en disk är $A = \pi r^2$. Vi vill tillverka en disk med arean $400\pi$ cm$^2$ med toleransen 5 cm$^2$. Hur nära den nominella radien 20 cm måste radien vara när vi svarvar disken?

Här är $A(r) = \pi r^2$, $L = 400\pi$, $a = 20$. Vi vill ha

$$|A(r) - L| = |\pi r^2 - 400\pi| < 5 = \epsilon.$$

Vi löser ut $r - 20$:

$$\begin{aligned} -5 < \pi r^2 - 400\pi < 5 \\ 400 - 5/\pi < r^2 < 400 + 5/\pi \\ \sqrt{400 - 5/\pi} < r < \sqrt{400 + 5/\pi} \\ 19.96017 < r < 20.03975 \\ -.03983 < r - 20 < 0.03975 \end{aligned}$$

Den snävaste gränsen är den högra, så vi tar $\delta = 0.03975$. Då gäller

$$|r - 20| < 0.03975 = \delta \quad \Rightarrow \quad |\pi r^2 - 400\pi| < 5 = \epsilon.$$

□

Med den formella definitionen av gränsvärde kan vi göra definitionen av kontinuitet formell: Funktionen $f$ är kontinuerlig i en inre punkt $a$ till $D(f)$ om $\forall \epsilon > 0\ \exists \delta > 0$ sådant att

$$|x - a| < \delta \quad \Rightarrow \quad x \in D(f) \text{ och } |f(x) - f(a)| < \epsilon. \tag{2.1}$$

Den formella definitionen behövs om man ska bevisa satser om gränsvärde och kontinuerliga funktioner. Till exempel, Adams 10.2 Theorem 2 om kombination av gränsvärden och satser om kontinuerliga funktioner i Adams 10.4. När man använder definitionen måste man bestämma $\delta$ som funktion av $\epsilon$. Det är ofta svårt. Därför kommer vi att genomföra bevisen endast för en speciell klass av kontinuerliga funktioner: Lipschitz-kontinuerliga funktioner. För dessa finns ett enkelt samband mellan $\epsilon$ och $\delta$, nämligen $\epsilon = L\delta$ för någon konstant $L$.

## 2.2 Lipschitz-kontinuitet

**Definition 2.** (Lipschitz-kontinuerlig funktion.) *Funktionen $f$ är Lipschitz-kontinuerlig på intervallet $I$ med Lipschitz-konstanten $L$ om*

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2| \quad \forall x_1, x_2 \in I. \tag{2.2}$$

Olikheten (2.2) kallas Lipschitz-villkor och vi säger ofta lite slarvigt att funktionen är Lipschitz istället för Lipschitz-kontinuerlig.

Notera vad definitionen säger. Det handlar om sambandet mellan noggrannheten i $x$ och noggrannheten i $y = f(x)$. Om t ex $\hat{x}$ är en approximation till $x$ med felet högst $10^{-6}$, dvs

$$|\hat{x} - x| \leq 10^{-6},$$

och $L = 10$, så blir felet i $y = f(x)$ högst $10^{-5}$,

$$|f(\hat{x}) - f(x)| \leq L|\hat{x} - x| \leq 10 \cdot 10^{-6} = 10^{-5}.$$

Och detta gäller oavsett var i intervallet $I$ talen $\hat{x}$ och $x$ ligger. Dvs vi behöver inte veta exakt vilka $\hat{x}$ och $x$ är, det räcker med en grov uppskattning om i vilket intervall de ligger. Vi får på detta vis en kvantitativ ( = som kan mätas) information om felet.

Observera också att Lipschitz-konstanten inte är unik: om vi har hittat en Lipschitz-konstant så är varje större konstant också en Lipschitz-konstant för funktionen $f$ på intervallet $I$. Det är bättre ju mindre konstant man hittar.

**Exempel 2.** En allmän linjär funktion

$$f(x) = mx + c \quad \text{med } I = \mathbf{R} = (-\infty, \infty).$$

Vi får

$$|f(x_1) - f(x_2)| = |(mx_1 + c) - (mx_2 + c)| = |m(x_1 - x_2)| = |m||x_1 - x_2| \quad \forall x_1, x_2 \in \mathbf{R}.$$

Vi kan alltså ta $L = |m|$. □

**Exempel 3.** En speciell linjär funktion

$$f(x) = -3x + 2 \quad \text{med } I = \mathbf{R} = (-\infty, \infty).$$

Vi får

$$|f(x_1) - f(x_2)| = |-3||x_1 - x_2| = 3|x_1 - x_2| \quad \forall x_1, x_2 \in \mathbf{R}.$$

Om vi vill att $|f(x_1) - f(x_2)| \leq 10^{-3}$ så kan vi ta

$$3|x_1 - x_2| \leq 10^{-3},$$

dvs

$$|x_1 - x_2| \leq \frac{1}{3}10^{-3}.$$

□

**Exempel 4.** En speciell kvadratisk funktion

$$f(x) = x^2 \quad \text{med } I = [-2, 2].$$

Vi får, med konjugatregeln och triangelolikheten,

$$\begin{aligned}|f(x_1) - f(x_2)| &= |x_1^2 - x_2^2| = |(x_1 + x_2)(x_1 - x_2)| = |x_1 + x_2||x_1 - x_2| \\ &\leq (\underbrace{|x_1|}_{\leq 2} + \underbrace{|x_2|}_{\leq 2})|x_1 - x_2| \leq (2+2)|x_1 - x_2| = 4|x_1 - x_2| \quad \forall x_1, x_2 \in [-2, 2].\end{aligned}$$

Alltså: $f(x) = x^2$ är Lipschitz med konstanten $L = 4$ på intervallet $[-2, 2]$. □

**Exempel 5.** $f(x) = x^2$ på $[2, 4]$. Vi får, som förut,

$$\begin{aligned}|f(x_1) - f(x_2)| = |x_1^2 - x_2^2| = |x_1 + x_2||x_1 - x_2| &\leq (\underbrace{|x_1|}_{\leq 4} + \underbrace{|x_2|}_{\leq 4})|x_1 - x_2| \\ &\leq (4+4)|x_1 - x_2| = 8|x_1 - x_2| \quad \forall x_1, x_2 \in [2, 4].\end{aligned}$$

Alltså: $f(x) = x^2$ är Lipschitz med konstanten $L = 8$ på intervallet $[2, 4]$. □

Notera att Lipschitz-konstanten beror både på funktionen och intervallet. Löst uttryckt: $L$ är maximala lutningen på intervallet $I$ tagen med absolutbelopp. Vi ska senare se att om funktionen är deriverbar så är

$$L = \max_{x \in I} |f'(x)|.$$

Men alla funktioner är inte deriverbara så vill inte använda detta nu.

**Exempel 6.** $f(x) = x^2$ på $\mathbf{R}$. Denna är ej Lipschitz på det angivna intervallet för

$$|f(x_1) - f(x_2)| = |x_1 + x_2||x_1 - x_2|,$$

där kvantiteten $|x_1 + x_2|$ kan bli hur stor som helst, "$L = \infty$". □

**Exempel 7.** $f(x) = \sqrt{x}$ på $[1, \infty)$. Vi får med konjugatregeln

$$\begin{aligned}|f(x_1) - f(x_2)| = |\sqrt{x_1} - \sqrt{x_2}| &= \Big|\frac{(\sqrt{x_1} - \sqrt{x_2})(\sqrt{x_1} + \sqrt{x_2})}{\sqrt{x_1} + \sqrt{x_2}}\Big| = \Big|\frac{x_1 - x_2}{\sqrt{x_1} + \sqrt{x_2}}\Big| \\ &= \frac{1}{\sqrt{x_1} + \sqrt{x_2}}|x_1 - x_2| \leq \frac{1}{2}|x_1 - x_2| \quad \forall x_1, x_2 \in [1, \infty).\end{aligned}$$

Här använde vi att $x_1, x_2 \geq 1$ så att $\sqrt{x_1} + \sqrt{x_2} \geq 2$ och därmed

$$\frac{1}{\sqrt{x_1} + \sqrt{x_2}} \leq \frac{1}{2}$$

Alltså är $\sqrt{x}$ Lipschitz på $[1, \infty)$ med konstanten $\frac{1}{2}$.

Men $\sqrt{x}$ är inte Lipschitz på $[0, 1]$, för på det intervallet kan $\frac{1}{\sqrt{x_1} + \sqrt{x_2}}$ bli hur stor som helst. Kvadratroten är inte Lipschitz på något intervall som innehåller punkten 0. □

Nu ska vi visa att Lipschitz-kontinuerliga funktioner verkligen är kontinuerliga enligt vår gamla definition (2.1) (se även Adams 1.4 Definition 4 och 7 och Adams 1.5 Definition 8.) Obs: enkelt bevis.

**Sats 1.** *Om $f$ är Lipschitz-kontinuerlig på $I$ så är $f$ kontinuerlig på $I$.*

*Bevis.* Vad vet vi? Jo, enligt antagandet vet vi att

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2| \quad \forall x_1, x_2 \in I. \tag{2.3}$$

Vad ska vi visa? Tag $c \in I$. Vi ska visa att $f$ är kontinuerlig i $c$, dvs

$$\lim_{x \to c} f(x) = f(c).$$

Om $c$ är en ändpunkt till intervallet så ska gränsvärdet vara enkelsidigt:

$$\lim_{x \to c-} f(x) = f(c) \quad \text{eller} \quad \lim_{x \to c+} f(x) = f(c).$$

Detta betyder enligt den formella definitionen (Adams 1.5, Definition 8): $\forall \epsilon > 0\ \exists \delta = \delta(\epsilon)$ sådant att

$$0 < |x - c| < \delta \text{ och } x \in I \Rightarrow |f(x) - f(c)| < \epsilon. \tag{2.4}$$

Vi tar då ett $\epsilon > 0$ och försöker hitta $\delta$ så att (2.4) gäller. Lipschitz-villkoret (2.3) ger

$$|f(x) - f(c)| \leq L|x - c| \quad \forall x \in I.$$

Vi vill att $L|x - c| < \epsilon$. Detta gäller om

$$|x - c| < \frac{1}{L}\epsilon.$$

Vi kan alltså ta $\delta = \frac{1}{L}\epsilon$. Då gäller (2.4). Beviset är klart. □

Observera i det föregående beviset att vi kan ta $\delta = \frac{1}{L}\epsilon$, dvs att $\delta$ är proportionell mot $\epsilon$. Det är anledningen till att det är lättare att räkna med Lipschitz-kontinuerliga funktioner. Nackdelen är att inte alla kontinuerliga funktioner är Lipschitz-kontinuerliga. Det är dock ingen stor nackdel i praktiken.

**Exempel 8.** Vi återvänder till Exempel 1 om att tillverka en disk. Vi betraktar då funktionen $A(r) = \pi r^2$ på intervallet $0 \leq r \leq 25$. Vi beräknar en Lipschitz-konstant som i Exempel 5:

$$|A(r_1) - A(r_2)| \leq \pi(25 + 25)|r_1 - r_2| = 50\pi|r_1 - r_2|, \quad \forall r_1, r_2 \in [0, 25],$$

dvs $L = 50\pi$. Vi vill ha

$$|A(r) - A(20)| = |\pi r^2 - 400\pi| < \epsilon = 5.$$

Vi ska åstadkomma detta genom att ta $|r - 20| < \delta$. Tillsammans med Lipschitz-villkoret ger detta

$$|A(r) - A(20)| \leq L|r - 20| < L\delta \leq \text{ (vi vill) } \leq \epsilon.$$

Vi löser ut $\delta$:

$$\delta \leq \frac{\epsilon}{L} = \frac{5}{50\pi} = \frac{1}{10\pi} = 0.0318309...$$

Vi tar $\delta = 0.03184$. Lite sämre tolerans än i den mer exakta beräkningen i Exempel 1 men mycket enklare. □

Ett intervall är begränsat ("bounded") om det inte når ut till oändligheten, dvs

$$(a, b),\ [a, b),\ (a, b] \text{ eller } [a, b]. \tag{2.5}$$

Ett sådant intervall kan också kallas ändligt ("finite") som i Adams 10.4 sid 80. En funktion är begränsad ("bounded") på intervallet $I$ om dess värden inte når ut till oändligheten (Adams 10.4, sid 81), dvs om det finns en begränsning $M$ sådan att

$$|f(x)| \leq M \quad \forall x \in I. \tag{2.6}$$

Vi ska nu visa att en Lipschitz-kontinuerlig funktion inte hinner ut till oändligheten på ett begränsat intervall. Enkelt bevis!

