! Do not Print yet !

Preliminary Version

Subject to final Proof-reading

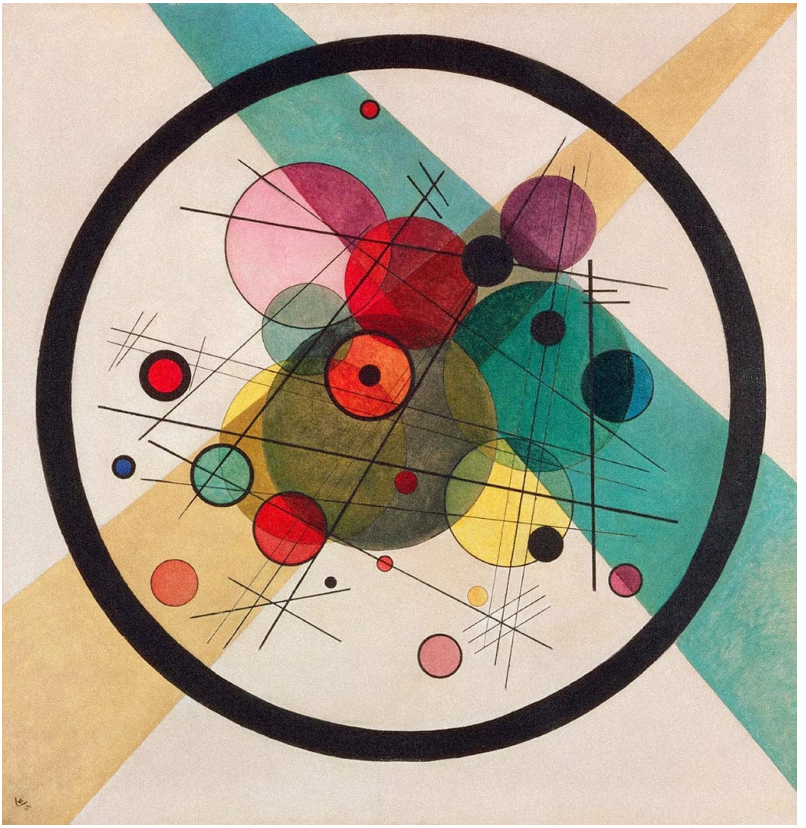
Final Version will be available
in time for SP'22

Mathematics for Computational Science

Lecture Notes for MATH 156

Alexander Hulpke
with
K. Bonney, H. Meade, M. Moy, T. Neighbors, R. Tremaine

December 1, 2021



Alexander Hulpke
Department of Mathematics
Colorado State University
1874 Campus Delivery
Fort Collins, CO, 80523

Title graphics:

Circles in a Circle (1923)
Wassily Kandinsky

Contents

Preface

For many decades, Calculus has been the epitome of College level mathematics. This is due to its undoubted usefulness in modelling the physical world for applications in Engineering and in the Physical Sciences.

But this apex role, and the associated standardization of the subject as seen in the manifold Calculus textbooks on the market, have led to a number of drawbacks:

- The underlying assumption of the course in topics and examples is that everyone ultimately wants to solve differential equations.
- Univariate calculus typically is spread over two semesters, delaying the point for students to take Linear Algebra.
- Material and presentation are firmly rooted in the 19th century, bypassing much of the developments that underly modern mathematics and its applications in information technology.
- The course (in particular as far as student learning is concerned) is heavily reliant on being able to execute recipe methods for finding (anti-)derivatives, tasks that nowadays are more than satisfactory solved by computer programs.
- With standard problems and scores providing an easy grade distribution, and asa a class with a significant failure rate, Calculus is often abused as a filter, standing in as a test for study skills and *grit*.

All of this makes the standard Calculus course an awkward choice for majors in disciplines that focus on data analysis and information processing – disciplines that barely existed when the standard Calculus sequence was created.

This book therefore takes a new approach to an introductory College mathematics course for students in Computational Science disciplines: It starts with

mathematical foundations, sufficient to enable students to take more advanced courses such as Linear Algebra, Combinatorics, Elementary Number Theory, or Abstract Algebra, which are highly relevant to their major. It then presents

The focus here is on functions as data, and what Calculus tools can do conceptionally, rather than on modelling physical phenomena or practicing manipulation of functions given through term expressions. Nor shall we delve into the borderline cases of the definitions – such as functions that are once but not twice differentiable, discontinuities for the sake of being discontinuous, or the behavior of series on the circle of convergence.

This does not mean that we will be superficial. We shall cover the concepts from univariate calculus, with applications to, and as relevant to, Computational Science, but we do not focus on solving the "standard" problems one will find on a typical Calculus exam.

Thanks

I am grateful to Kirk Bonney, Harley Meade, Michael Moy, Tristan Neighbors and Rachel Tremaine, graduate students at CSU, who corrected mistakes, provided examples and applications, as well as alternative text for graphics. Their work was supported by an Open Educational Materials grant from the Colorado Department of Higher Education, administered through the CSU Libraries, which is gratefully acknowledged.

Chapter

I

Sets and Logic

I.1 Sets and Elements

The basic "data type" of mathematics is the *set*. A set is a collection (really just a fancy word for a container that can hold things) of objects, which are called the *elements* of the set. We typically use squiggly parentheses { } to denote a set. For example, imagine we have three objects, the number 2, and two other objects we call a and b. Then

$$\{2, a, b\}$$

is the set containing exactly these three objects. To refer to it we can give it a name, $S = \{2, a, b\}$. We can now make statements about which objects are elements of this set. For example 2 is an element of the set, while 3 is not. We write this in symbols (with $\in$ reading as "is element of") in the form

$$2 \in S, \qquad 3 \notin S$$

We might also say "2 is in S" or "S contains 2", meaning exactly the same.

Sets are characterized by their membership with two sets being equal if and only if they have the same elements.

Note I.1: Mathematicians try to be very exact in the language used. The expression *if and only if* means that the property (or what is defined) – here *two sets are equal* – holds under the given condition – here *they have the same elements* – but not if the condition is violated.

An analog would be to describe an animal as an elephant if (and only if) it is large, has large floppy ears and a trunk in place of the nose.

A single *if* instead shows that one condition implies another, but is not the only reason:

If it is snowing outside, I wear gloves.

(but I also might be wearing them when it is cold and rainy).

When we describe sets by enumerating elements it does not matter in which order we write down the elements, nor if we write them down multiple times[1]. Thus

$$S = \{b, 1, a\} = \{a, 1, a, 1, b, b, a\}$$

as sets, but

$$S \neq \{1, 2, 3\}, \quad S \neq \{a, b\}, \quad S \neq \{2, a, a\}, \quad S \neq \{1, 2, a, b\}.$$

Describing Sets

In many cases, describing a set by enumerating all its elements is hard or impossible. Thus two other techniques that are used. The first of these is continuation with ... to indicate a pattern to be continued. For example, we might write

$$E = \{\ldots, -4, -2, 0, 2, 4, 6, 8, 10, \ldots\}$$

to describe the set of even integers. Similarly $R = \{5, 6, \ldots, 10\}$ would describe the integers from 5 to 10, that is $R = \{5, 6, 7, 8, 9, 10\}$. The drawback of this method is that it relies on the reader making the correct assumptions about the rule used to extend the listed numbers. It thus – and this is the second method – is often easier and better to spell out this rule explicitly as a property the objects must have. We thus can write:

$$R = \{x \mid x \text{ is integer and } 5 \leq x \leq 10\},$$

read as "The set of x such that x is an integer and $5 \leq x \leq 10$".

This notation can be quite powerful. The general pattern is to first give a variable that stands for the elements of the set, possibly indicating an(other) set that these elements are taken from. Next comes a separator, we use a vertical line | (but other punctuation marks such as ; or : are used as well). One can read this separator as "such as". Finally follows the rule or condition that the elements need to satisfy to be elements of the set.

Thus, if we use $\mathbb{Z}$ to denote the set of all integers, we could have described the above examples as $E = \{x \in \mathbb{Z} \mid x \text{ even}\}$ (read as "The set of those integers x, such that x is even") and $R = \{x \in \mathbb{Z} \mid 5 \leq x \leq 10\}$.

The reason for specifying a set from which elements are chosen is to avoid any ambiguity what kinds of objects are in the set (do we allow rational numbers between 5 and 10 in the set R?), and to make the specification clear.

[1] Of course, sometimes one might want to be able to describe objects with multiplicities. We will see how to describe this below ⇝ I.6

Formally, this specification is called a *predicate*, using language from grammar, in which a predicate is a part of a sentence that gives information about the subject. For example, in the following sentence:

$$\underbrace{\text{The house}}_{\textit{subject}}\ \underbrace{\text{is painted green.}}_{\textit{predicate}}$$

Definition I.2: A *predicate* (for a given set Y) is a sentence, involving a variable x, such that if we substitute x by a particular element $a \in Y$, the sentence becomes a statement that is either (and unambiguously) true or false.

For example, *x has brown fur*, would be a possible predicate for the set Y of all animals, this predicate would be true for a brown bear, but false for a frog.

Note that determining the truth value of a predicate for a particular element might be hard, or even impossible at a given time. (Imagine for example the property for a sequence of words to occur in at least one book in a huge library.) But there cannot be ambiguity about whether the property is true for a particular x. Thus, for the set of pictures, *x is art* would not be a predicate, while *x is letter size* would be one.

Finally, there is a variant that describes elements from transforming elements of another set. (Sometimes it is easier to describe what to do with elements, than to give a property.) We thus can write $E = \{2y \mid y \in \mathbb{Z}\}$, using the property that even integers are exactly the multiples of 2.