**Sats 2.** *Om $f$ är Lipschitz på ett begränsat intervall $I$, så är $f$ begränsad på $I$.*

*Bevis.* Eftersom $I$ är begränsat så är det av formen (2.5). Vi måste hitta en konstant $M$ så att (2.6) gäller. Tag en punkt $c \in I$. Lipschitz-villkoret ger för alla $x \in I$:

$$|f(x)| = |f(x) - f(c) + f(c)| \leq |f(x) - f(c)| + |f(c)| \leq L|x - c| + |f(c)| \leq L(b - a) + |f(c)|.$$

Vi tar alltså $M = L(b - a) + |f(c)|$. □

Vi kan kombinera Lipschitz-kontinuerliga funktioner. Detta är Lipschitz-motsvarigheten till Adams 10.4 Theorem 6.

**Sats 3.** *Antag att $f$ och $g$ är Lipschitz på $I$ med konstanter $L_f$ och $L_g$ och $\alpha, \beta \in \mathbf{R}$. Då är följande kombinationer också Lipschitz på $I$. I (b) och (c) antar vi dessutom att $f$ och $g$ är begränsade på $I$ med begränsningar $M_f$ och $M_g$.*

*(a) linjär kombination: $\alpha f + \beta g$ med $L = |\alpha|L_f + |\beta|L_g$;*

*(b) produkt: $fg$ med $L = M_f L_g + M_g L_f$;*

*(c) kvot: $\dfrac{f}{g}$, om $|g(x)| \geq a$ $\forall x \in I$ och något $a > 0$, med $L = (M_f L_g + M_g L_f)/a^2$;*

*(d) komposition: $f \circ g$ med $L = L_f L_g$.*

*Bevis.* (Det räcker om du lär ett av dessa.) Bevis av (a):

$$\begin{aligned}
|(\alpha f + \beta g)(x_1) - (\alpha f + \beta g)(x_2)| &= |(\alpha f(x_1) + \beta g(x_1)) - (\alpha f(x_2) + \beta g(x_2))| \\
&= |\alpha(f(x_1) - f(x_2)) + \beta(g(x_1) - g(x_2))| \\
&\leq |\alpha||f(x_1) - f(x_2)| + |\beta||g(x_1) - g(x_2)| \\
&\leq |\alpha|L_f|x_1 - x_2| + |\beta|L_g|x_1 - x_2| \\
&= (|\alpha|L_f + |\beta|L_g)|x_1 - x_2|.
\end{aligned}$$

Lipschitz-konstanten blir $L = |\alpha|L_f + |\beta|L_g$.

Bevis av (b):

$$\begin{aligned}
|(fg)(x_1) - (fg)(x_2)| &= |f(x_1)g(x_1) - f(x_2)g(x_2)| \\
&= |f(x_1)g(x_1) - f(x_1)g(x_2) + f(x_1)g(x_2) - f(x_2)g(x_2)| \\
&\leq |f(x_1)g(x_1) - f(x_1)g(x_2)| + |f(x_1)g(x_2) - f(x_2)g(x_2)| \\
&\leq |f(x_1)||g(x_1) - g(x_2)| + |f(x_1) - f(x_2)||g(x_2)| \\
&\leq M_f L_g|x_1 - x_2| + L_f|x_1 - x_2|M_g \\
&\leq (M_f L_g + M_g L_f)|x_1 - x_2|.
\end{aligned}$$

Lipschitz-konstanten blir $L = M_f L_g + M_g L_f$.

Bevis av (c):

$$\begin{aligned}
\Big|\frac{f}{g}(x_1)-\frac{f}{g}(x_2)\Big| &= \Big|\frac{f(x_1)}{g(x_1)}-\frac{f(x_2)}{g(x_2)}\Big| \\
&= \Big|\frac{f(x_1)g(x_2)-f(x_2)g(x_1)}{g(x_1)g(x_2)}\Big| \\
&= \Big|\frac{f(x_1)g(x_2)-f(x_2)g(x_2)+f(x_2)g(x_2)-f(x_2)g(x_1)}{g(x_1)g(x_2)}\Big| \\
&= \Big|\frac{(f(x_1)-f(x_2))g(x_2)+f(x_2)(g(x_2)-g(x_1))}{g(x_1)g(x_2)}\Big| \\
&= \frac{|(f(x_1)-f(x_2))g(x_2)+f(x_2)(g(x_2)-g(x_1))|}{|g(x_1)||g(x_2)|} \\
&\le \frac{|f(x_1)-f(x_2)||g(x_2)|+|f(x_2)||g(x_2)-g(x_1)|}{|g(x_1)||g(x_2)|} \\
&\le \frac{(L_fM_g+M_fL_g)|x_1-x_2|}{a^2} \\
&= \frac{(M_fL_g+M_gL_f)}{a^2}|x_1-x_2|.
\end{aligned}$$

Lipschitz-konstanten blir $L=(M_fL_g+M_gL_f)/a^2$.

Bevis av (d):

$$\begin{aligned}
|(f\circ g)(x_1)-(f\circ g)(x_2)| &= |f(g(x_1))-f(g(x_2))| \\
&\le L_f|g(x_1)-g(x_2)| \le L_fL_g|x_1-x_2|.
\end{aligned}$$

Lipschitz-konstanten blir $L=L_fL_g$. □

Om funktionen är deriverbar kan man beräkna Lipschitz-konstanten enkelt.

**Sats 4.** (Beräkning av Lipschitz-konstant med hjälp av derivata) *Antag att funktionen $f$ är kontinuerlig på $[a,b]$ och deriverbar på $(a,b)$ med begränsad derivata,*

$$|f'(x)|\le M \quad \forall x\in(a,b).$$

*Då gäller*

$$|f(x_1)-f(x_2)|\le M|x_1-x_2| \quad \forall x_1,x_2\in[a,b].$$

Slutsatsen är alltså att $f$ är Lipschitz-kontinuerlig på $[a,b]$ med konstanten $L\le M$.

*Bevis.* Tag två punkter $x_1,x_2\in[a,b]$ med $x_1<x_2$. Tillämpa Medelvärdessatsen (Adams 2.6 Theorem 11) på intervallet $[x_1,x_2]$. Vi får en (okänd) punkt $c\in(a,b)$ sådan att

$$\frac{f(x_2)-f(x_1)}{x_2-x_1}=f'(c),$$

dvs

$$|f(x_2)-f(x_1)|=|f'(c)(x_2-x_1)|=|f'(c)||x_2-x_1|\le M|x_2-x_1|.$$

□

## Övningar

Bestäm en Lipschitz-konstant för följande funktioner. Gör både ett direkt bevis (om det går) och ett som baseras på Sats 3.

1. $h(x) = x^3$ på $[0, 2]$
2. $h(x) = \dfrac{1}{x}$ på $[1, 10]$
3. $h(x) = \dfrac{1}{x^2}$ på $[1, 10]$
4. $h(x) = \sqrt{x}$ på $[0.01, 1]$
5. $h(x) = 4x^2 - 3x$ på $[-1, 1]$
6. $h(x) = \frac{x^2}{x+1}$ på $[0, 1]$

## Svar

1. $L = 12$
2. $L = 1$
3. grov uppskattning $L = 20$, bästa möjliga $L = 2$

$$\Big|\frac{1}{x^2} - \frac{1}{y^2}\Big| = \Big|\frac{(x+y)(y-x)}{x^2y^2}\Big| = \frac{|x+y||y-x|}{x^2y^2} = \frac{x+y}{x^2y^2}|x-y| \le \frac{10+10}{1\cdot 1}|x-y| = 20|x-y|$$

$$\Big|\frac{1}{x^2} - \frac{1}{y^2}\Big| = \Big|\frac{(x+y)(y-x)}{x^2y^2}\Big| = \frac{x+y}{x^2y^2}|x-y| = \Big(\frac{1}{x} + \frac{1}{y}\Big)|x-y| \le (1+1)|x-y| = 2|x-y|$$

4. $L = 5$
5. $L = 11$
6. $L = 5$, $a = 1$, $M_g = 2$, $L_g = 1$, $M_f = 1$, $L_f = 2$

Obs: dessa konstanter är kanske inte de bästa (minsta) möjliga, men det räcker hitta en $L$ som inte är överdrivet stor.

# Kapitel 3

# Talföljder

Vi introducerar talföljder och konvergens av talföljder. Se också Adams 9.1.

## 3.1 Definition av talföljd

En talföljd är en oändlig följd av tal som räknas upp i en bestämd ordning, t ex,

$$1, 2, 3, 4, ...,$$
$$1, -\tfrac{1}{2}, \tfrac{1}{4}, -\tfrac{1}{8}, \tfrac{1}{16}, ...$$

Vi betecknar hela följden med en bokstav och numrerar termerna med de naturliga talen

$$a = \{a_1, a_2, a_3, ...\} = \{a_k\}_{k=1}^{\infty}.$$

Exemplen ovan kan då skrivas

$$a = \{1, 2, 3, ...\} = \{k\}_{k=1}^{\infty},$$
$$b = \{1, -\tfrac{1}{2}, \tfrac{1}{4}, -\tfrac{1}{8}, \tfrac{1}{16}, ...\} = \{(-\tfrac{1}{2})^{k-1}\}_{k=1}^{\infty}.$$

Här ges alltså $a_k$ av en formel, $a_k = k$, men också av en algoritm (en rekursion) $a_1 = 1$, $a_k = a_{k-1} + 1$. På samma vis har vi $b_k = (-\frac{1}{2})^k$ och rekursionen $b_1 = 1$, $b_k = -\frac{1}{2}b_{k-1}$. Anledningen till att vi studerar talföljder är att våra matematiska (och numeriska) algoritmer ofta genererar talföljder.

De första, säg 100, termerna av följden $b$ genereras enkelt med MATLAB:

```
>> k=1:100;
>> b=(-0.5).^(k-1);
>> b=b'
```

Följden $a$ ovan är ett exempel på en *aritmetisk talföljd med differensen d*:

$$a_k = a_{k-1} + d,$$

medan $b$ är exempel på en *geometrisk talföljd med kvoten q*:

$$a_k = qa_{k-1}.$$

En talföljd kan också ses som en funktion från de naturliga talen till de reella:

$$f : \mathbf{N} \to \mathbf{R}, \quad f(n) = a_n.$$

## 3.2 Konvergens, gränsvärde

**Definition 3.** *Vi säger att följden* $a = \{a_n\}_{n=1}^{\infty}$ *är konvergent med gränsvärdet* $L$*,*

$$\lim_{n\to\infty} a_n = L,$$

*om för varje tal* $\epsilon > 0$ *finns ett naturligt tal* $N$ *sådant att*

$$n \geq N \Rightarrow |a_n - L| < \epsilon.$$

Vi skriver också

$$a_n \to L \quad \text{då } n \to \infty,$$

vilket utläses "$a_n$ går mot $L$ då $n$ går mot oändligheten". Observera att $N$ beror på $\epsilon$. Att $a = \{a_n\}_{n=1}^{\infty}$ är konvergent med gränsvärdet $L$ betyder i praktiken att vi kan approximera $L$ med godtycklig noggrannhet med hjälp av följden $a$. Talet $\epsilon$ är approximationsfelet, t ex, $\epsilon = 10^{-6}$ betyder att vi har 5 decimalers noggrannhet. Talet $N$ anger hur många steg av algoritmen som genererar $a_n$ som vi måste utföra för att uppnå denna noggrannhet.

Tre grundläggande exempel:

**Exempel 9.** Konstant följd: $a_n = 1$. Då gäller

$$\lim_{n\to\infty} 1 = 1.$$

*Bevis.* Tag ett tal $\epsilon > 0$ och bestäm $N = N(\epsilon)$. Vi har

$$|a_n - L| = |1 - 1| = 0 < \epsilon$$

för alla $n$. Vi kan alltså ta $N = 1$. □

**Exempel 10.** Följden $a_n = 1/n$, dvs $a = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, ....$ Då gäller

$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

*Bevis.* Tag ett tal $\epsilon > 0$ och bestäm $N = N(\epsilon)$. Vi har

$$|a_n - L| = |\tfrac{1}{n} - 0| = \tfrac{1}{n} < \epsilon$$

om $n > 1/\epsilon$. Vi kan ta $N = \lceil \epsilon^{-1} \rceil$. (Här är $\lceil x \rceil$ heltalstakfunktionen, som avrundar uppåt till heltal, se Adams P.5, i MATLAB `ceil(x)`.) □

**Exempel 11.** Geometrisk följd: $a_n = q^n$, dvs $a = \{q, q^2, q^3, ...\}$. Då gäller

$$\lim_{n\to\infty} q^n = 0 \quad \text{om och endast om } |q| < 1.$$

*Bevis.* Antag $|q| < 1$. Tag ett tal $\epsilon > 0$ och bestäm $N = N(\epsilon)$. Vi har

$$|a_n - L| = |q^n - 0| = |q^n| = |q|^n < \epsilon,$$

vilket är ekvivalent med $n \ln(|q|) = \ln(|q|^n) < \ln(\epsilon)$, där vi använt att logaritmen är en strängt växande funktion. Att $|q| < 1$ medför att $\ln(|q|) < 0$ så att vårt villkor är ekvivalent med

$$n > \frac{\ln(\epsilon)}{\ln(|q|)}.$$

Vi kan ta $N = \lceil \frac{\ln(\epsilon)}{\ln(|q|)} \rceil$. Alltså: $q^n \to 0$.

Antag sedan $|q| \geq 1$. Vi har

$$|a_n - L| = |q^n - 0| = |q^n| = |q|^n \geq 1,$$

vilket aldrig kan bli mindre än ett litet $\epsilon$. Alltså: $q^n$ går inte mot 0. □

Till exempel: $q = -\frac{1}{2}$, $\epsilon = 10^{-6}$ ger

$$N = \left\lceil \frac{\ln(\epsilon)}{\ln(|q|)} \right\rceil = \left\lceil \frac{\ln(10^{-6})}{\ln(\frac{1}{2})} \right\rceil = \left\lceil \frac{-6\ln(10)}{-\ln(2)} \right\rceil \approx \lceil 19.3 \rceil = 20.$$

Det vill säga: $n \geq 20 \Rightarrow |(-\frac{1}{2})^n| \leq 10^{-6}$. □

En följd kallas *divergent* om den inte är konvergent.

**Exempel 12.** Följden $\{(-1)^n\}_{n=1}^{\infty} = \{-1, 1, -1, 1, ...\}$ har inget gränsvärde, den är divergent. □

Denna följd divergerar eftersom den hoppar mellan två värden och därför inte närmar sig något. Ett annat sätt att divergera är att följden växer obegränsat, dvs divergerar mot oändligheten.

**Definition 4.** *Följden $a = \{a_n\}_{n=1}^{\infty}$ divergerar mot oändligheten,*

$$\lim_{n\to\infty} a_n = \infty \quad (\textit{eller } a_n \to \infty),$$

*om för varje naturligt tal $M$ finns ett naturligt tal $N$ sådant att*

$$n \geq N \Rightarrow a_n > M.$$

På liknande sätt definieras divergens mot minus oändligheten,

$$\lim_{n\to\infty} a_n = -\infty \quad (\text{eller } a_n \to -\infty).$$

**Exempel 13.**

$$\lim_{n\to\infty} n = \infty.$$

*Bevis.* Tag $M$ och bestäm $N = N(M)$. Vi har

$$a_n = n > M$$

om $n \geq N = \lceil M \rceil$. □

## 3.3 Kombination av gränsvärden

**Sats 5.** *Antag att $\{a_n\}$ och $\{b_n\}$ är konvergenta följder och $\alpha, \beta \in \mathbf{R}$. Då gäller*

$$\lim_{n\to\infty} (\alpha a_n + \beta b_n) = \alpha \lim_{n\to\infty} a_n + \beta \lim_{n\to\infty} b_n \qquad \textit{(linjär kombination)},$$

$$\lim_{n\to\infty} (a_n b_n) = \Big(\lim_{n\to\infty} a_n\Big)\Big(\lim_{n\to\infty} b_n\Big) \qquad \textit{(produkt)},$$

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n} \quad \textit{om } \lim_{n\to\infty} b_n \neq 0 \qquad \textit{(kvot)}.$$

Vi bevisar inte denna sats. Den gör det möjligt att kombinera de grundläggande gränsvärdena i Exempel 9–11 och få stort antal nya gränsvärden.

**Exempel 14.** $\lim_{n\to\infty} n^{-2} = 0$. Bevis: Exempel 10 och produktregeln ger

$$\lim_{n\to\infty} n^{-2} = (\lim_{n\to\infty} n^{-1})^2 = 0^2 = 0.$$

□

**Exempel 15.** $\lim_{n\to\infty}(a_n - L) = 0$ är ekvivalent med $\lim_{n\to\infty} a_n = L$. Det räcker alltså att beräkna gränsvärden som är noll. Bevis: Linjärkombination och Exempel 9 ger

$$\lim_{n\to\infty} (a_n - L) = \lim_{n\to\infty} (a_n - L \cdot 1) = \lim_{n\to\infty} a_n - L \lim_{n\to\infty} 1 = \lim_{n\to\infty} a_n - L.$$

□

# Kapitel 4

# Bisektionsalgoritmen

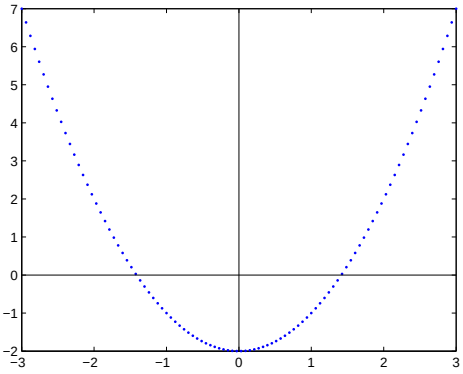
Vi ska konstruera lösningar till algebraiska ekvationer av formen $f(x) = 0$ med hjälp av *bisektionsalgoritmen* (intervallhalveringsmetoden). På samma gång ska vi se hur man definierar de reella talen och bevisar *Bolzanos sats* och *Satsen om mellanliggande värden* (Adams 1.4, Theorem 9). Vi gör detta först i form av ett exempel: kvadratroten ur 2.

## 4.1 Kvadratroten ur 2

Figur 4.1 visar grafen till funktionen $f(x) = x^2 - 2$. Grafen antyder att ekvationen

$$x^2 - 2 = 0$$

har exakt en positiv lösning (och en negativ). Den positiva lösningen är naturligtvis $\sqrt{2}$. Men vi




Figur 4.1: Funktionen $f(x) = x^2 - 2$.

börjar med att visa att $\sqrt{2}$ inte är ett rationellt tal och därför inte kan representeras exakt på datorn. Sedan skall vi konstruera och definiera talet $\sqrt{2}$ som ett nytt slags tal: reellt tal.

**Påstående.** *Ekvationen $x^2 - 2$ har ingen rationell lösning.*

*Bevis.* Antag att $x \in \mathbf{Q}$ (rationellt tal) uppfyller $x^2 - 2 = 0$. Vi skriver $x = p/q$ där $p, q \in \mathbf{Z}$ (hela tal) och där vi förkortat så att $p$ och $q$ inte har några gemensamma faktorer. Ekvationen $x^2 - 2 = 0$ ger då

$$p^2 = 2q^2,$$

vilket betyder att $p$ innehåller faktorn 2, dvs $p = 2r$ för något heltal $r$. Men då blir $x = 2r/q$ och $x^2 = 2$ ger

$$\frac{4r^2}{q^2} = 2, \quad \text{dvs} \quad q^2 = 2r^2,$$

så att även $q$ är delbart med 2. Men detta är en motsägelse till vårt antagande att vi har förkortat $p/q$. Alltså kan inte $x$ vara ett rationellt tal. □

## Konstruktion av $\sqrt{2}$

Låt $f(x) = x^2 - 2$. Vi söker $\bar{x}$ sådan att $f(\bar{x}) = 0$. Vi ser att

$$f(1) = -1 < 0, \quad f(2) = 2 > 0,$$

och drar slutsatsen att $\bar{x} \in [1, 2]$. Alltså borde mittpunkten 1.5 vara närmare $\bar{x}$. Låt nu

$$x_0 = 1, \quad X_0 = 2, \quad \hat{x}_0 = 1.5.$$

Vi ser att $f(\hat{x}_0) = 2.25 - 2 = 0.25 > 0$ och roten bör ligga i intervallet $[1, 1.5]$. Låt då

$$x_1 = x_0 = 1, \quad X_1 = \hat{x}_0 = 1.5.$$

Det aktuella intervallet är nu $[x_1, X_1] = [1, 1.5]$. Vi bildar mittpunkten

$$\hat{x}_1 = (x_1 + X_1)/2 = 1.25.$$

Nu är $f(\hat{x}_1) = (1.25)^2 - 2 = -0.4375 < 0$. Vi sätter då

$$x_2 = \hat{x}_1 = 1.25, \quad X_1 = X_1 = 1.5.$$

Vi har tabellen

| $i$ | $x_i$ | $X_i$ |
|---|---|---|
| 0 | 1 | 2 |
| 1 | 1 | 1.5 |
| 2 | 1.25 | 1.5 |

Proceduren kan upprepas hur många gånger som helst. Den kallas *bisektionsalgoritmen*. Kör programmet `bisectdemo.m` med funktionsfilen `funk.m` för att göra en längre tabell. (Filerna finns under länken "matlab/facit" på kurshemsidan.) Det verkar som om en decimalutveckling växer fram.

| $x_i$ | $X_i$ |
|---|---|
| 1.00000000000000 | 2.00000000000000 |
| 1.00000000000000 | 1.50000000000000 |
| 1.25000000000000 | 1.50000000000000 |
| 1.37500000000000 | 1.50000000000000 |
| 1.37500000000000 | 1.43750000000000 |
| 1.40625000000000 | 1.43750000000000 |
| 1.40625000000000 | 1.42187500000000 |
| 1.41406250000000 | 1.42187500000000 |
| 1.41406250000000 | 1.41796875000000 |
| 1.41406250000000 | 1.41601562500000 |
| 1.41406250000000 | 1.41503906250000 |
| 1.41406250000000 | 1.41455078125000 |
| 1.41406250000000 | 1.41430664062500 |
| 1.41418457031250 | 1.41430664062500 |
| 1.41418457031250 | 1.41424560546875 |
| 1.41418457031250 | 1.41421508789063 |
| 1.41419982910156 | 1.41421508789063 |
| 1.41420745849609 | 1.41421508789063 |
| 1.41421127319336 | 1.41421508789063 |
| 1.41421318054199 | 1.41421508789063 |
| 1.41421318054199 | 1.41421413421631 |
| 1.41421318054199 | 1.41421365737915 |
| 1.41421341896057 | 1.41421365737915 |
| 1.41421353816986 | 1.41421365737915 |
| 1.41421353816986 | 1.41421359777451 |
| 1.41421353816986 | 1.41421356797218 |
| 1.41421355307102 | 1.41421356797218 |

Man kan ställa sig följande frågor:

1. Hur kan vi veta att det är en decimalutveckling? (Vi kan ju aldrig beräkna alla decimalerna, bara ändligt många.)
2. Hur kan vi veta att decimalutvecklingen löser ekvationen?
3. Finns det någon annan decimalutveckling som löser ekvationen?