For a set A, we define the *cardinality*, denoted by $|A|$ or $\#A$ as the number of elements in A.[2].

Some common sets

With much of mathematics working with numbers, it will be convenient to give special names for some sets of numbers:

$\mathbb{Z}$ The integers, $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, 4, \ldots\}$

$\mathbb{Q}$ The rational numbers, fractions of integers. $\mathbb{Q} = \{\frac{a}{b} \mid a, b \in \mathbb{Z}, b \neq 0\}$.

$\mathbb{N}$ The natural numbers $1, 2, 3, \ldots$. Often it is ambiguous between different authors whether 0 should be part of this, thus we will write $\mathbb{N}_0$ (or $\mathbb{Z}_{\geq 0}$) if we want to guarantee that 0 is an element, respectively $\mathbb{N}_{>0}$ (or $\mathbb{Z}_{>0}$) if we explicitly want to exclude 0.

$\mathbb{R}$ The real numbers on the number line. These numbers can be described by possibly infinite decimal expansions. The proper, formal definition however is more complicated and it will take us quite a while (at the end of Chapter IV) to describe their formal definition.

[2] This is a somewhat vague definition. We will give a more formal definition below in ⇝ III.6

I.2 Subsets

With the basic operation for sets being a test for membership, an obvious property for two sets A, B is that one set contains every element of the other.

DEFINITION I.3: If A, B are sets, we call A a *subset* of B if every element of A is also an element of B. That is $x \in A$ implies that $x \in B$ (formally: $x \in A \Rightarrow x \in B$). We write $A \subset B$. When talking about sets being subsets of others, this is also called *inclusion* of subsets.

NOTE I.4: Some authors distinguish between *subset, could be equal* (symbol $\subseteq$) and *proper subset, not equal* ($\subset$). We do not do this and will state explicitly ($\subsetneq$) if a subset is proper (that is, not equal).

To provide for examples in this section, let

$$\begin{aligned} C &= \{x \in \mathbb{Z} \mid 0 \leq x \leq 9\} = \{0,1,2,3,4,5,6,7,8,9\} \\ D &= \{0,2,4\} \\ E &= \{x \in \mathbb{Z} \mid x \text{ is even}\} \\ F &= \{0,1,2\}. \end{aligned}$$

Then (for example) $D \subset C$, $D \subset E$, $C \not\subset D$, $C \not\subset E$.

Since membership test is the basic operation for sets, one often reduces equality of sets to two subset test:

LEMMA I.5: Let A, B two sets. Then $A = B$ if and only if $A \subset B$ and $B \subset A$.

Proof: First assume that $A = B$. We want to show that $A \subset B$. For this, let $x \in A$. Then $x \in B = A$, so $A \subset B$. By swapping the role of A and B we get $B \subset A$ as well.

Vice versa, assume that $A \subset B$ and $B \subset A$. Then $x \in A$ implies $x \in B$ and $x \in B$ implies $x \in A$, that is both sets have the same elements and thus are equal. □

The Empty Set

It is often useful (for example for constructing certain sets, or to handle borderline cases) to refer to the *empty set* $\varnothing = \{\}$, that is the set which contains no element. It is a subset of every set.

Sets of Sets and Subtleties

Sets can contain anything and thus one set can be an element of another set. Indeed, this will be used later to build more complicated structures from sets. For example, if we have

$$A = \{1, 5, \{2, 3\}\}$$

then $1 \in A$ and $\{2,3\} \in A$. Or consider

$$B = \left\{ S \subset \{1,2,3\} \;\middle|\; |S| = 2 \right\} = \{\{1,2\},\{1,3\},\{2,3\}\}\,.$$

In such situations, the wrapping level is important: Being in a subset contained in another set is not the same as being an element, i.e. $2 \notin A$. Nor is element the same as subset, we have $\{2,3\} \in A$ but $\{2,3\} \not\subset A$, and indeed (note the extra parentheses $\{\{2,3\}\} \subset A$. And of course $A \neq \{1,2,3,5\}$.

The Power Set, Hasse diagrams

The *power set* of a set X is the set

$$\mathcal{P}(X) = \{S \mid S \subset X\}$$

whose elements are the subsets of X, including the empty set and X itself. For example, if $X = \{1,2,3\}$, then

$$\mathcal{P}(X) = \{\{\},\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},X\}\,.$$

For a finite set X with $|X| = n$, we can describe the elements of the power set by the bit[3]-lists of length n: Given a subset $S \subset X$, we form a bit-list by setting the i-th bit to 1 if and only if the i-th element of X is contained in S. (Indeed, such bit-lists are good way of representing the subsets of a set on a computer.) Thus there are as many subsets as there are bit-lists, namely 2^n subsets of a set of size n.

A convenient way of depicting the subsets of a set is the *Hasse diagram*[4]: sets are represented by dots (maybe labeled with the set name or the set itself). If a set A is contained in another set B, we place the dot for B higher than the dot for A, and connect the two dots by a line, indicating that the lower placed set is contained in the higher placed one. Finally, we leave out (or delete) lines, that indicate a connection that is already implied by connections to an intermediate set. That is, if $A \subset B$ and $B \subset C$ (and thus also $A \subset C$), we draw lines $A - B$ and $B - C$, but not $A - C$. Figure I.1 shows the Hasse diagram for the 8 subsets of $X = \{1,2,3\}$. The actual diagram is given by the solid lines. The grey dashed lines indicate set inclusions that will not be drawn, as they are implied already by connections with intermediate sets.

The same idea can be used, of course, for arbitrary sets and subsets. Figure I.2 depicts the Hasse diagram for three different collections of subsets of the 4-element set $\{0,a,b,c\}$.

[3] A bit is a variable that can have only two values, often denoted by 0 and 1, or by "false" and "true".

[4] Named after the German mathematician Helmut Hasse (1898-1979), who made effective use of such diagrams, but did not invent them.

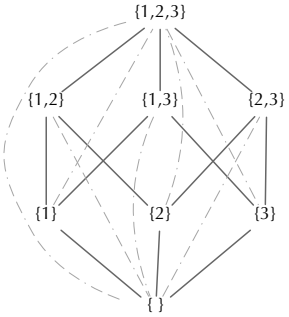


Figure I.1: The Hasse diagram for all subsets of $\{1, 2, 3\}$.

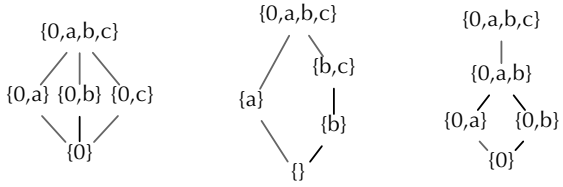


Figure I.2: The Hasse diagram for certain subsets of $\{0, a, b, c\}$.

I.3 Intersection, Union, Difference, and Complement

Next, we define a number of operations that construct new sets from old ones, for example by taking common elements.

Definition I.6: Let A, B be two sets. The

intersection of A and B is the set of those elements that are in A and in B:

$$A \cap B = \{x \in A \mid x \in B\} = \{x \in B \mid x \in A\}$$

union of A and B is the set of elements that are in A, together with the elements in B:

$$A \cup B = \{x \mid x \in A \quad \text{or} \quad x \in B\}$$

difference of A and B is the set of elements that are in A but not in B:

$$A \setminus B = \{x \in A \mid x \notin B\}.$$

(Note that some authors simply write $A - B$.)

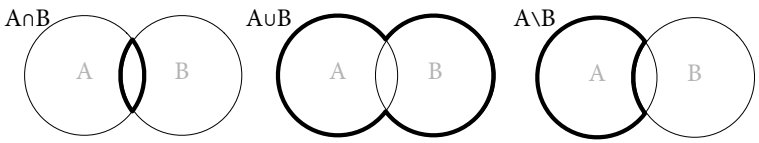


Figure I.3: Intersection, Union, and Difference

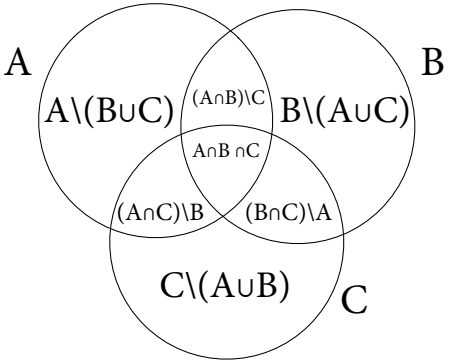


Figure I.4: A Venn diagram for three sets

In the case that $B \subset A$ is understood from the context, this difference is sometimes called the *complement* of B in A and denoted by $B^{\complement}$. Clearly

$$(B^{\complement})^{\complement} = A \setminus (A \setminus B) = B$$

In the examples from Section I.2, we have that $C \cap E = \{0, 2, 4, 6, 8\}$, $C \cap D = D$, $D \cap F = \{0, 2\}$, $D \cup F = \{0, 1, 2, 4\}$, $C \cup D = C$, $C \setminus D = \{1, 3, 5, 6, 7, 8, 9\}$, $D \setminus F = \{4\}$. And if we assume that all sets are subsets of C (such a set containing everything in a given context is sometimes called an *universe*), we have that $D^{\complement} = \{1, 3, 5, 6, 7, 8, 9\}$ and $C^{\complement} = \varnothing$.

If they are included in a Hasse diagram (which does not hold for all diagrams in Figure I.2!), the union of two sets A and B will be the (minimal) dot above A and B, the intersection the maximal dot below A and B.

Another nice way of illustrating sets and the intersections and unions is by using a *Venn diagram*, in which sets are represented by areas in the plane. Figure I.3 illustrates intersection, union and difference in such a diagram, Figure I.4 labels the 7 areas of a 3-set Venn diagram.