## Svar på Fråga 1

Bisektionsalgoritmen konstruerar två följder $\{x_i\}_{i=0}^{\infty}$ och $\{X_i\}_{i=0}^{\infty}$ av rationella tal. Om $j > i$ förhåller de så här:

$$x_i \leq x_j \leq \bar{x} \leq X_j \leq X_i.$$

Vi drar slutsatsen att vi har följande avstånd:

$$|x_i - X_i| = 2^{-i} \quad \text{(vi har halverat intervallet } [1,2] \text{ } i \text{ gånger)} \tag{4.1}$$

$$|x_i - x_j| \leq 2^{-i} \quad \text{för } j \geq i, \tag{4.2}$$

$$|X_i - X_j| \leq 2^{-i} \quad \text{för } j \geq i. \tag{4.3}$$

Olikheterna (4.2) och (4.3) innebär att $\{x_i\}_{i=0}^{\infty}$ och $\{X_i\}_{i=0}^{\infty}$ bildar decimalutvecklingar: om $2^{-i} < 10^{-N-1}$ så har vi $N$ fixerade decimaler. Låt oss räkna vi ut hur många steg av algoritmen vi ska köra för att få $N$ decimaler. Eftersom $2^{10} = 1024 > 10^3$ så får vi

$$2^{-i} = (2^{-10})^{i/10} = 10^{-3i/10} \leq \text{ (vi vill) } \leq 10^{-N-1}.$$

Om vi tar $i \geq 3(N+1)/10$ så får vi $2^{-i} < 10^{-N-1}$, dvs vi har $N$ fixerade decimaler (vi vinner ungefär 3 decimaler per 10 steg).

Vi har alltså två decimalutvecklingar. Vi betecknar dem med

$$\bar{x} = \lim_{i\to\infty} x_i = 1.4142135...,$$
$$\bar{X} = \lim_{i\to\infty} X_i = 1.41421356...$$

Likheten (4.1) innebär att decimalutvecklingarna är lika. Med $2^{-i} < 10^{-N-1}$ överensstämmer $x_i$ och $X_i$ till $N$ decimaler. Alltså:

$$\bar{x} = \bar{X} = 1.41421356...$$

Vi skriver

$$\sqrt{2} = \bar{x} = \bar{X} = 1.41421356...$$

Vilket innebär att vi satt ett namn, $\sqrt{2}$, på den decimalutveckling som vi konstruerat.

## Svar på Fråga 2

Vi vill visa att $f(\sqrt{2}) = 0$ dvs $(\sqrt{2})^2 = (1.41421356...)^2 = 2$. Vi visar att detta gäller i form av gränsvärdet

$$\lim_{i\to\infty} f(x_i) = 0, \tag{4.4}$$

dvs $\forall \epsilon > 0$ $\exists N$ sådant att

$$i \geq N \quad \Rightarrow \quad |f(x_i) - 0| < \epsilon. \tag{4.5}$$

Tag då ett $\epsilon > 0$ och försök bestämma $N$ sådant att (4.5) gäller. Kom ihåg att $f(x) = x^2 - 2$ är Lipschitz-kontinuerlig på intervallet $[1, 2]$ med Lipschitz-konstanten $L = 4$. Med Lipschitz-villkoret och (4.1) får vi

$$|f(x_i) - 0| = |\underbrace{f(x_i)}_{<0}| = -f(x_i) < \underbrace{f(X_i)}_{>0} - f(x_i) = |f(X_i) - f(x_i)| \leq 4|X_i - x_i| = 4 \cdot 2^{-i} < \epsilon,$$

om $i > \ln(\epsilon/4)/\ln(1/2) = \ln(4/\epsilon)/\ln(2)$. Vi kan ta $N = \lceil \ln(4/\epsilon)/\ln(2) \rceil$ (heltalstaket=avrunda uppåt). Detta visar (4.5).

## Svar på Fråga 3

Finns det fler lösningar? Vi noterar först att funktionen $f(x) = x^2 - 2$ är strängt växande för $x \geq 0$, dvs

$$0 \leq x < y \quad \Rightarrow \quad f(x) < f(y).$$

(Bevis av detta: Antag först $0 < x < y$. Multiplikation med $x$ och $y$ ger $x^2 < xy$ och $xy < y^2$, så att $x^2 < xy < y^2$ och därmed $x^2 - 2 < y^2 - 2$. Om $0 = x < y$ så är $-2 = f(x) < f(y) = y^2 - 2$.)

Antag nu att det finns två icke-negativa lösningar $x, y$ med $0 \leq x < y$. Men då är $0 = f(x) < f(y) = 0$ vilket är omöjligt. Alltså är $\sqrt{2} = 1.41421356...$ den enda icke-negativa lösningen.

## Sammanfattning

1. Vi har konstruerat en approximerande följd $x_i$.
2. Vi har visat att följden bildar en decimalutveckling $\bar{x} = \lim_{i\to\infty} x_i$.
3. Vi har visat att $\lim_{i\to\infty} f(x_i) = 0$, dvs $f(\bar{x}) = 0$.
4. Vi har visat att det finns bara en icke-negativ lösning, dvs alla konstruktioner ger samma resultat.

## 4.2 Reellt tal=decimalutveckling=Cauchy-följd

De reella talen är mängden av alla decimalutvecklingar (ändliga, periodiska eller icke-periodiska). Vi har sett att olikheterna (4.2) och (4.3),

$$|x_i - x_j| \leq 2^{-i}, \quad |X_i - X_j| \leq 2^{-i} \quad \text{för } j \geq i,$$

garanterar att följderna bildar decimalutvecklingar. Antag att vi har en decimalutveckling,

$$a = p.q_1q_2q_3q_4...,$$

där $p$ är heltalsdelen. Om vi trunkerar efter $n$ decimaler,

$$a_n = p.q_1q_2q_3q_4...q_n,$$

så får vi en följd $\{a_n\}_{n=1}^{\infty}$, sådan att

$$|a_i - a_j| < 10^{-i} \quad \text{för } j \geq i.$$

Det beror på att $a_j - a_i$ ges av

$$\begin{array}{rlll} p & .q_1 \dots q_i & q_{i+1} \dots q_j \\ -p & .q_1 \dots q_i & \\ \hline 0 & .0 \dots 0 & q_{i+1} \dots q_j \end{array}$$

dvs $|a_i - a_j| = |0.q_{i+1} \dots q_j| \cdot 10^{-i} < 10^{-i}$.

Å andra sidan, om en följd uppfyller

$$|a_i - a_j| < 10^{-n-1} \quad \text{för } j \geq i,$$

så stämmer de $n$ första decimalerna i $a_i$ överens med de $n$ första decimalerna i $a_j$ för alla $j \geq i$. Dvs en decimalutveckling växer fram. Vi skriver då

$$a = \lim_{i \to \infty} a_i,$$

där $a$ är decimalutvecklingen = reella talet. En sådan följd kallas Cauchy-följd.

**Definition 5.** *(Cauchy-följd) En talföljd $\{a_j\}_{j=1}^{\infty}$ kallas Cauchy-följd om $\forall \epsilon > 0$ $\exists N$ sådant att*

$$i, j \geq N \quad \Rightarrow \quad |a_i - a_j| < \epsilon.$$

Som vi sett betyder detta att följden genererar en decimalutveckling. Toleransen $\epsilon$ anger antalet decimaler: $\epsilon = 10^{-n-1}$ betyder $n$ korrekta decimaler.

**Exempel 16.** $a_n = 1/n$. Antag $j \geq i$. Då får vi

$$|a_i - a_j| = \left|\frac{1}{i} - \frac{1}{j}\right| = \frac{1}{i} - \frac{1}{j} < \frac{1}{i} \leq \text{(vi vill)} \leq \epsilon.$$

Vi löser ut $i$. Vi får $i \geq 1/\epsilon$. Vi tar $N = \lceil 1/\epsilon \rceil$. Då gäller

$$j \geq i \geq N \quad \Rightarrow \quad |a_i - a_j| < 1/i \leq 1/N \leq \epsilon.$$

□

### Övningar

Visa att följande är Cauchy-följder.

1. $a_n = 1/n^2$
2. $a_n = 1/\sqrt{n}$
3. Visa: $a_n$ och $b_n$ är Cauchy medför att $a_n + b_n$ är Cauchy.
4. Visa: $f$ är Lipschitz på $[a, b]$ och $a_n \in [a, b]$ är Cauchy medför att $f(a_n)$ är Cauchy.

## 4.3 Bisektionsalgoritmen

Vi formulerar nu bisektionsalgoritmen i sin allmänna form, dvs för en allmän ekvation av formen $f(x) = 0$. Algoritmen är:

(a) Givet: en funktion $f$ som är kontinuerlig på $[a, b]$ och med $f(a)f(b) < 0$ (olika tecken).

(b) Sätt $x_0 = a$, $X_0 = b$, $i = 0$.

(c) Bilda mittpunkten $\hat{x}_i = (x_i + X_i)/2$.
Om $f(\hat{x}_i) = 0$, sätt $x_{i+1} = X_{i+1} = \hat{x}_i$ och stoppa.
Om $f(\hat{x}_i)f(x_i) < 0$, sätt $x_{i+1} = x_i,\ X_{i+1} = \hat{x}_i$.
Om $f(\hat{x}_i)f(X_i) < 0$, sätt $x_{i+1} = \hat{x}_i,\ X_{i+1} = X_i$.
Sätt $i = i + 1$ och upprepa (c).

Algoritmen genererar två följder $\{x_i\}_{i=0}^{\infty}$ och $\{X_i\}_{i=0}^{\infty}$. (Om algoritmen stoppar i steg nummer $i$ tänker vi oss att vi fortsätter följderna som konstanta följder: $x_j = X_j = \hat{x}_i$, $j \geq i + 1$.)

I praktiken räcker det att spara en av följderna. Och i praktiken avbryter vi algoritmen när tillräcklig noggrannhet uppnåtts: $|x_i - X_i| \leq \text{TOL}$, där TOL är en given feltolerans.

Med hjälp av denna algoritm kan vi bevisa Bolzanos sats och Satsen om mellanliggande värden.

## 4.4 Bolzanos sats

**Sats 6.** *(Bolzano's theorem) Assume that $f : [a, b] \to \mathbf{R}$ is continuous and that $f(a)f(b) < 0$ (opposite signs). Then there is a real number $\bar{x} \in [a, b]$ such that $f(\bar{x}) = 0$. If $f$ is strictly monotone (increasing or decreasing), then $\bar{x}$ is unique (the only such number).*

The proof is a *constructive proof*, in contrast to, for example, a proof by contradiction. This means that the proof describes how the number $\bar{x}$ is constructed. It is organized in the following four steps:

1. an algorithm which produces an approximating sequence $\{x_i\}$;
2. a proof that $\{x_i\}$ is a Cauchy sequence so that we get a real number (decimal expansion) $\bar{x} = \lim_{i\to\infty} x_i$.
3. a proof that $\bar{x}$ solves the equation: $\{f(x_i)\}$ is also a Cauchy sequence and $f(\bar{x}) = \lim_{i\to\infty} f(x_i) = 0$;
4. a proof that $\bar{x}$ is unique (in the case of strictly monotone function).

Remember these steps, we will use the same kind of proof many times. The first three steps give the *existence* of a solution. The last step gives *uniqueness* of the solution of the equation $f(x) = 0$. One important consequence of uniqueness is that all approximating sequences will converge to the same solution $\bar{x}$. That is, the solution is independent of the choice of algorithm or approximating sequence (independent of the construction).

*Bevis.* We write the proof under the extra assumption that $f$ is Lipschitz continuous.

Step 1. We use the bisection algorithm with starting points $a$ and $b$.