We observe the following basic relations for these operations:

Lemma I.7: Let A, B, C be sets. Then

1. $A \cap B = B \cap A$.
2. $A \cup B = B \cup A$.
3. $A \cap B \subset A$.
4. $A \subset A \cup B$.
5. $A \setminus B \subset A$.
6. $(A \setminus B) \cup (A \cap B) = A$.
7. $(A \setminus B) \cap (A \cap B) = \varnothing$.
8. $(A \cap B) \cap C = A \cap (B \cap C)$ (so we can write $A \cap B \cap C$ without ambiguity).
9. $(A \cup B) \cup C = A \cup (B \cup C)$ (so we can write $A \cup B \cup C$ without ambiguity).

Proofs are left as exercise for the reader.

Distributive Laws

The following properties are not trivial, but can be easily seen in a Venn diagram:

THEOREM I.8 (Distributive laws): Let A, B, C be sets. Then (Figure I.5):
a) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Note that these rules mimic what happens if we multiply by a sum: $a(b + c)$.
Proof: Since we have to show equality of sets, we need to show two-sided subset inclusion.
a) We show first that $A \cap (B \cup C) \subset (A \cap B) \cup (A \cap C)$: For this, let $x \in A \cap (B \cup C)$. This means that $x \in A$ and also $x \in B$ or $x \in C$. In the first of these cases we have that $x \in A \cap B$, in the second that $x \in A \cap C$. Thus in either case $x \in (A \cap B) \cup (A \cap C)$. For the reverse inclusion, let $x \in (A \cap B)$. Then $x \in A$ and $x \in B \subset B \cup C$, so $x \in A \cap (B \cup C)$. Similarly (swap B and C) we see that $x \in A \cap C$ implies $x \in A \cap (B \cup C)$ as well. This show that $A \cap (B \cup C) \supset (A \cap B) \cup (A \cap C)$.
b) Exercise □

De Morgan's Laws

Next we look at rules to simplify complement operations

THEOREM I.9 (De Morgan's laws): Let U be a set (a universe) with $A, B \subset U$. Then (Figure I.6):
a) $(A \cup B)^{\complement} = A^{\complement} \cap B^{\complement}$.
b) $(A \cap B)^{\complement} = A^{\complement} \cup B^{\complement}$.

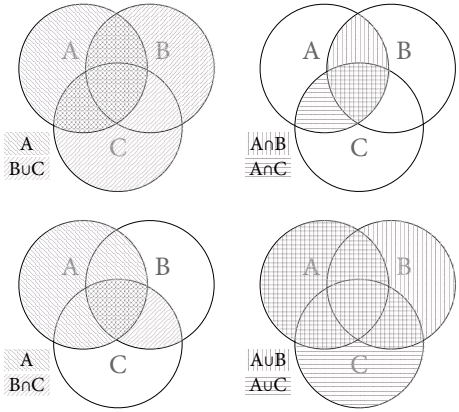


Figure I.5: Distributive Laws for Union and Intersection

Proof: Again, to show equality of sets, we need to show inclusion in two directions:
a) Let $x \in (A \cup B)^{\complement}$. That means that $x \in U$ but $x \notin A \cup B$. But that means that x can be neither in A, nor in B, so $x \in A^{\complement}$ and $x \in B^{\complement}$ and thus $x \in A^{\complement} \cap B^{\complement}$. Conversely, let $x \in A^{\complement} \cap B^{\complement}$. That means that $x \in U$ but $x \notin A$ and $x \notin B$, and thus $x \notin A \cup B$. This implies that $x \in (A \cup B)^{\complement}$.
b) Exercise □

Disjunctive Normal Form

Using the laws introduced in the previous two sections, it is possible to simplify a complicated expression involving sets into a simpler form. In particular, it is possible to transform any such expression into a

Union of

Intersections of

Sets or Complements of sets.

Such a form is called *disjunctive normal form* (*DNF*). For example,

$$(A \cap B^{\complement} \cap C) \cup (A^{\complement} \cap D)$$

is in DNF, while

$$(A \cup B^{\complement} \cup C) \cap (A^{\complement} \cap D)^{\complement}$$

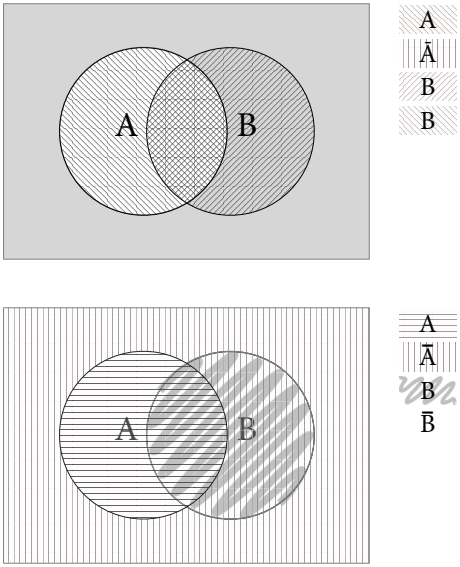


Figure I.6: De Morgan's Laws for Sets

is not. However we can transform this expression stepwise into DNF:

$$\begin{aligned}
& (A \cup B^{\complement} \cup C) \cap (A^{\complement} \cap D)^{\complement} \\
= \; & (A \cup B^{\complement} \cup C) \cap ((A^{\complement})^{\complement} \cup D^{\complement}) \\
= \; & (A \cup B^{\complement} \cup C) \cap (A \cup D^{\complement}) \\
= \; & (A \cap (A \cup D^{\complement})) \cup (B^{\complement} \cap (A \cup D^{\complement})) \cup (C \cap (A \cup D^{\complement})) \\
= \; & (A \cap A) \cup (A \cap D^{\complement}) \cup (B^{\complement} \cap A) \cup (B^{\complement} \cap D^{\complement}) \cup (C \cap A) \cup (C \cap D^{\complement}) \\
= \; & (A) \cup (A \cap D^{\complement}) \cup (B^{\complement} \cap A) \cup (B^{\complement} \cap D^{\complement}) \cup (C \cap A) \cup (C \cap D^{\complement})
\end{aligned}$$

If we imagine a Venn diagram of the sets, the disjunctive normal form describes how the set can be composed from minimal intersecting parts.

I.4 Connections to Logic

The distributive laws and De Morgan laws might to some readers look very similar to statements about operations in logic. Here we have truth values that can be true or false, and we can combine them with *and* (symbol $\wedge$), *or* (symbol $\vee$ [5]), and negate

[5] This is actually the origin of these symbols. The Latin word for "or" is "vel".

them (symbol $\neg$). If we take distributive laws or De Morgan laws and replace $\cap$ by $\wedge$, $\cup$ by $\vee$ and $^{\complement}$ by $\neg$ (placed before instead of exponents), we get the following valid logic laws (we call the variables P, Q, and R here):

1. $P \wedge (Q \vee R) = (P \wedge Q) \vee (P \wedge R)$.
2. $P \vee (Q \wedge R) = (P \vee Q) \wedge (P \vee R)$.
3. $\neg(P \vee Q) = (\neg P) \wedge (\neg Q)$.
4. $\neg(P \wedge Q) = (\neg P) \vee (\neg Q)$.

The reason for this is easy, if we consider P, Q, R as predicates (that is functions that give truth values) for objects in a set U that are true for some elements and false for others. We then define:

$$\begin{aligned} A &= \{x \in U \mid P(x) = \text{true}\}, \\ B &= \{x \in U \mid Q(x) = \text{true}\}, \\ C &= \{x \in U \mid R(x) = \text{true}\}, \end{aligned}$$

and observe that $A \cap B$ is the set of objects for which $P \wedge Q$ is true, $A \cup B$ the set for which $P \vee Q$ is true and $A^{\complement}$ the set for which $\neg P$ is true.

In the same way as with sets we have a disjunctive normal form. That is every logical expression can be written as an "or" combination of "and" combinations of variables or their negations. Such a form can be useful in evaluating the truth value of a more complicated logical expression, and there is a method (called the *Quine-McCluskey algorithm*) to convert a logical expression into a unique, minimal disjunctive normal form.

I.5 Predicate Logic and Quantifiers

Using predicates, we can start to construct mathematical statements. Such statements typically assume that if some properties hold for an object (that is some predicate P has true value $P(x)$ on an element x), then some other property holds (i.e. some other predicate Q will have value $Q(x)$ as true. We can write this a $P(x) \Rightarrow Q(x)$, and combine predicates with logical operations.

To write mathematical theorems, we however need two more operations, called *quantifiers*. They specify whether a predicate holds for all objects in a set (e.g. *all cats are grey*), or if there is (at least) one object for which the property holds (e.g. *there is a highest mountain*). The former is called a *universal statement*, the latter an *existence statement*. We write down a universal statement by writing down "for all" (often using the symbol $\forall$), followed by naming an element from a set for which the property is to be stated. This element stands for any (or all) elements of the set. In the case of an existence statement we write "there exists" (using the symbol $\exists$), again followed by selecting one element of a set, which is to be the one element

for which the following property is claimed. For example, using $E \subset \mathbb{Z}_{\geq 0}$ for the set of nonnegative even numbers, and $D = \mathbb{Z}_{\geq 0} \setminus E$ for the set of nonnegative odd numbers:

- The sum of an even number and 1 is odd: $\forall x \in E : x + 1 \in D$.
- The sum of two odd numbers is even $\forall x, y \in D : x + y \in E$
- The number 1234567 is composite: $\exists x, y \in \mathbb{Z}; x, y > 1 : 1234567 = x \cdot y$.
- Every even number is a multiple of 2: $\forall x \in E \exists y \in \mathbb{Z} : x = 2y$. (The : here can be read as "such that").
- Every even number is (strictly) smaller than another even number: $\forall x \in E \exists y \in E : y > x$.