Step 2. We obtain two sequences $\{x_i\}_{i=0}^{\infty}$ and $\{X_i\}_{i=0}^{\infty}$ with

$$|x_i - X_i| \leq (b-a)2^{-i}, \tag{4.6}$$

$$|x_i - x_j| \leq (b-a)2^{-i}, \quad j > i, \tag{4.7}$$

$$|X_i - X_j| \leq (b-a)2^{-i}, \quad j > i. \tag{4.8}$$

(We have equality in (4.6) if the algorithm does not stop.) The inequalities (4.7) and (4.8) mean that $x_i$ and $X_i$ are Cauchy sequences: $|x_i - x_j| \to 0$, $|X_i - X_j| \to 0$ as $i, j \to \infty$, and we get decimal expansions (real numbers)

$$\bar{x} = \lim_{i\to\infty} x_i, \quad \bar{X} = \lim_{i\to\infty} X_i.$$

The inequality (4.6) means that the decimal expansions are the same:

$$\bar{x} = \lim_{i\to\infty} x_i = \lim_{i\to\infty} X_i = \bar{X}.$$

Step 3. The Lipschitz continuity of $f$ gives

$$|f(x_i) - f(x_j)| \leq L|x_i - x_j| \leq L(b-a)2^{-i}, \quad j > i,$$

which means that $f(x_i)$ is a Cauchy sequence, $|f(x_i) - f(x_j)| \to 0$ as $i, j \to \infty$. It gives a decimal expansion (real number) which we denote $f(\bar{x})$:

$$f(\bar{x}) = \lim_{i\to\infty} f(x_i).$$

We must show that this is equal to 0. To do this we note that the distance between $f(x_i)$ and 0 is less than or equal to the distance between $f(x_i)$ and $f(X_i)$. This is because $f(x_i)$ and $f(X_i)$ have opposite signs. For example, if $f(x_i) < 0$ and $f(X_i) > 0$:

$$|f(x_i) - 0| = -f(x_i) < -f(x_i) + f(X_i) = |f(x_i) - f(X_i)|.$$

In any case:

$$|f(x_i) - 0| \leq |f(x_i) - f(X_i)| \leq L|x_i - X_i| \leq L(b-a)2^{-i} \to 0,$$

where we also used the Lipschitz condition and (4.6). This means that

$$\lim_{i\to\infty} f(x_i) = 0,$$

or in other words $f(\bar{x}) = 0$.

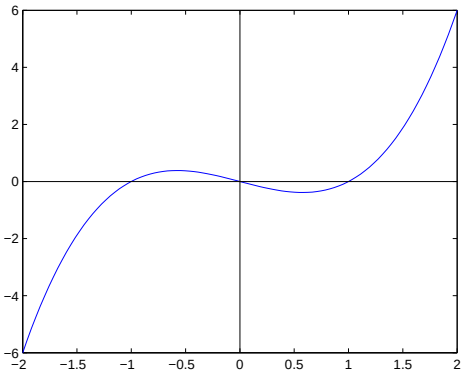
Step 4. Assume now that $f$ is strictly increasing. (The case of decreasing function can be handled in a similar way.) This means that $x < y$ implies $f(x) < f(y)$. Assume also that we have two different solutions, i.e., $\bar{x}_1 < \bar{x}_2$ with $f(\bar{x}_1) = f(\bar{x}_2) = 0$. But this is a contradiction to the strict monotonicity. Therefore, there is only one solution. □

We now consider equations of the form $f(x) = y$.

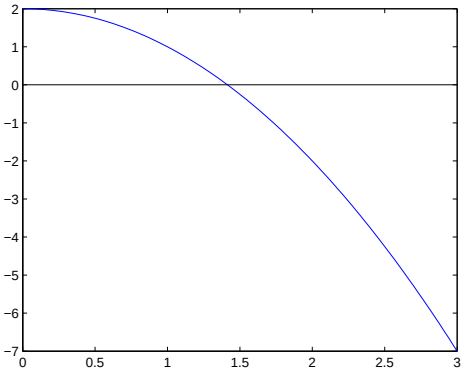
**Sats 7.** *(The Intermediate-Value Theorem) (Adams 1.4, Theorem 9) If $f : [a, b] \to \mathbf{R}$ is continuous and the real number $y$ is between $f(a)$ and $f(b)$, then there is a real number $\bar{x} \in [a, b]$ such that $f(\bar{x}) = y$. If $f$ is strictly monotone, then $\bar{x}$ is unique.*

The theorem says that a continuous function takes all values between its end-point values. A discontinuous function may skip some value.

*Bevis.* If $f(a) = f(b)$ then $y = f(a) = f(b)$ and we can choose $\bar{x} = a$ or $\bar{x} = b$. Otherwise we apply Bolzano's theorem to the function $F(x) = f(x) - y$. Clearly, $F : [a, b] \to \mathbf{R}$ is Lipschitz with the same constant as $f$, and $F(a)F(b) < 0$ because $y$ is between $f(a)$ and $f(b)$. Also, $F$ is strictly monotone if $f$ is strictly monotone. The conclusion now follows from Bolzano's theorem: there is $\bar{x}$ such that $F(\bar{x}) = f(\bar{x}) - y = 0$, and it is unique if $f$ is strictly monotone. □




Figur 4.2: Non-monotone function with three roots.




Figur 4.3: Monotone function with unique root.

## 4.5 Inverse function

Assume now that the function in the Intermediate-Value Theorem is strictly increasing. (Decreasing functions can be discussed in the same way.) Let us write $A = f(a)$, $B = f(b)$. We note the following consequences of the theorem.

The function takes all values $y$ in the interval $[A, B]$ and it takes no values outside $[A, B]$. This means that the range of the function is exactly $R(f) = [A, B]$.

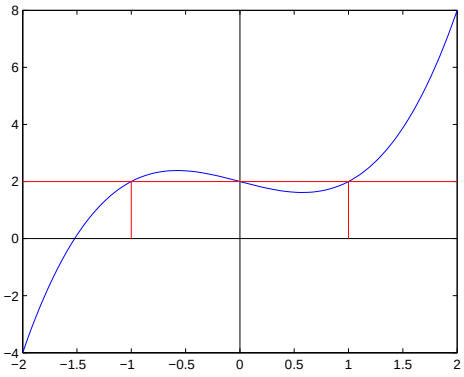
The equation $f(x) = y$ has a unique solution $x$ for any $y \in [A, B]$. Since $x$ is unique, this defines a function $y \mapsto x$. More precisely, we can define a function

$$\begin{aligned} &g : [A, B] \to \mathbf{R} \\ &x = g(y), \quad \text{where } x \text{ is the unique solution of } f(x) = y. \end{aligned}$$

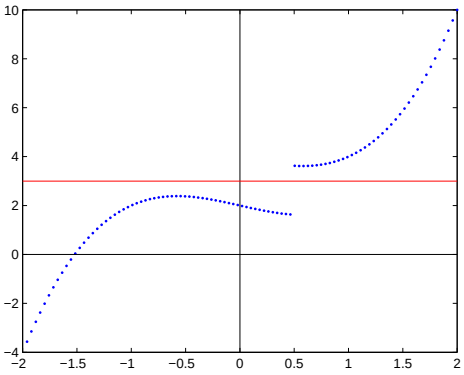
(Remember that a function must have a *unique* value *for each* element of the domain of definition.) We then say that $f$ is invertible ("inverterbar") and the function $g$ is called the the inverse of $f$. It is denoted $g = f^{-1}$,

$$\begin{aligned} &f^{-1} : [A, B] \to \mathbf{R} \\ &x = f^{-1}(y), \quad \text{where } x \text{ is the unique solution of } f(x) = y. \end{aligned}$$

Of course, we may interchange the roles of $x$ and $y$ and write $y = f^{-1}(x)$, $x \in [A, B]$.




Figur 4.4: Equation $f(x) = y$ with multiple solutions.




Figur 4.5: The discontinuous function skips the value $y = 3$.

Note that

$$D(f^{-1}) = R(f) = [A, B], \qquad R(f^{-1}) = D(f) = [a, b],$$
$$f(f^{-1}(y)) = y, \ \forall y \in [A, B]; \qquad f^{-1}(f(x)) = x, \ \forall x \in [a, b].$$

The last two identities are called the cancellation property of the inverse function. They are proved like this:

$$y = f(x) = f(f^{-1}(y)), \quad \forall y \in [A, B],$$
$$x = f^{-1}(y) = f^{-1}(f(x)), \quad \forall x \in [a, b].$$

A strictly decreasing function can be discussed in the same way if we note that $R(f) = [B, A]$. We have now proved the following.

**Sats.** *If the function $f : [a, b] \to \mathbf{R}$ is continuous and strictly monotone, then it has an inverse function $f^{-1}$.*

Warning: $f^{-1}$ is pronounced "f inverse", and it is not the same as "f to the minus one" $(f)^{-1} = 1/f$. It is confusing that "inverse" and "raised to $-1$" are written in the same way, but it is a tradition in mathematics, and we have to live with it and be careful when we use it.

**Exempel.** The function $f(x) = x^2$, is not invertible because the equation $x^2 = y$ has two solutions $x = \pm\sqrt{y}$. In the next section we shall see that $f$ becomes invertible if we restrict its domain of definition to the nonnegative numbers.

## 4.6 The square root function

We introduce the square root function and investigate its properties.

The function $f : [0, b] \to \mathbf{R}$, $f(x) = x^2$, is strictly increasing because $0 \le x < y$ implies $x - y < 0$ and $x + y > 0$ so that $x^2 - y^2 = (x - y)(x + y) < 0$. Hence the equation $x^2 = y$ has a unique solution $x = \sqrt{y}$ for any $y \in [0, b^2]$. Therefore, $f$ is invertible with $f^{-1}(y) = \sqrt{y}$.

This can be done for any $b$, the uniqueness implies that the bisection algorithm gives the same result no matter what starting points $a = 0$ and $b > 0$ we use. So the function $f : [0, \infty) \to \mathbf{R}$, $f(x) = x^2$, is invertible with inverse $f^{-1} : [0, \infty) \to \mathbf{R}$, $f^{-1}(x) = \sqrt{x}$. Note: the domain of the square root $D(\sqrt{\ }) = [0, \infty)$ and the range $R(\sqrt{\ }) = [0, \infty)$.

Warning: $f^{-1}(x) = \sqrt{x}$ is not the same as $(f)^{-1}(x) = (f(x))^{-1} = x^{-2} = 1/x^2$ although they are written in almost the same way.

Since $x^2 = a$ and $y^2 = b$ implies $(xy)^2 = ab$ and $(x/y)^2 = a/b$, we conclude

$$\sqrt{ab} = \sqrt{a}\sqrt{b}, \quad \sqrt{\frac{a}{b}} = \frac{\sqrt{a}}{\sqrt{b}}.$$

We now investigate the Lipschitz continuity of $\sqrt{x}$:

$$\begin{aligned}
|\sqrt{x} - \sqrt{y}| &= \Big\{\text{multiply by the "conjugate expression" } \sqrt{x} + \sqrt{y}\Big\} \\
&= \Big|\frac{(\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y})}{\sqrt{x} + \sqrt{y}}\Big| = \Big|\frac{x - y}{\sqrt{x} + \sqrt{y}}\Big| \quad \Big\{\text{by the conjugate rule}\Big\} \\
&= \frac{1}{\sqrt{x} + \sqrt{y}}\,|x - y| \le \frac{1}{\sqrt{\delta} + \sqrt{\delta}}\,|x - y| = \frac{1}{2\sqrt{\delta}}\,|x - y| \quad \text{for } x, y \ge \delta.
\end{aligned}$$

We conclude that the square root function is Lipschitz continuous with constant $L = 1/(2\sqrt{\delta})$ on any interval of the form $[\delta, \infty)$ with $\delta > 0$. That is, on any interval that stays away from 0. But it is not Lipschitz on its whole domain of definition $\mathbf{R}_+ = [0, \infty)$. This is clear from the previous calculation, but is also seen in the graph where the slope is infinite at 0.

# Kapitel 5

# Fixpunktsiteration

## 5.1 Fixed point equation

An algebraic equation can be written in two equivalent ways (meaning that they have the same solutions).

**1.** $f(x) = 0$**.**

A solution $\bar{x}$ is called a *root* of $f$ or a *zero* of $f$ ("nollställe till $f$").