Note that the order, in which we write the quantifiers of different type is important, and that changing the ordering will change the meaning of the statement: For example $\forall x \in \mathbb{Z} \exists y \in \mathbb{Z} : y = x + 1$ (*for every integer there exists one that is larger by one*, clearly a true statement) versus $\exists y \in \mathbb{Z} \forall x \in \mathbb{Z} : y = x + 1$ (*there is an integer (y) that is one larger than every other integer*, a nonsensical claim). (Adjacent quantifiers of the same type can be exchanged.)

Most mathematical statements (either as part of a definition, or in the claim of a theorem) can be formulated as a combination of $\exists$, $\forall$, :, and $\Rightarrow$ (implication) statements. A formal proof of such a statement then will follow this pattern:

If the statement starts with:	Then the proof:
An existence statement: $\exists x \in S \ldots$	Construct (or describe a way how to find; effectively it will be an algorithm) an element $x \in S$ that has the required property (which will be given in the following part of the statement). A rarer, more complicated, construction is to imply that such an element must exist, without giving an explicit way of finding it. Such a proof would be called *non-constructive*, and not allow translation to an effective algorithm.
A universal statement: $\forall x \in S \ldots$.	The proof will start with a sentence: "Let $x \in S$" (and the implicit fact that x is to be an arbitrary element, or that any element of S could be selected here). Then the proof will need to show that the remaining part of the statement holds for x. For this we can use only the fact that $x \in S$ (and the properties implied by it).
An implication between predicates $P(x) \Rightarrow Q(x)$.	We assume that the element x satisfies the predicate P, and need to shows that x also satisfies the predicate Q.

I.6 Pairs, Tuples, Cartesian Product

In general, we do not store all items we use in big bags that get everything thrown in together. Instead we often use more structured storage, say boxes in drawers. The same is true in mathematics, where we will often find it convenient to have objects stored in a form with more structure than a set. This section describes some of the ways how this can be done.

We start with the definition of pairs: A *pair* is an object (a, b), that holds two objects (namely a and b) in two distinct positions, enclosed by parentheses (or sometimes brackets: $[a, b]$). The pair $(1, 2)$ is different from the pair $(2, 1)$, while the sets $\{1, 2\}$ and $\{2, 1\}$ are the same. (Note that sets and tuples formally are different: $\{1, 2\} \neq (1, 2)$.)

Formally, we can construct pairs as sets. We simply define (a, b) as a shorthand for the set $\{a, \{a, b\}\}$. Note that this form allows us to always extract the first (a) and second (b) entry unambiguously. (The reason for building pairs from sets is that it allows to re-use existing language and theorems, rather than having to re-do everything for the new kind of objects.)

If we have two sets A, B, we sometimes want to look at the set of all possible pairs whose first entry is from A and whose second entry is from b. This set is

called the *Cartesian product*[6] of A with B and written as

$$A \times B = \{(a, b) \mid a \in A, b \in B\}$$

For example, if $A = \{1, 2, 3\}$ and $B = \{5, 6\}$, then

$$A \times B = \{(1, 5), (1, 6), (2, 5), (2, 6), (3, 5), (3, 6)\}$$

It is easy to see that if $|A| = m$ and $|B| = n$, then $|A \times B| = m \cdot n$.

The same process that leads to pairs can be extended or repeated (e.g by forming pairs, whose entries are pairs again) and thus form ordered collections of more than two objects, such as (a, b, c) or $(3, 1, 4, 1, 5, 9)$. We call such objects *tuples*, typically together with a number that indicates the number of entries. Thus $(7, 3, 5)$ is a 3-tuple and $(8, 2, 5, 1, 4, 1, 2)$ an 7-tuple; pairs could be called 2-tuples. And of course the set of all k-tuples can be considered as an iterated Cartesian product $A \times B \times C \times \cdots \times K$.

Conceptually, pairs and tuples will allow us to group information consisting of different parts (that may not be mixed up) together. It is the first step towards structured data.

Indexing and index sets

If we have a tuple, we might want to refer to a particular entry of it. The easiest way to do so is by referring to the place in the tuple, at which the entry lies. We often do so by adding an *index* to the name of the tuple, that is a number put on the bottom right of the object. For example, if $t = (p, r, q)$, we have $t_1 = p$, $t_2 = r$ and $t_3 = q$. (Sometimes people like to start at 0 instead. Which convention is used needs to be stated or be clear from the context.) We thus could write

$$t = (t_1, t_2, \ldots, t_k)$$

for a particular k-tuple. Programmers might prefer to write $t[i]$ instead. The index notation simply uses less ink and space.

We can put the possible positions (i.e. the possible indices) into an *index set*, and use this to refer to the entries: Let $I = \{1, 2, \ldots, k\}$ and we will talk about t_i for $i \in I$.[7]

This concept (and notation) generalizes easily to other, even infinite, index sets. We might take an index set P of persons and then talk of the first name f_p for a

[6] Named after the French mathematician René Descartes, who invented the standard x/y coordinate set, the *Cartesian coordinates*, in which points are described by pairs.

[7] The reader who already has some programming experience should think at this point of a `for`-loop. The i (for "index") from mathematics is also the reason that the standard name of a loop variable is `i` as well.

person $p \in P$. Or we look at entries t_i, $i \in \mathbb{N}$ for an infinite tuple (which we will call a sequence ⇝ IV.1) $(t_1, t_2, t_3, \ldots)$.

Mathematics here only cares about the fact that the indexed object t_i is determined clearly and unambiguously from the index i, while implementing such objects on a computer, in particular for more complicated index sets, can bring up questions of efficiency – finding the object associated to a particular index. But that is a topic you will learn about in courses on algorithms and data structures.

The entries of a tuple can be tuples again, and we can refer to the entries of entries by multiple indices. Here one typically will write $t_{i,j}$ instead of a more clumsy $(t_i)_j$. For example, suppose we have a 2-tuple t (depicted here written vertically), whose entries are 3-tuples. We then can refer to the entries of this object (also called a *matrix*) as follows

$$\begin{pmatrix} (t_{1,1}, & t_{1,2}, & t_{1,3}), \\ (t_{2,1}, & t_{2,2}, & t_{2,3}) \end{pmatrix}.$$

Contrary to the x/y coordinates of geometry, here the first entry typically indexes the row and the second entry the column. Again this can be generalized to objects of higher dimension

Generating Tuples and Subsets

In understanding tuples and index sets, it can be helpful to think about how one would generate such objects in algorithms. First, lets look how an algorithm would generate all pairs in $A \times B$: We select all possibilities for the first entry and for each such choice select all possibilities for the second entry. If the sets A and B are given as input, this is simply a set of iterated `for` loops:

```
for a in A do
  for b in B do
    print (a,b); # or other processing, as desired
  od;
od;
```

If we replaced the second `for` loop by `for b in A do`, we would get the pairs in $A \times A$. To enumerate $A \times A \times A$ would be three iterated for loops and so on.

To describe all subsets of a set, we re-use the idea we already encountered in Section I.2, of indicating by bits whether an element lies in the set or not. If we have a set A with n elements, the subsets thus can be described as n-tuples in $\underbrace{B \times \cdots \times B}_{n \text{ factors}}$ where $B = \{0,1\}$. For example, if $n = 3$, the following algorithm would construct these subsets

```
for a in {0,1} do
  for b in {0,1} do
```

```
    for c in {0,1} do
      print (a,b,c);
    od;
  od;
od;
```

(Writing such an algorithm for an *arbitrary* (user-selected) value of n is a bit more difficult, and recursion might be the easiest way to do so.)

In using bit lists to describe subsets we of course need to fix an ordering of the elements of A. For example if we have $A = \{\text{dog}, \text{bird}, \text{cat}\}$, and consider the elements in this order, we get the following correspondences:

$$\begin{aligned}(0,0,0) &\longleftrightarrow \varnothing\\ (1,0,1) &\longleftrightarrow \{\text{dog}, \text{cat}\}\\ (0,1,0) &\longleftrightarrow \{\text{bird}\}\\ (1,1,1) &\longleftrightarrow \{\text{dog}, \text{bird}, \text{cat}\}.\end{aligned}$$

But if we used $A = \{\text{dog}, \text{cat}, \text{bird}\}$ (this is the *same* set, we just change the convention in how we order elements), the tuple $(1,0,1)$ would describe the subset $A = \{\text{dog}, \text{bird}\}$. (Formally, each arrangement give us a different bijective function between the binary tuples and the subsets.)

Multisets

We can use the construct of an index set to associate counts to objects of a set and thus represent objects being in a set multiple times. The resulting object is called a *multiset*. Formally, a multiset can be defined as an ordinary set S, together with a counting set $C = \{c_s \in \mathbb{N} \mid s \in S\}$ indexed by S. This pair (S, C) then represents a collection in which object $s \in S$ occurs c_s times. For example, we could describe a wallet's content by the set $S = \{p, n, d, q\}$ [8] and have e.g.

$$W = (S, C), \quad C = \{c_p = 3, c_n = 1, c_d = 0, c_q = 3\}.$$

[8] US-centric *penny* (1 cent) ,*nickel* (5 cent),*dime* (10 cent), *quarter* (25 cent).

Chapter

II Relations

II.1 Connecting Elements

In the same way that skis without slopes are of limited excitement, just talking about individual sets will not get us very far. Instead, we want to connect elements of one set with elements if other sets (could be different or the same). This can be used to describe information that is more than just an accumulation of objects: we can describe relations (such as Parent/Child, or Sibling), properties associated to objects (such as age or color), or describe more complicated structures (a travel network, composed from point-to-point connections).

The tool for doing this is that of a relation, described in this section.

Definition II.1: Let A, B be sets. A *relation* between A and B is a subset $R \subset A \times B$, that is R is a set of pairs (a, b) with $a \in A$ and $b \in B$. We say that $a \in A$ is in relation to $b \in B$ (sometimes written $a \sim_R b$ (or even just aRb) if and only if $(a, b) \in R$.

Similarly, we write $a \not\sim_R b$ (or $a \not R b$) to denote that $(a, b) \notin R$.

We could for example take A the set of all students and B the set of all majors with the relation describing the major(s) of every student.[1]

Another example of a relation would be A the set of natural numbers and B the set of prime numbers with the relation R defined as a number a being in relation to a prime b if and only b divides a. Then for example $4 \sim_R 2$ but $4 \not\sim_R 3$. Also note that $2 \not\sim_R 4$. Then some of the elements of R are

$$(2,2), (4,2), (6,2), (3,3), (6,3), \ldots (20,2), (20,5), \ldots \in R.$$

[1] Note that multiple students might have the same major and that some students might have multiple majors.

Note that the relation is just what we defined. For example we have that $(-10, 2) \notin R$ and $(12, 6) \notin R$, since neither of these two pairs would be in $A \times B$.[2]

What is important to remember is that a relation is a subset of $A \times B$. It can be as small as the empty set (no elements in relation, not particularly interesting), and as large as all of $A \times B$ (every element of A in relation to every element of BN, again not very interesting), but typically will be a proper nonempty subset.

The examples above show two basic uses of relations. One is to associate objects of a set with objects from another set. If each objects gets associated with a single object (e.g. the competition number on a bib, assigned to each runner in a race), this can be interpreted as an assignment and will later on get us to the concept of a function.

The other use is a relation among objects in the same set, which can be used to establish clusters, families or hierarchies. Such a relation amongst objects of a set is often called a *binary relation* on the set, in particular if using a notation similar to $a \sim b$. For example, consider the well known operations $=, <, \leq$ on the rational numbers: For the relation $\leq$, say, we have that $(3, 5)$ is in the relation, but $(5, 3)$ is not. As we will see, relations thus allow us to generalize concepts of equality or order. For example, if we wanted to model rounding to integers, we could define a relation $\sim$ on the rational numbers by defining $a \sim b$ if $-10^{-5} < a - b < 10^{-5}$.

Describing Relations

We have already seen two ways of defining a relation. The first is to describe the elements in relation as a list of pairs in the cartesian product. For example, if we take the set $A = \{1, 2, \ldots, 6\}$ and the relation "strictly smaller", we get

$$\{(1, 2), (1, 3), \ldots, (1, 6), (2, 3), (2, 4), \ldots, (5, 6)\}$$

A variant of this is to lists the pairs in relation in a table:

a	b
1	2
1	3
⋮	⋮

This notation indicates that it is possible to add further columns, leading to the concept of an n-ary relation. Such relations are the underlying concept of a *relational database*, but we will not study them further here.

A further variant of this description is the *digraph* (short for "directed graph"). We draw the sets A and B on two sides and connect element $a \in A$ to $b \in B$ by an arrow, whenever $(a, b) \in R$. Figure II.1, left depicts this for the example.

[2] One could of course extend the divisibility relation to larger sets, and then have these pairs in the new, larger, relation.

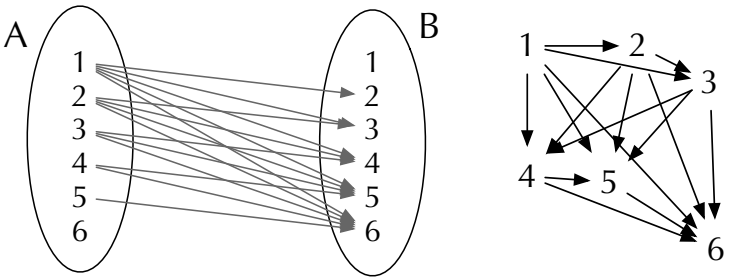


Figure II.1: The "strictly smaller" relation described by digraphs

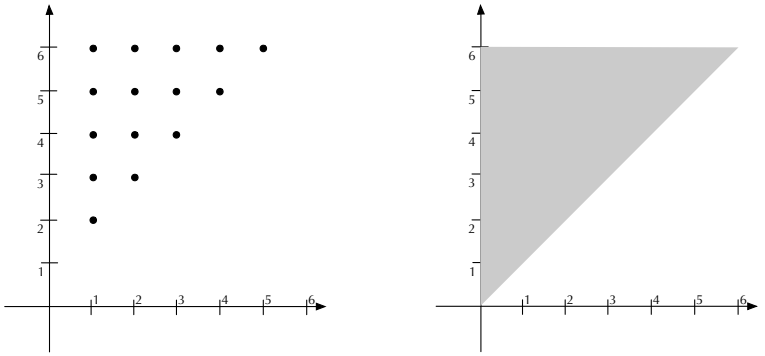


Figure II.2: The strictly smaller relation on integers and on real numbers

If, as in this example, we have that $A = B$, we could also draw arrows between elements of $A = B$.

The second way of description is to give a predicate that identifies the pairs in relation. In the example, this predicate would be: $S(a, b)$ if $a < b$. It then is often convenient to write the predicate as a connecting symbol, i.e. $3S5$. You have used symbols such as $\leq$, $\subset$, or $\in$ before, formally they all denote relations.

The third way of describing relations is one you will have seen before, but maybe not under this name. Assume that we can arrange the sets A and B along a line. (This is easy if A and B are subsets of the rational or the real numbers.) We then can interpret the pairs (a, b) in relation as coordinates of points in the plane $A \times B$. We call this the *graph* of the relation. Figure II.2, left, shows the graph of the "strictly smaller" relation on $A = \{1, 2, \ldots, 6\}$, the right image then shows the graph of the same relation on the set $B = \{0 \leq x \leq 6\} \subset \mathbb{R}$.

A number of further examples are shown in Figure II.3. In all of these examples

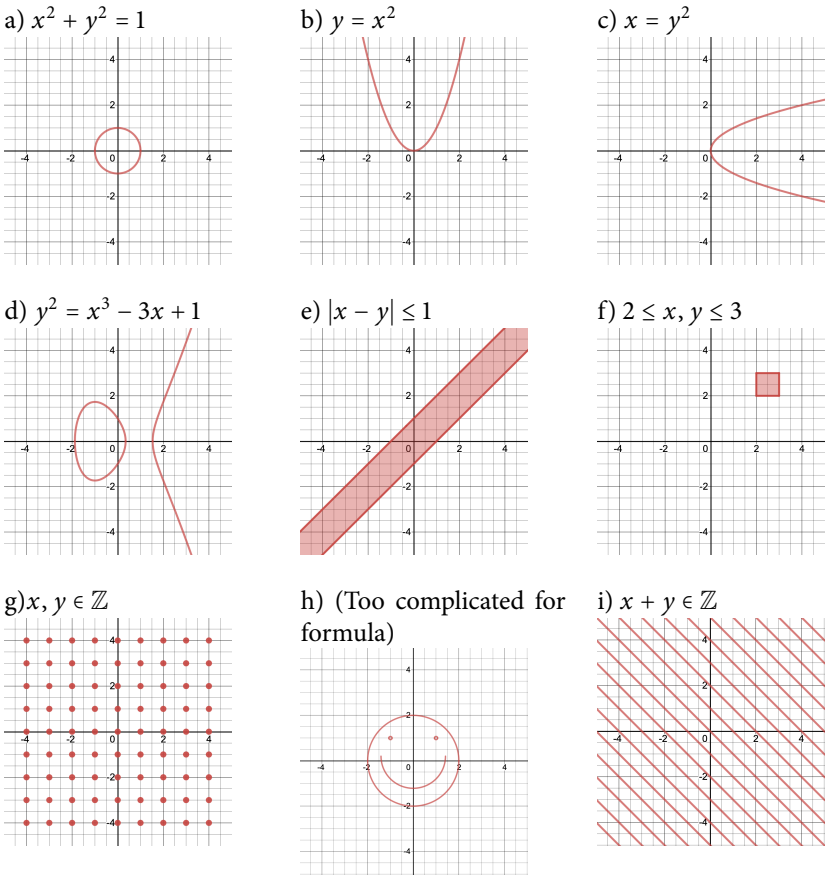


Figure II.3: Some relations on $\mathbb{R} \times \mathbb{R}$

we have $A = B = \mathbb{R}$ and indicate the condition for a pair (x, y) to be in the relation.

Domain, Range, Source and Target

To talk about relations, it will be useful to define a number of terms.

DEFINITION II.2: Let $R \subset A \times B$ a relation. We call A the *source* of R and B the *target* of R.

The set

$$\{a \in A \mid (a, b) \in R \text{ for some } b \in B\} \subset A$$

is called the *domain* of R, while

$$\{b \in B \mid (a, b) \in R \text{ for some } a \in A\} \subset B$$

is called the *range* of R.

In the above example of the strictly smaller relation on $A = B = \{1, \dots, 6\}$, we have that source and target are both equal to $A = B$. The domain is $\{1, 2, 3, 4, 5\}$ (as 6 is not smaller than any number), while the range is $\{2, 3, 4, 5, 6\}$.

II.2 Complements, Converse and Composition

Sometimes it can be convenient to build new relations from existing ones – e.g. the relation of "parent" implies also the relations of "child" and of "grandparent". Here are some ways to build new relations from old ones.