**Exempel.** The function $f(x) = x^2 - 2$ has two roots $\bar{x}_1 = \sqrt{2}$, $\bar{x}_2 = -\sqrt{2}$.

Algorithm: For equations of this form we have the bisection algorithm.

**2.** $x = g(x)$**.**

A solution $\bar{x}$ is called a *fixed point* ("fixpunkt") of $g$. The equation is called a fixed point equation.

**Exempel.** The function $g(x) = 2/x$ has two fixed points, $\bar{x}_1 = \sqrt{2}$, $\bar{x}_2 = -\sqrt{2}$, because $\pm\sqrt{2} = \pm\frac{\sqrt{2}\sqrt{2}}{\sqrt{2}} = \frac{2}{\pm\sqrt{2}}$.

Algorithm: a natural algorithm for a fixed point equation is to to choose a starting point $x_0$ and then compute $x_i$ according to $x_i = g(x_{i-1})$. This is called the fixed point iteration. We hope that the sequence $x_i$ converges to a fixed point. This works sometimes and sometimes not.

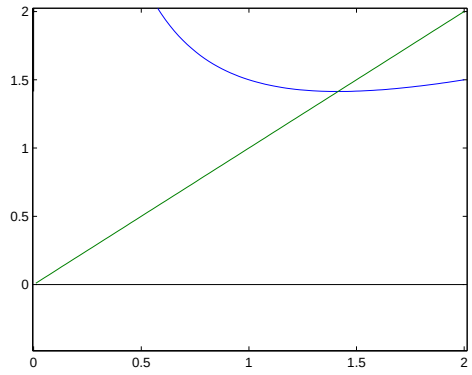
**Exempel.** With $g(x) = x/2 + 1/x$ and $x_0 = 1$ we get $x_1 = g(x_0) = 3/2$, $x_2 = 15/12$ and so on. This is easy to try in MATLAB:

```
>> format long
>> x=1
>> x=x/2+1/x
>> x=x/2+1/x
>> x=x/2+1/x
>> x=x/2+1/x
```

Do this now!! Does it converge? Do you recognize a decimal expansion? Se Figure 5.1.

**Exempel.** With $g(x) = 2/x$ and $x_0 = 1$ we get $x_1 = g(x_0) = 2$, $x_2 = 1$, $x_3 = 2$, i.e., we get the sequence $\{1, 2, 1, 2, \dots\}$ which is divergent.

Note that the equations $x^2 - 2 = 0$, $x = x/2 + 1/x$, and $x = 2/x$ are equivalent (multiply the last two equations by $x$ to see this). We can rewrite equations between the two forms in many ways. For example, $x = g(x)$ can be written as $x - g(x) = 0$. On the other hand, $f(x) = 0$ can




Figur 5.1: The functions $y = x/2 + 1/x$ and $y = x$.

be written $x = x + \alpha f(x)$ where $\alpha$ or $\alpha(x)$ is any nonzero number or function. The challenge is to find a "good" choice of $\alpha$ so that the fixed point iteration is convergent. Later, when we discuss Newton's method, we shall find an optimal choice of $\alpha$. In fact, $x = x/2 + 1/x$ is obtained from $x^2 - 2 = 0$ by using $\alpha(x) = -1/(2x)$ and we will learn why this is a good choice.

So when does the fixed point iteration work? It turns out that one important condition is that the Lipschitz constant $L = L_g$ of $g$ is strictly smaller than 1. More precisely, we shall assume that $g$ is Lipschitz on an interval $I$ with constant $L < 1$, i.e.,

$$|g(x) - g(y)| \le L|x - y| \quad \text{for } x, y \in I \text{ and with } L < 1.$$

Such a function is called a *contraction mapping* (or just contraction). Note that the distance between the images $g(x)$ and $g(y)$ is smaller than the distance between $x$ and $y$.

**Exempel.** For $g(x) = 2/x$ we have

$$|g(x) - g(y)| = 2\Big|\frac{y - x}{xy}\Big| = \frac{2}{|x|\,|y|}\,|x - y| \le \frac{1}{2}|x - y|, \quad \forall x, y \ge 2,$$

so that $g$ is a contraction with $L = 1/2$ on $I = [2, \infty)$. Note that we used that $z = xy \in [4, \infty)$ so that $\frac{2}{|x|\,|y|} = \frac{2}{z} \in [0, \frac{1}{2}]$.

**Exempel.** For $g(x) = x/2 + 1/x$ we have

$$|g(x) - g(y)| = \Big|\frac{x - y}{2} + \frac{y - x}{xy}\Big| = \Big|\frac{1}{2} - \frac{1}{xy}\Big|\,|x - y|.$$

Now let us consider $x, y \in [1, 2]$ for example. Write $z = xy$. Then $z \in [1, 4]$ and we get $\frac{1}{2} - \frac{1}{xy} = \frac{1}{2} - \frac{1}{z} \in [-\frac{1}{2}, \frac{1}{4}]$ with absolute value $|\frac{1}{2} - \frac{1}{xy}| \le \frac{1}{2}$. Therefore

$$|g(x) - g(y)| = \Big|\frac{1}{2} - \frac{1}{xy}\Big|\,|x - y| \le \frac{1}{2}\,|x - y|, \quad \forall x, y \in [1, 2],$$

so that $g$ is a contraction with $L = 1/2$ on $I = [1, 2]$.

## 5.2 The contraction mapping theorem

Remember that the bisection algorithm leads to Bolzano's theorem. The fixed point iteration also leads to a theorem.

Remember that a closed interval $I$ is an interval that contains its endpoints (if it has any). A closed interval can be of the following types:

$$I = [a, b] \quad \text{(closed and bounded interval)}$$
$$I = [a, \infty),\ I = (-\infty, b],\ I = (\infty, \infty) = \mathbf{R}, \quad \text{(closed and unbounded intervals)}$$

**Sats.** *(The Contraction Mapping Theorem) Assume that $I$ is a closed interval and that $g : I \to I$ is a contraction mapping. Then $g$ has a unique fixed point $\bar{x} \in I$. The fixed point is obtained as the limit of the fixed point iteration, $x_i = g(x_{i-1})$, for any starting point $x_0 \in I$.*

It is important that the target set $I$ is the same as the domain of definition, $g : I \to I$; it guarantees that the sequence does not jump out of the interval $I$ where $g$ is a contraction. It is also important that $I$ is closed; it guarantees that $\bar{x} = \lim x_i \in I$.

*Bevis.* The proof follows the four steps of a constructive proof that we mentioned before.

Step 1. An algorithm: we use the fixed point iteration. Take an arbitrary point $x_0 \in I$ and compute $x_i = g(x_{i-1})$.

Step 2. A proof that $\{x_i\}$ is a Cauchy sequence. We must estimate $|x_i - x_j|$ for $j > i$. Consider first the distance between two consecutive elements of the sequence:

$$|x_{k+1} - x_k| = |g(x_k) - g(x_{k-1})| \le L|x_k - x_{k-1}|.$$

Here we used the fact that the $x_k$ stay in $I$ and $g$ is a contraction on $I$. Therefore

$$|x_{k+1} - x_k| \le L|x_k - x_{k-1}|.$$

Since $L < 1$ this means that $x_{k+1}, x_k$ are closer to each other than $x_k, x_{k-1}$. In the same way:

$$|x_k - x_{k-1}| \le L|x_{k-1} - x_{k-2}|.$$

By repeating this we get

$$\begin{aligned} |x_{k+1} - x_k| &\le L|x_k - x_{k-1}| \\ &\le L^2|x_{k-1} - x_{k-2}| \\ &\le L^3|x_{k-2} - x_{k-3}| \\ &\le \cdots \le L^k|x_1 - x_0|, \end{aligned}$$

that is

$$|x_{k+1} - x_k| \le L^k|x_1 - x_0|. \tag{5.1}$$

Now consider $|x_i - x_j|$ for $j > i$. We have

$$\begin{aligned} x_i - x_j &= x_i - x_{i+1} + x_{i+1} - x_{i+2} + x_{i+2} - \cdots - x_{j-2} + x_{j-2} - x_{j-1} + x_{j-1} - x_j \\ &= \sum_{k=i}^{j-1} (x_k - x_{k+1}). \end{aligned}$$

Such a sum is called a telescope sum because all terms cancel except the first and the last. Applying the triangle inequality to the sum and using (5.1) we have

$$|x_i - x_j| \le |x_i - x_{i+1}| + \cdots + |x_{j-1} - x_j| = \sum_{k=i}^{j-1} |x_k - x_{k+1}| \le |x_1 - x_0| \sum_{k=i}^{j-1} L^k.$$

This is a geometric sum given by the well-known formula:

$$\sum_{k=i}^{j-1} L^k = L^i(1 + L + \cdots + L^{j-i-1}) = L^i \frac{1 - L^{j-i}}{1 - L}.$$

Therefore

$$|x_i - x_j| \leq |x_1 - x_0| L^i \frac{1 - L^{j-i}}{1 - L} \leq |x_1 - x_0| L^i \frac{1}{1 - L}, \tag{5.2}$$

because $0 \leq 1 - L^{j-i} \leq 1$ for $j > i$. Since $L < 1$ we have $L^i \to 0$ and hence $|x_i - x_j| \to 0$ as $i \to \infty$ with $j > i$. Thus $x_i$ is a Cauchy sequence and we get a decimal expansion (real number)

$$\bar{x} = \lim_{i \to \infty} x_i,$$

which belongs to $I$ because $I$ is closed.

Step 3. Proof that $\bar{x}$ is a fixed point. We have

$$|g(x_i) - g(x_j)| \leq L|x_i - x_j| \to 0, \quad i, j \to \infty,$$

so that $g(x_i)$ is a Cauchy sequence. We get a real number which we denote $g(\bar{x})$:

$$g(\bar{x}) = \lim_{i \to \infty} g(x_i).$$

We must show that $\lim_{i \to \infty} g(x_i) = \bar{x}$ so that $g(\bar{x}) = \bar{x}$. But

$$|\bar{x} - g(x_i)| = |\bar{x} - x_{i+1}| \to 0, \quad i \to \infty.$$

This means that $\lim_{i \to \infty} g(x_i) = \bar{x}$ and hence $g(\bar{x}) = \bar{x}$.

Step 4. Uniqueness. Assume that we have two fixed points $\bar{x}_1, \bar{x}_2 \in I$. Then

$$|\bar{x}_1 - \bar{x}_2| = |g(\bar{x}_1) - g(\bar{x}_2)| \leq L|\bar{x}_1 - \bar{x}_2|.$$

which implies

$$(1 - L)|\bar{x}_1 - \bar{x}_2| \leq 0.$$

But $1 - L > 0$ so the only possibility is $|\bar{x}_1 - \bar{x}_2| = 0$. In other words: $\bar{x}_1 = \bar{x}_2$. So there is only one fixed point in $I$.

Note that the uniqueness of $\bar{x}$ implies that we get the same limit no matter which starting point $x_0$ we choose. □

**Exempel.** We have seen that $g(x) = 2/x$ is a contraction on the closed interval $I = [2, \infty)$. But $x_0 = 3$ gives $x_1 = 2/3 \notin I$ so the sequence jumps out. The sequence jumps back and forth between 3 and 2/3; it does not converge.

**Exempel.** We have seen that $g(x) = x/2 + 1/x$ is a contraction with $L = 1/2$ on $I = [1, 2]$. We check that $g : I \to I$. If $x \in I = [1, 2]$, i.e., $1 \leq x \leq 2$, then $x/2 \leq 1$ and $1/x \leq 1$ so that $x/2 + 1/x \leq 1 + 1 = 2$. Also $x/2 \geq 1/2$ and $1/x \geq 1/2$ so that $x/2 + 1/x \geq 1$. Therefore $g(x) \in I = [1, 2]$. The contraction mapping theorem says that $g$ has a unique fixed point in $I = [1, 2]$. What is it?