The first observation is that a relation is a set and thus subject to set operations. If $R \subset A \times B$ is a relation, the complement

$$R^{\complement} = \{(a, b) \mid (a, b) \notin R\} \subset A \times B$$

is the logical negation with $aR^{\complement}b$ if and only if $a \not\sim_R b$. For example, if we take for R the relation "parent", then $R^{\complement}$ is the relation "is not parent".

The next operation is that we swap the pairs in R around (or reverse the direction of the arrows in the digraph representation). This is called the *converse* of R. Formally we define the converse as

$$\{(b, a) \in B \times A \mid (a, b) \in R\}.$$

For example if R is the relation "is parent of", then its converse is the relation "is child of".

The third operation turns out the most useful, but also maybe most complicated one. Here we take two relations $R \subset A \times B$ and $S \subset B \times C$ such that the target of the first relation is the source of the first. The *composition* of R with S is the relation

$$\begin{aligned} S \circ R &= \{(a, c) \in A \times C \mid \exists b \in B : (a, b) \in R \text{ and } (b, c) \in S\} \\ &= \{(a, c) \in A \times C \mid (a, b) \in R \text{ and } (b, c) \in S \text{ for an element } b \in B\} \end{aligned}$$

Note that we write the composition in *reverse order* as $S \circ R$, with a circle $\circ$ as connection. (Why so? Because it fits with how functions are used, as we will see later ↝ III.3)

For example, if R is the relation "is parent of" and S is the relation "is spouse of", then $S \circ R$ is "parent of spouse" or "parent-in-law" – aRb and bSc means that a is parent of b and b is spouse of c. On the other hand, $R \circ S$ is "spouse of parent" (that is parent or step-parent),

Composition is probably easiest visualized in the digraph model. We are connecting $a \in A$ to all $c \in C$ that can be reached by following arrows from a through elements $b \in B$.

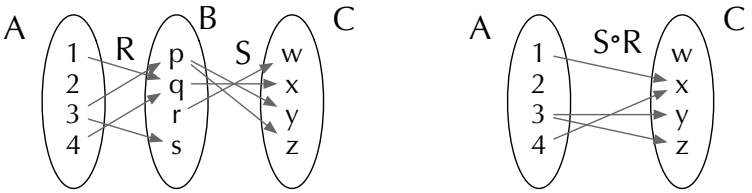


Figure II.4: Composition of relations

For example, with $A = \{1, 2, 3, 4\}$, $B = \{p, q, r, s\}$ and $C = \{w, x, y, z\}$, Figure II.4, left, shows the two relations $R = \{(1, q), (3, p), (3, s), (4, q)\}$ as well as $S = \{(p, y), (p, z), (q, x), (r, w)\}$. On the right then is seen the composition $S \circ R = \{(1, x), (3, y), (3, z), (4, x)\}$. Note that the fact that $3Rs$ or rSw do not contribute to the composition.

Composition is useful in that it can form genuinely new connections between objects.

(In the case of higher order relations (and relational databases), composition generalizes to an operation *join*.)

II.3 Properties of Relations

We shall focus, for a while, on binary relations on a set A (that is relations amongst elements of one set $A = B$). We first define a number of properties that such a relation might have:

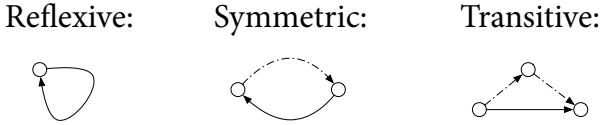


Figure II.5: Possible Properties of a Relation

Definition II.3: Let $\sim$ be a binary relation on a set A. Then $\sim$ is called

reflexive, if $a \sim a$ for every $a \in A$.

symmetric, if (for $a, b \in A$) $a \sim b$ implies that $b \sim a$.

antisymmetric, if (for $a, b \in A$) $a \sim b$ and $b \sim a$ together imply that $a = b$. (This means that, apart from a trivial case we might have either $a \sim b$ or $b \sim a$, but not both. Think of the $\leq$ relation on numbers.)

transitive , if (for $a, b, c \in A$) $a \sim b$ and $b \sim c$ imply that $a \sim c$.

In the digraph model (with one set, identifying A and B, a relation is reflexive if there is an arrow from every vertex to itself, it is symmetric, if for every arrow there is an arrow in the opposite direction, and it is transitive, if for every pair of arrows following each other, there is a "composite arrow". Figure II.5 illustrates this (the full lines are required, if the dashed lines exist).

We give some examples:

1. Let A be the set of rational numbers and R be ordinary equality. (That is

$$R = \{(a, b) \in \mathbb{Q} \times \mathbb{Q} \mid a = b\}.)$$

   This relation is reflexive, symmetric and transitive

2. Let A be the set of all people in a country with two people in relation if they have the same last name. This relation is reflexive, symmetric and transitive.

3. Let A be the set of all people living in the United States[3] with two people in relation if they live in the same state. Again this relation is reflexive, symmetric and transitive.

4. Let $A = \mathbb{R} \times \mathbb{R}$ the set of points in the plane, with two points in relation if they have the same x coordinate or the same y coordinate. (That is, one could draw a horizontal line or a vertical line through the two points.) We practice the set notation for relations by writing this down formally, noting that we have to describe pairs of pairs:

$$R = \{((x, y), (a, b)) \in A \times A \mid x = a \text{ or } y = b\}.$$

   This relation is reflexive and symmetric, but not transitive (go first horizontal, then vertical).

5. We slightly modify example 4 by requiring same x or same y coordinate, but not both the same. Then the relation is only symmetric, but not reflexive any more.

6. Let A be a set and $R = A \times A$ (i.e. all elements are in relation). This relation is reflexive, symmetric and transitive.

7. Let A be an arbitrary nonempty set and $R = \varnothing$ (that is no elements are in relation). Then R is symmetric and transitive (both conditions are true[4] since they are "if-then" with a condition that never can be fulfilled.), but not reflexive.

[3] Constitutional scholars should take the continental US without Washington DC here

[4] This situation of true statements is sometimes called *vacuously true*

8. Let $A = \mathbb{Q}$ with the usual "smaller or equal" relation $\leq$. This relation is reflexive, antisymmetric and transitive, but not symmetric.

9. Let $A = \mathbb{Q}$ with the "strictly smaller" relation to be smaller but not equal. This relation is antisymmetric (again vacuously as it is not possible that $a < b$ and $b < a$ for inequal a, b) and transitive, but not reflexive.

10. The relation $x^2 + y^3 = 1$ on $\mathbb{R}$ has none of these properties.

For the graph of a relation defined on (subsets) of $\mathbb{R}$, reflexivity (that is $(x, x) \in R$) means that the (increasing) diagonal through the origin must be part of the graph. Symmetry means that the graph is symmetric when reflecting along this diagonal. (Transitivity is somewhat more complicated and probably less helpful to visualize.)

II.4 Equivalence Relations, Equivalence Classes and Partitions

Restricting our focus even more, we now consider relations that could be used to represent a concept of equality. This is an idea you have used before easily. For example you probably would claim that $3 = 3$, even though the digits are of different size[5]. But they represent the same magnitude. Some of the examples in the last section share this characteristic, and we will characterize it with the properties we just defined:

Definition II.4: A binary relation $\sim$ on a set A is an *equivalence relation* if it is reflexive, symmetric and transitive.

Equivalence relations can be though of as a "less picky" version of *equality* that allow us to forget about differences between objects (say the color or size of (physical) numbers). Often one deliberately wants to consider formally different things as the same.

This concept of different levels of "being the same" occurs naturally in everyday life. For example, if we say that *all persons are equal*, we do not mean that they are identical (and that there is but one person in the world), but that they have the same natural rights and privileges.

For a more mathematical example, consider the expressions $1 + 1$ and 2, they are formally different objects (and a typesetter certainly will consider them as not the same. But if we consider them as expressing magnitudes, we say that $1 + 1 = 2$. This can be described as an equivalence relation on algebraic expressions.

An important characterization of equivalence relations is that they chop a set into parts, allowing for example for clustering large sets of data into a smaller number of cases. We shall investigate this next.

[5] On the other hand, if your business is in selling house numbers, you might consider them different, as you will be charging more for the larger digit.

Definition II.5: Let A be a set. A *partition* of A is a set P consisting of nonempty subsets $S \subset A$ (often called *cells*), such that:

1. No two different subsets share an element[6]:

$$\forall S, T \in P : (S \cap T \neq \varnothing \Rightarrow S = T).$$

2. Every element of A is in a subset:

$$\forall a \in A \exists S \in P : a \in S.$$

For example, we could take $A = \{1, 2, 3, 4, 5\}$ and

$$P = \{\{1, 3\}, \{2, 4, 5\}\}.$$

Note that P is not a subset of A, but if $S \in P$ then $S \subset A$.

Lemma II.6: Let P be a partition of A. Then every element of A is in exactly one $S \in P$.

Proof: By the second property it needs to be in at least one set, by the first property it cannot be in more than one. □

If $S \in P$ and $a \in S$, we call a a *representative* of S. It often is convenient to describe cells by giving such a representative. Typically representatives are not unique. Indeed, by the prior lemma, any element of a cell can serve as its representative.

We now want to show that partitions and equivalence relations are closely related.

Given a partition P of A, we define a relation on A by defining $a \sim b$ if and only if a and b are in the *same* set of the partition. In the example above, this would yield the relation

$$\begin{aligned} RP = \{&(1,1), (2,2), (3,3), (4,4), (5,5), \\ &(1,3), (3,1), (2,4), (4,2), (2,5), (5,2), (4,5), (5,4)\} \end{aligned}$$

Such a relation is clearly an equivalence relation: an element is in the same set as itself. If a and b are in the same set, so are b and a, and if a and b, as well as b and c are in the same set, this set contains a and c.

The same holds in reverse. For this, assume we are given an equivalence relation $\sim$ on a set A. For every $a \in A$ we define its *equivalence class* as the set of all elements in relation to a:

$$[a] = \{b \in A \mid b \sim a\}.$$

[6] This is a somewhat slick definition. You probably would have written ($S = T$ or $S \cap T = \varnothing$). Doings so would describe the same (logic), but the way we write it down here makes verifying the property less work to write.

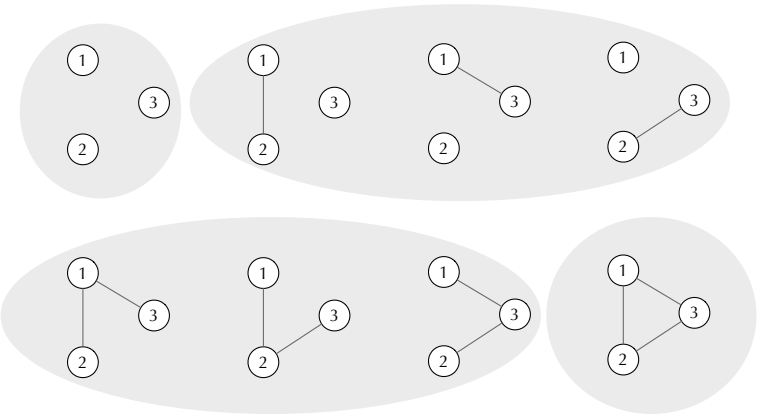


Figure II.6: The possible graphs on 3 vertices

Clearly $a \in [a]$ is in its own equivalence class, and equivalence classes thus are not empty, and every element of A lies in an equivalence class. We also note that if $a \sim b$, then $[a] = [b]$, since any $c \in [a]$ satisfies $c \sim a$ and thus $c \sim b$ by transitivity (and vice versa for $c \in [b]$). Thus, if two equivalence classes have non-empty intersection, they are equal: let $c \in [a] \cap [b]$, then $c \sim a$, $c \sim b$ (and thus $b \sim c$ because of symmetry) and by transitivity $b \sim a$ and thus $[a] = [b]$.

This means that the set of equivalence classes forms a partition of the set A. If a is the representative of a class S, then $S = \{b \in A \mid b \sim a\}$.

In the above example, the relation RP creates the equivalence classes

$$\{1, 3\} \text{ and } \{1, 4, 5\},$$

and thus the partition P.

For another, prototypical, example, suppose we define a relation $\sim$ on the set $\mathbb{Z}$ of integers by defining $a \sim b$ if 2 divides $a - b$. Then we get two (infinitely large) equivalence classes, namely the even and the odd numbers. If we change the equivalence to $a - b$ being divisible by 3, we get three (infinite) equivalence classes:

$$\{0, 3, -3, 6, -6, \ldots\}, \quad \{1, 4, -2, 7, -5, \ldots\}, \quad \{2, 5, -1, 8, -4, \ldots\}.$$

Equivalence classes will be important tools for us to construct new classes of objects, in particular arithmetic objects.

For a more elaborate example of equivalence classes, and why we care about them, consider graphs. A graph consists of *vertices* (basically points), together with *edges* that each connect two vertices.

To represent a graph, we need to represent the vertices – say for a graph on n vertices with the numbers $1, 2, \ldots, n$. An edge then is a set of two vertices. For

example (Figure II.6) consider graphs on $n = 3$ vertices. There are 3 potential edges and each edge can be selected or not, so we get $2^3 = 8$ possibilities for *labelled graphs*, that is graphs where the labeling of the vertices matters.

But often we do not care about this labeling but only about the possible connection patterns. We thus can define an equivalence relation (called *graph isomorphism*, this is an important, hard, algorithmic problem) as two graphs being equivalent if they become the same after a relabeling of the vertices. For example, the graph with the edge set $\{\{1,2\}\}$ can be transformed into the graph with edge set $\{\{2,3\}\}$ by relabeling[7] the vertices $\begin{matrix}1\to 2\\ 2\to 3\\ 3\to 1\end{matrix}$.

We thus get (for the example of three vertices) 4 equivalence classes (as indicated by shaded areas in the figure). These equivalence classes are what one considers typically as (unlabeled) graphs and the objects one would like to classify. On the other hand, when storing a graph on the computer, one needs to identify vertices in some way, which (implicitly) gives them labels. What is stored is thus a labeled graph as representative of its conjugacy class.

II.5 The Integers and the Rationals

In the remainder of this chapter we study a number of examples in which equivalence relations or equivalence classes are used to construct interesting objects.

The first example is to construct (all, including the negative) integers from the positive numbers (that correspond to counting objects). This requires more than adding a possible minus-sign, since we need that $-0 = 0$. Instead we form equivalence classes on pairs:

Let $\mathbb{N}_0 = \{0,1,2,3,\ldots\}$ be the set of nonnegative integers. We form the set of pairs

$$S = \mathbb{N}_0 \times \mathbb{N}_0 = \{(a,b) \mid a,b \in \mathbb{N}_0\}.$$

Our goal is to have the integer z to be represented by the set $\{(a,b) \in S \mid b-a = z\}$. To define this set as an equivalence class, we want to define a relation on S that (a,b) is related to (c,d), if $b-a = d-c$. Since this might involve negative numbers (which we are just constructing), we use instead: $b+c = a+d$:

$$(a,b) \sim (c,d) \Leftrightarrow b+c = a+d.$$

This relation is clearly reflexive (as $b+a = a+b$) and symmetric (if $b+c = a+d$ then $d+a = c+b$). For transitivity, observe that if $(a,b) \sim (c,d)$ then $b+c = a+d$, and if $(c,d) \sim (e,f)$ then $d+e = c+f$. We add f to the first equation and b to the

[7] This is not unique. Another option would be $\begin{matrix}1\to 3\\ 2\to 2\\ 3\to 1\end{matrix}$.

second and obtain

$$b + c + f = a + d + f \text{ and } b + d + e = b + c + f,$$

and thus $a + d + f = b + d + e$. But that implies that $a + f = b + e$, and thus $(a, b) \sim (e, f)$.

We then define arithmetic on the equivalence classes using representatives:

$$[(a, b)] + [(c, d)] = [(a + c, b + d)]$$

Formally, we need also to show that this definition is independent of the choice of representative, that is if $(a, b) \sim (e, f)$ then $[(a, b)] + [(c, d)] = [(e, b)] + [(c, d)]$, but we skip this somewhat technical step here.

Constructing the rational numbers

The rational numbers, $\mathbb{Q}$, similarly can be constructed as pairs of integers, representing fractions. We start with the set

$$S = \mathbb{Z} \times (\mathbb{Z} \setminus \{0\}) = \{(n, d) \mid n, d \in \mathbb{Z}, d \neq 0\}.$$

Now observe that $n/d = m/e$ if and only if $ne = md$. We thus define a relation

$$(a, b) \sim (c, d) \Leftrightarrow ad = bc.$$

As above, one shows easily that this is an equivalence relation. The details of this are left as an exercise for the reader. One then defines arithmetic operations on the equivalence classes that mimic the arithmetic rules for fractions.

II.6 Remainders and Modulo

For the last example, we choose a positive integer $m > 1$ and (similar to an example above) define a relation on $\mathbb{Z}$ by

$$a \sim b \Leftrightarrow m \mid (b - a)$$

using the vertical line $\mid$ as a shorthand for "divides". Dividing means that there exists a q (depending on $b - a$) such that $mq = b - a$. For example, if we had chosen $m = 7$ we would have $3 \sim 10$, but $3 \not\sim 5$.

Again, we show first that this defines an equivalence relation: Reflexivity holds as $m \cdot 0 = 0 = a - a$. Symmetry follows from the fact that if $mq = b - a$ then $m(-q) = a - b$. And transitivity from the fact that if $a \sim b$ there exists q such that $mq = b - a$ and if $b \sim c$ there exists r such that $mr = c - b$. But then

$$c - a = c - b + b - a = mr + mq = m(r + q)$$

and thus $m \mid (c - a)$, i.e. $a \sim c$.

We thus can form equivalence classes. To understand what these classes are, we make a number of observations:

If $a \in \mathbb{Z}$, we have that the class containing a is

$$[a] = \{a, a+m, a-m, a+2m, a-2m, a+3m, \ldots\} = \{a + k \cdot m \mid k \in \mathbb{Z}\},$$

since the elements equivalent to a are exactly those that differ from a by a multiple of m.

In every class C there is a non-negative integer For if $x \in C$ we also have that $x + m \in C$.

In every class C there is an element $r \in C$ with $0 \leq r \lneq m$ Take $r \geq 0$ to be the smallest nonnegative element of C. Then $r < m$, as otherwise $r - m \in C$.

This element r is unique, that is in the class C there cannot be two different elements $r_1, r_2 \in C$ with $0 \leq r_1, r_2 \lneq m$. Since if this was, we would have (without loss of generality[8] $r_1 < r_2$, but then $0 < r_2 - r_1$ and $m \mid r_2 - r_1 \lneq m$, which is impossible.

There are m equivalence classes, namely $[r]$ for $0 \leq r \lneq m$. This is since each number $0 \leq r \lneq m$ must be in an equivalence class, but no two of them are in the same class. And each class must contain such an r.

Part of these observations are summarized in the following theorem:

THEOREM II.7 (Division with remainder): For any integer $m > 1$ and any integer a, there exist unique $q, r \in \mathbb{Z}$ such that $a = qm + r$ with $0 \leq r \lneq m$.