## 5.3 When do we stop the iteration?

We stop the iteration when the distance between two consecutive iterates is less that a given tolerance, $|x_i - x_{i+1}| \leq \text{TOL}$. Then we expect that a certain number of decimals have been fixed in the decimal expansion $\bar{x}$. For example, with $|x_i - x_{i+1}| \leq 10^{-N-1}$ we expect approximately $N$ decimals to be fixed.

This is justified by the calculation (with $j > i$ as usual)

$$|x_i - x_j| \le \sum_{k=i}^{j-1} |x_k - x_{k+1}| \le |x_i - x_{i+1}| \sum_{k=i}^{j-1} L^{k-i}$$
$$\le \frac{1 - L^{j-i}}{1 - L} |x_i - x_{i+1}| \le \frac{1}{1 - L} |x_i - x_{i+1}|,$$

which is done in the same way as (5.2) but using $|x_{k+1} - x_k| \le L^{k-i}|x_i - x_{i+1}|$ instead of (5.1). Therefore

$$|x_i - x_j| \le \frac{1}{1 - L} |x_i - x_{i+1}| \le \frac{1}{1 - L} \text{ TOL}, \quad j > i.$$

or by letting $j \to \infty$

$$|x_i - \bar{x}| \le \frac{1}{1 - L} |x_i - x_{i+1}| \le \frac{1}{1 - L} \text{ TOL}\,.$$

So the number of fixed decimals after $i$ steps is determined by $\frac{1}{1-L}$ TOL. Note that this number is bigger than TOL, much bigger if $L$ is near 1, so we get fewer decimals than TOL itself indicates. The reason for this is that we are looking at the *residual* $x_i - x_{i+1} = x_i - g(x_i)$ which measures how well $x_i$ satifies the equation $x - g(x) = 0$. The size of the residual is then magnified by the factor $\frac{1}{1-L}$ when we use it to estimate the error in $x_i$.

## 5.4 How fast is the convergence?

From the construction of the sequence $x_i$ we have

$$|x_i - \bar{x}| = |g(x_{i-1}) - g(\bar{x})| \le L|x_{i-1} - \bar{x}|.$$

Therefore the error is reduced by the factor $L < 1$ in each step. The smaller $L$ is, the faster the convergence. This is called *linear convergence*. The bisection algorithm also converges linearly: the error is reduced by a factor 1/2 in each step.

**Exempel.** For $g(x) = x - x^2/2 + 1$ on $I = [1, 3/2]$ we have $L = 1/2$. This follows from

$$g(x) - g(y) = (1 - \tfrac{1}{2}(x + y))(x - y).$$

Here $2 \le x + y \le 3$ so that $-\frac{1}{2} \le 1 - \frac{1}{2}(x + y) \le 0$ with absolute value $|1 - \frac{1}{2}(x + y)| \le 1/2$.

If we compute a few iterations in MATLAB with $x_0 = 1$ and for each iteration compute $|x_i - \sqrt{2}|$ we see that that the error is reduced approximately by a factor 1/2 in each step. This is what I got. The first column is $x_i$ and the second is $|x_i - \sqrt{2}|$.

```
1.00000000000000   0.41421356237310
1.50000000000000   0.08578643762690
1.37500000000000   0.03921356237310
1.42968750000000   0.01547393762690
1.40768432617188   0.00652923620122
1.41689674509689   0.00268318272380
1.41309855196381   0.00111501040929
1.41467479318270   0.00046123080961
1.41402240794944   0.00019115442365
1.41429272285787   0.00007916048478
```

Note that $\sqrt{2}$ is computed by the MATLAB function `sqrt(2)` which is also an approximation but with approximately 16 correct decimals. So we can use it to test the accuracy of our computation.

The convergence is sometimes much faster than this. As the example $g(x) = x/2 + 1/x$ shows. Compute and check this!! This is what I got: again the first column is $x_i$ and the second is $|x_i - \sqrt{2}|$.

```
1.00000000000000   0.41421356237310
1.50000000000000   0.08578643762690
1.41666666666667   0.00245310429357
1.41421568627451   0.00000212390141
1.41421356237469   0.00000000000159
1.41421356237309   0.00000000000000
1.41421356237309   0.00000000000000
1.41421356237309   0.00000000000000
1.41421356237309   0.00000000000000
1.41421356237309   0.00000000000000
```

In fact it is possible to show that

$$|x_i - \bar{x}| \leq K|x_i - \bar{x}|^2,$$

where $K$ is some number. This is called *quadratic convergence* and means that the error is reduced by a factor which is proportional to the error itself. So the speed of converges increases as $x_i$ approaches $\bar{x}$. We will learn later why it is so fast. See Chapter Newton's method.

## 5.5 Advantages and disadvantages

A disadvantage with the fixed point iteration is that it is often difficult to find a suitable interval $I$ where the iteration converges. This is usually very easy for the bisection algorithm: just plot the function and pick two points where the function has opposite signs.

There are two advantages:

1. the fixed point iteration can be very fast. The bisection algorithm always converges linearly but not faster than that.

2. the fixed point iteration also works for systems of equations. This is not true for bisection.

We shall return to these advantages later.

# Kapitel 6

# Newtons metod

## 6.1 Numerisk beräkning av derivata

Låt $f : I \to \mathbf{R}$ vara en deriverbar funktion ($I$ är ett intervall). Derivatans definition är

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{6.1}$$

Det betyder att vi kan approximera derivatan med en differenskvot:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \quad \text{med } h \approx 0. \tag{6.2}$$

Vi ska nu diskutera felet i denna approximation. Vi måste då ta hänsyn till att datorn beräknar med ändlig precision, dvs räknar med ändligt många siffror och att detta leder till avrundningsfel. Datorn beräknar alltså en approximation

$$\tilde{f}(x) \approx f(x) \tag{6.3}$$

där avrundningsfelet är

$$e_f(x) = \tilde{f}(x) - f(x) \tag{6.4}$$

med begränsningen

$$|e_f(x)| = |\tilde{f}(x) - f(x)| \leq \delta_f, \quad x \in I. \tag{6.5}$$

I MATLAB, som räknar med cirka 16 siffror, och om $|f(x)| \approx 1$, kan vi antaga att $\delta_f \approx 10^{-15}$. Det vi beräknar är alltså

$$f'(x) \approx \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} \quad \text{med } h \approx 0. \tag{6.6}$$

Felet är

$$\begin{aligned} \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x) &= \frac{f(x+h) - f(x)}{h} - f'(x) + \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - \frac{f(x+h) - f(x)}{h} \\ &= \Big(\frac{f(x+h) - f(x)}{h} - f'(x)\Big) + \frac{e_f(x+h) - e_f(x)}{h}. \end{aligned}$$

Triangelolikheten ger

$$\Big|\frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x)\Big| \leq \Big|\frac{f(x+h) - f(x)}{h} - f'(x)\Big| + \Big|\frac{e_f(x+h) - e_f(x)}{h}\Big|. \tag{6.7}$$

Den första termen är diskretiseringsfelet och den andra är avrundningsfelet. För avrundningsfelet använder vi (6.5)

$$\Big|\frac{e_f(x+h)-e_f(x)}{h}\Big| \le \frac{|e_f(x+h)|+|e_f(x)|}{h} \le \frac{2\delta_f}{h}. \tag{6.8}$$

För diskretiseringsfelet använder vi en formel för linjäriseringsfelet (se Definition 8 och Theorem 9 i Adams 4.7)

$$f(x) = f(a) + f'(a)(x-a) + E(x) \quad \text{med felet } E(x) = \frac{1}{2}f''(s)(x-a)^2, \tag{6.9}$$

där $s$ är en (obekant) punkt mellan $x$ och $a$. Vi använder denna genom att byta ut $x$ mot $x+h$ och $a$ mot $x$:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(s)h^2 \quad \text{med } s \text{ mellan } x \text{ och } x+h. \tag{6.10}$$

Genom att dividera med $h$ får vi

$$\Big|\frac{f(x+h)-f(x)}{h} - f'(x)\Big| = \frac{1}{2}|f''(s)|h. \tag{6.11}$$

Eftersom $s$ är okänd måste vi antaga att vi har en begränsning för $f''$,

$$|f''(x)| \le K_f, \quad x \in I. \tag{6.12}$$

Vi får då:

$$\Big|\frac{f(x+h)-f(x)}{h} - f'(x)\Big| = \frac{1}{2}|f''(s)|h \le \frac{1}{2}K_f h. \tag{6.13}$$

För totala felet i (6.7) får vi med (6.13) och (6.8):

$$\Big|\frac{\tilde{f}(x+h)-\tilde{f}(x)}{h} - f'(x)\Big| \le \frac{1}{2}K_f h + \frac{2\delta_f}{h}. \tag{6.14}$$

Detta gäller för alla $x \in I$ och $x+h \in I$. Vi vill välja $h$ så att felet blir minimalt. Vi ser att den första termen i $\frac{1}{2}K_f h + \frac{2\delta_f}{h}$ minskar då $h$ minskar medan den andra ökar. Vi får minimum då båda är lika:

$$\frac{1}{2}K_f h = \frac{2\delta_f}{h}, \tag{6.15}$$

dvs

$$h^2 = \frac{4\delta_f}{K_f}, \quad h = 2\sqrt{\frac{\delta_f}{K_f}}. \tag{6.16}$$

(Man kan också visa detta genom att derivera $\frac{1}{2}K_f h + \frac{2\delta_f}{h}$ med avseende på $h$ och sätta derivatan $= 0$.) Minimala felet blir då

$$\Big|\frac{\tilde{f}(x+h)-\tilde{f}(x)}{h} - f'(x)\Big| = \frac{1}{2}K_f h + \frac{2\delta_f}{h} = K_f h = 2\sqrt{K_f}\sqrt{\delta_f}. \tag{6.17}$$

I MATLAB med $\delta_f \approx 10^{-15}$ och $|f(x)| \approx 1$ och $K_f \approx 1$ (till exempel) får vi ungefär

$$h \approx 2\sqrt{\delta_f} \approx 2 \cdot 10^{-7.5} \approx 10^{-7}, \tag{6.18}$$

$$\Big|\frac{\tilde{f}(x+h)-\tilde{f}(x)}{h} - f'(x)\Big| \approx 2\sqrt{\delta_f} \approx 10^{-7}. \tag{6.19}$$

Dvs vi får ungefär 7 korrekta decimaler.

## En bättre approximation

Vi får en bättre approximation om vi ersätter den ensidiga differenskvoten i (6.2) med den symmetriska differenskvoten

$$f'(x) \approx \frac{f(x+h)-f(x-h)}{2h} \quad \text{med } h \approx 0. \tag{6.20}$$

Det totala felet blir nu istället för (6.7)

$$\Big|\frac{\tilde f(x+h)-\tilde f(x-h)}{2h} - f'(x)\Big| \le \Big|\frac{f(x+h)-f(x-h)}{2h} - f'(x)\Big| + \Big|\frac{e_f(x+h)-e_f(x-h)}{2h}\Big|. \tag{6.21}$$

Man kan visa att detta begränsas av

$$\Big|\frac{\tilde f(x+h)-\tilde f(x-h)}{h} - f'(x)\Big| \le \frac{1}{6}M_f h^2 + \frac{\delta_f}{h}, \tag{6.22}$$

där $\delta_f$ är som förut och $M_f$ är en begränsning av $f'''$:

$$|f'''(x)| \le M_f, \quad x \in I. \tag{6.23}$$

Vi gör en överslagsberäkning baserad på antagandena $\delta_f \approx 10^{-15}$, $M_f \approx 1$. Feluppskattningen i (6.22) blir approximativt minimum då båda termerna är approximativt lika:

$$h^2 \approx \frac{\delta_f}{h}, \quad h^3 \approx \delta_f, \quad h \approx \delta_f^{1/3} \approx 10^{-5}, \tag{6.24}$$

och då blir minimala felet ungefär

$$\Big|\frac{\tilde f(x+h)-\tilde f(x-h)}{2h} - f'(x)\Big| \approx h^2 \approx \delta_f^{2/3} \approx 10^{-10}. \tag{6.25}$$

Jämfört med (6.19) har vi cirka 3 decimaler noggrannare approximation av derivatan. Men man ska komma ihåg att (6.19) och (6.25) beror också på $K_f$ och $M_f$, vilket kan påverka jämförelsen om någon av dessa är stor. (Exakt värde i (6.24) är $h = (3\delta_f/M_f)^{1/3}$ och i (6.25) $cM_f^{1/3}\delta_f^{2/3}$ med $c \approx 1$.)

## Bevis av (6.22)

(Överkurs, kan skippas.) Avrundningsfelet uppskattas som förut i (6.8). För diskretiseringsfelet använder vi Taylors formel (se Theorem 10 i Adams 4.8)

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + E(x) \quad \text{med felet } E(x) = \frac{1}{6}f'''(s)(x-a)^3, \tag{6.26}$$

där $s$ är en (obekant) punkt mellan $x$ och $a$. Vi använder denna genom att byta ut $x$ mot $x+h$ och $a$ mot $x$:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{6}f'''(s_1)h^3, \tag{6.27}$$

$$f(x-h) = f(x) + f'(x)(-h) + \frac{1}{2}f''(x)(-h)^2 + \frac{1}{6}f'''(s_2)(-h)^3, \tag{6.28}$$

med $s_1, s_2$ mellan $x$ och $x+h$. Detta ger

$$\frac{f(x+h)-f(x-h)}{2h} = f'(x) + \frac{1}{12}f'''(s_1)h^2 + \frac{1}{12}f'''(s_2)h^2. \tag{6.29}$$

Diskretiseringsfelet blir

$$\Big|\frac{f(x+h)-f(x-h)}{2h} - f'(x)\Big| \le \frac{1}{12}|f'''(s_1)|h^2 + \frac{1}{12}|f'''(s_2)|h^2. \tag{6.30}$$

Eftersom $s_1, s_2$ är okända måste vi antaga att vi har en begränsning för $f'''$,

$$|f'''(x)| \leq M_f, \quad x \in I. \tag{6.31}$$

Vi får då:

$$\Big|\frac{f(x+h) - f(x-h)}{2h} - f'(x)\Big| \leq \frac{1}{6} M_f h^2. \tag{6.32}$$

Detta visar (6.22).

## 6.2 Newtons metod

(Adams 4.6) Låt $f : I \to \mathbf{R}$ vara en deriverbar funktion ($I$ är ett intervall). Vi erinrar oss att ekvationen $f(x) = 0$ kan skrivas som en fixpunktsekvation $x = g(x)$ genom omskrivningen

$$x = x - \alpha(x) f(x),$$

dvs med

$$g(x) = x - \alpha(x) f(x).$$

Det gäller att hitta $\alpha(x)$ så att $g : I \to I$ blir en kontraktion. Då konvergerar fixpunktsiterationen

$$x_{k+1} = g(x_k). \tag{6.33}$$

Detta är inte lätt att åstadkomma, men Newtons metod ger oss ett systematiskt sätt att göra detta.

Antag att vi har en approximativ lösning $x_k$ och vi vill hitta en bättre approximation $x_{k+1}$ som i (6.33). Vi bildar linjäriseringen av funktionen $f$ i $x_k$ (Definition 8 i Adams 4.7):

$$L(x) = f(x_k) + f'(x_k)(x - x_k) \tag{6.34}$$

och löser $L(x) = 0$ istället för $f(x) = 0$. Dvs

$$f(x_k) + f'(x_k)(x - x_k) = 0 \tag{6.35}$$

med lösningen

$$x = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{6.36}$$

Detta får bli nästa approximation:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{6.37}$$

Detta är på formen (6.33) med

$$g(x) = x - \frac{f(x)}{f'(x)}, \quad \text{dvs } \alpha(x) = \frac{1}{f'(x)}.$$

Iterationen (6.37) kallas Newtons metod (eller Newton-Raphsons metod). För att Newtons metod ska fungera måste $f'(x_k) \neq 0$. Vi antar därför att

$$f'(x) \neq 0 \quad \text{för alla } x \in I.$$

Kom ihåg att ekvationen för tangenten till grafen $y = f(x)$ i $(x_k, f(x_k))$ är $y = L(x)$, dvs

$$y = f(x_k) + f'(x_k)(x - x_k).$$

Geometriskt betyder (6.35) att vi följer tangenten och hittar $x_{k+1}$ där denna skär $x$-axeln.

**Exempel 17.** Med $f(x) = x^2 - 2$ får vi

$$g(x) = x - \frac{x^2-2}{2x} = x - \frac{1}{2}x + \frac{1}{x} = \frac{1}{2}x + \frac{1}{x}.$$

Newtons iteration blir

$$x_{k+1} = \frac{1}{2}x_k + \frac{1}{x_k}.$$

Vi har tidigare sett att $g$ är en kontraktion med $L = 1/2$ på $[1, 2]$ och denna iteration konvergerar snabbt mot $\sqrt{2}$ om vi väljer $x_0 \in [1, 2]$. □

Antag nu att $\bar{x} \in I$ är en rot, dvs

$$f(\bar{x}) = 0.$$

Om vi deriverar $g(x)$ får vi

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

så att

$$g'(\bar{x}) = \frac{f(\bar{x})f''(\bar{x})}{f'(\bar{x})^2} = 0$$

och $g'(x) \approx 0$ på varje litet intervall nära roten. Det betyder att $g$ har en mycket liten Lipschitzkonstant på varje sådant intervall. Det innebär att Newtons metod konvergerar mycket snabbt.

Följande sats handlar om detta. Det är Theorem 7 i Adams 4.6.

**Sats 8.** *Antag att $x_k, x_{k+1}, \bar{x} \in I$ och att vi har begränsningarna*

$$\frac{1}{|f'(x)|} \leq M, \quad \forall x \in I, \tag{6.38}$$

$$|f''(x)| \leq K, \quad \forall x \in I. \tag{6.39}$$

*Då gäller*

$$|x_k - \bar{x}| \leq M|f(x_k)|, \tag{6.40}$$

$$|x_{k+1} - \bar{x}| \leq \frac{1}{2}MK|x_{k+1} - x_k|^2, \tag{6.41}$$

$$|x_{k+1} - \bar{x}| \leq \frac{1}{2}MK|x_k - \bar{x}|^2. \tag{6.42}$$

Begränsningarna (6.38) och (6.39) innebär att tangentens lutning inte får vara alltför nära 0 och inte ändra sig för mycket på intervallet $I$.

Feluppskattningen (6.40) relaterar felet i $x$ till felet i $f(x)$, dvs relaterar felet $x_k - \bar{x}$ till residualen $f(x_k)$. Residualen anger hur väl $x_k$ uppfyller ekvationen $f(x) = 0$. Om $f'(x)$ är liten på $I$ så blir $M$ stor och då kan felet vara stort även om residualen $f(x_k)$ är liten. Det betyder att ekvationen är svår att lösa om $f'(x)$ är liten.

De andra uppskattningarna visar att $x_k$ konvergerar mycket snabbt om den konvergerar alls. Till exempel, om $x_k$ har kommit så nära roten att $|x_k - \bar{x}| < 1$, så ser vi i (6.42) att felet i nästa steg är proportionellt mot kvadraten på det tidigare felet, dvs det minskar mycket snabbt. Vi säger att felet konvergerar kvadratiskt mot 0. Vi erinrar oss att bisektionsmetoden och fixpunktsiterationen i allmänhet bara konvergerar linjärt.

En nackdel med Newtons metod är att det är svårt att ange ett intervall $I$ där ovanstående begränsningar gäller och sådant att iterationen stannar kvar i intervallet. I praktiken får man

oftast nöja sig med att säga att metoden konvergerar om man väljer $x_0$ tillräckligt nära en rot och prova sig fram med olika $x_0$.

Beviset kan skippas om du inte orkar mer teori.

**Bevis.** Medelvärdessatsen (Theorem 11 i Adams 2.6) ger

$$f(x_k) - f(\bar{x}) = f'(s_1)(x_k - \bar{x}), \tag{6.43}$$

där $f(\bar{x}) = 0$ och $s_1$ är en punkt mellan $x$ och $\bar{x}$. Vi får med hjälp av begränsningen (6.38)

$$|x_k - \bar{x}| = \frac{|f(x_k)|}{|f'(s_1)|} \leq M|f(x_k)|, \tag{6.44}$$

vilket är (6.40). Linjärisering med felterm (Theorem 9 i Adams 4.7) ger

$$f(x_{k+1}) = f(x_k) + f'(x_k)(x_{k+1} - x_k) + \frac{1}{2}f''(s_2)(x_{k+1} - x_k)^2 = \frac{1}{2}f''(s_2)(x_{k+1} - x_k)^2 \tag{6.45}$$

där vi använde

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0$$

från (6.37). Tillsammans med (6.40) och begränsningen (6.39) ger detta

$$|x_{k+1} - \bar{x}| \leq M|f(x_{k+1})| = \frac{1}{2}M|f''(s_2)|(x_{k+1} - x_k)^2 \leq \frac{1}{2}MK(x_{k+1} - x_k)^2, \tag{6.46}$$

vilket är (6.41). Linjärisering med felterm ger även

$$f(\bar{x}) = f(x_k) + f'(x_k)(\bar{x} - x_k) + \frac{1}{2}f''(s_3)(\bar{x} - x_k)^2, \tag{6.47}$$

där $f(\bar{x}) = 0$ så att

$$x_k - \bar{x} = \frac{f(x_k)}{f'(x_k)} + \frac{1}{2}\frac{f''(s_3)}{f'(x_k)}(x_k - \bar{x})^2. \tag{6.48}$$

Enligt (6.37) har vi

$$\frac{f(x_k)}{f'(x_k)} = -(x_{k+1} - x_k), \tag{6.49}$$

så att

$$x_k - \bar{x} = -(x_{k+1} - x_k) + \frac{1}{2}\frac{f''(s_3)}{f'(x_k)}(x_k - \bar{x})^2, \tag{6.50}$$

vilket leder till

$$x_{k+1} - \bar{x} = \frac{1}{2}\frac{f''(s_3)}{f'(x_k)}(x_k - \bar{x})^2 \tag{6.51}$$

och sedan

$$|x_{k+1} - \bar{x}| = \frac{1}{2}\frac{|f''(s_3)|}{|f'(x_k)|}(x_k - \bar{x})^2 \leq \frac{1}{2}MK(x_k - \bar{x})^2, \tag{6.52}$$

vilket är (6.42). □

## När stoppar man?

Vi vill stoppa iterationen när felet är mindre än en given tolerans:

$$|x_k - \bar{x}| \leq \text{TOL}\,. \tag{6.53}$$

Feluppskattningarna i satsen är inte användbara eftersom vi inte vill använda de svårbestända och grova begränsningarna $M$ och $K$. Men (6.50) ger

$$x_k - \bar{x} = -(x_{k+1} - x_k) + \frac{1}{2}\frac{f''(s_3)}{f'(x_k)}(x_k - \bar{x})^2,$$

så att

$$|x_k - \bar{x}| \leq |x_{k+1} - x_k| + \frac{1}{2}MK|x_k - \bar{x}|^2.$$

Om iterationen konvergerar så är den sista termen snart mycket mindre än de andra termerna och vi har

$$|x_k - \bar{x}| \approx |x_{k+1} - x_k|. \tag{6.54}$$

Stoppvilkoret

$$|x_{k+1} - x_k| \leq \text{TOL} \tag{6.55}$$

garanterar alltså (6.53) approximativt. Det betyder att vi accepterar $x_k$ om ändringen i nästa iteration är mindre än toleransen.

## Algoritmen

Algoritmen är mycket enkel.

while $|h| > \text{TOL}$
beräkna residualen: $b = -f(x)$
beräkna derivatan: $a = f'(x)$
beräkna ändringen: $h = b/a$
uppdatera $x$: $x = x + h$

Derivatan beräknas lämpligen numeriskt. Newtons metod är nämligen okänslig för fel i derivatan.