For example, if we choose $m = 3$ there will be 3 equivalence classes, namely $[0] = [3]$, $[1]$, and $[2]$.

If the number m is chosen and fixed, sometimes the convention is used to denote this remainder r by $\overline{a}$.

Modular Arithmetic

We now define arithmetic on the equivalence classes by the following rules:

$$\begin{aligned} [a] + [b] &:= [a + b] \\ [a] \cdot [b] &:= [a \cdot b] \end{aligned}$$

To ensure there is no ambiguity in which representative we chose, we show that the result is the same, even if we chose different representatives. Recall, that the elements of $[a]$ are of the form $a + k \cdot m$ for $k \in \mathbb{Z}$. We thus need to show that the

[8] This is an example of a useful tool in proofs. We could have $r_1 < r_2$ or $r_2 > r_1$ and would need to consider both cases. But there is freedom in labeling the two numbers and we use this freedom to dictate that r_1 shall be the smaller of the two numbers. Such a choice does not conflict with any other requirement, and thus is permissible. This means that such a choice does not restrict the argument to a special case, but still is applicable to all situations.

results of addition and of multiplication yield the same results, even if we replace a by $a + k \cdot m$ and b by $b + l \cdot m$:

$$\begin{aligned} [a + km] + [b + lm] &= [(a + km + b + lm)] = [a + b + m(l + k)] = [a + b] \\ [a + km][b + lm] &= [(a + km)(b + lm)] = [ab + m(al + bk + klm)] = [ab] \end{aligned}$$

These results imply that the standard arithmetic rules we know for integers also hold for these operations on equivalence classes.

We have thus defined a new kind of arithmetic, involving addition, subtraction (from addition of the negative) and multiplication, on the m equivalence classes $[0], [1], \ldots, [m]$. It is called *modular arithmetic* (or *modulo* arithmetic). One often writes i mod m instead of $[i]$.

Using the $\bar{\cdot}$ notation for remainders, the same result can be interpreted as $\overline{a + b} = \overline{\overline{a} + \overline{b}}$ and $\overline{a \cdot b} = \overline{\overline{a} \cdot \overline{b}}$.

Modular arithmetic is particularly well suited for computers, because the set of objects is finite. Binary arithmetic on bits is just arithmetic modulo 2. And standard arithmetic on a 64-bit processor is simply arithmetic modulo 2^{64}. Modular arithmetic also has important applications in data transmission and information security. You will encounter it again in more advanced classes.

Examples of Modular Arithmetic

One way of describing modular arithmetic is by giving tables for addition and multiplication. We do this here, for example, for $m = 7$:

$+$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$
$\bar{1}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$	$\bar{0}$
$\bar{2}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$	$\bar{0}$	$\bar{1}$
$\bar{3}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{4}$	$\bar{4}$	$\bar{5}$	$\bar{6}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$
$\bar{5}$	$\bar{5}$	$\bar{6}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$
$\bar{6}$	$\bar{6}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$

$\cdot$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{6}$
$\bar{2}$	$\bar{0}$	$\bar{2}$	$\bar{4}$	$\bar{6}$	$\bar{1}$	$\bar{3}$	$\bar{5}$
$\bar{3}$	$\bar{0}$	$\bar{3}$	$\bar{6}$	$\bar{2}$	$\bar{5}$	$\bar{1}$	$\bar{4}$
$\bar{4}$	$\bar{0}$	$\bar{4}$	$\bar{1}$	$\bar{5}$	$\bar{2}$	$\bar{6}$	$\bar{3}$
$\bar{5}$	$\bar{0}$	$\bar{5}$	$\bar{3}$	$\bar{1}$	$\bar{6}$	$\bar{4}$	$\bar{2}$
$\bar{6}$	$\bar{0}$	$\bar{6}$	$\bar{5}$	$\bar{4}$	$\bar{3}$	$\bar{2}$	$\bar{1}$

These tables contain many interesting patterns. In the addition table, every row and every column contain every entry exactly once. The multiplication table (this is because m is prime, but we shall not prove this here) restricted to $\bar{1}$ to $\bar{6}$ has the same property. Having $\bar{1}$ in every row and every column means that we can invert every nonzero number, for example $\bar{1}/\bar{3} = \bar{5}$. This means that one can solve equations in modular arithmetic as one would do normally. We illustrate this in the following examples: Consider the equation

$$4x + 3 = 6 \pmod 7.$$

We subtract 3, getting $4x = 3 \pmod 7$. We then note from the multiplication table that $\overline{4} \cdot \overline{2} = \overline{1}$. Multiplying by 2 thus gives $x = 6 \pmod 7$ as solution.

We can handle systems of equations similarly:

$$\begin{aligned} \overline{3}x + \overline{5}y &= \overline{2} \\ x + \overline{2}y &= \overline{3} \end{aligned} \quad \Rightarrow \quad \begin{aligned} (\overline{5} - \overline{3} \cdot \overline{2})y = \overline{6}y &= \overline{2} - \overline{3} \cdot \overline{3} = \overline{0} \\ x + \overline{2}y &= \overline{3} \end{aligned}$$

$$\Rightarrow \quad \begin{aligned} y &= (\overline{1}/\overline{6}) \cdot \overline{0} = \overline{0} \\ x + \overline{2}y &= \overline{3} \end{aligned} \quad \Rightarrow \quad \begin{aligned} y &= \overline{0} \\ x &= \overline{3} \end{aligned}$$

Chapter

III

Functions

III.1 Functions

Functions are arguably at the heart of mathematics and are the most important definition of the whole course.

Definition III.1: A *function* $f\colon A \to B$ (for two sets A and B) is a relation $R \subset A \times B$ with the properties that

1. For every $a \in A$ there is a pair $(a, b) \in \mathbb{R}$: The domain of R equals the source A.
2. There are no two pairs $(a, b), (a, c) \in R$ that share the same first entry. (Formally: For all $a \in A$ and $b, c \in B$, if $(a, b) \in R$ and $(a, c) \in R$, then $b = c$.)

If this is the case, we can interpret the relation entry $(a, b) \in R$ (which will be the only one for a given a) as *assigning* the *value* $f(a) = b$ to the *argument* a. As with relations, we call

$$I = \{b \in B \mid \exists a \in A : (a, b) \in R\}$$

the *range* or *image* of f. See Figure III.1 for illustration.

In short, a function is a rule that assigns for every element of its domain a unique element of its range.

Sometimes the term *map* or *mapping* is used synonymously with *function*.

For relations on $\mathbb{Q}$ or $\mathbb{R}$ (or subsets thereof), the condition of being a function can be visualized in the graph if the relation by the property that no two points of the relation lie on a vertical line. In the examples in Figure II.3, only the relation b) is a function, the others are not.

What is important is to distinguish function f (though sometimes we will write $f(x)$ to denote the function), argument a, and value $f(a)$ of an argument as three

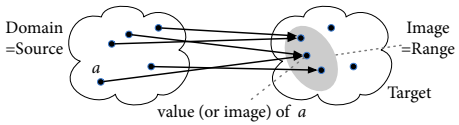


Figure III.1: A function

different (but related) entities. The argument is an input, the function a machine (or an algorithm), and the value the output.

Two functions are equal if they have the same domain, the same target, and (and this is the most important property) if they assign to every element of the domain the same value. To show that two functions f, g are equal, we need to compare their domains, their targets[1], and finally show that for **every** x in the domain we have that $f(x) = g(x)$. But if $f(a) = g(a)$ for just one argument a, this does not imply equality of the functions.

Thus the function $f\colon \mathbb{Z} \to \mathbb{Z}$, that maps every $x \in Z$ to $(-1)^x$ is equal to the function $g\colon \mathbb{Z} \to \mathbb{Z}$ that maps $x \in Z$ to 1, if x is even, and to -1 otherwise. The functions are equal, since they give the same values, even though the mechanism how they do so is different.

Describing Functions

Since functions are relations, they can be specified by text, formula, table, distinction, picture, to name just a few ways.

In most cases, the easiest way of describing a function is by using the interpretation of assigning values to elements of their domain. That is, we specify its domain, its target, and a rule (this can be a formula, or an algorithm) that specifies the value of the function for every element of the domain. This value must be an element of the target.

Concretely, we write the name of the function, a colon :, the domain, an arrow, and the target. If the value is given by a formula, often the notation $a \mapsto f(a) = \text{formula}(a)$ is used. In many cases, it is convenient to describe a function informally by text and makes for easier understanding. On the other hand, a formal description avoids ambiguity and often makes it easier to formally prove statements. And of course formal definitions are often easier to implement on a computer.

The following thus are all perfectly good definitions of functions:

1. Let $A = \{\text{frog}, \text{horse}, \text{sheep}\}$ and B the set of colors. We define $f_1\colon A \to B$ by

[1] In the context of functions, the target is sometimes called *codomain*.

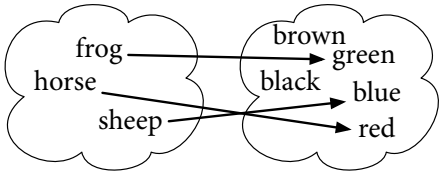


Figure III.2: A function assigning colors to animals

assigning to each animal its outer color:

$$f_1: \begin{cases} \text{frog} & \to & \text{green} \\ \text{horse} & \to & \text{brown} \\ \text{sheep} & \to & \text{black} \end{cases}.$$

2. Another function f_2 with same domain and target might assign to each animal its favorite color, say

$$f_2: \begin{cases} \text{frog} & \to & \text{green} \\ \text{horse} & \to & \text{red} \\ \text{sheep} & \to & \text{blue} \end{cases}.$$

This same function f_2 could be specified as

$$f_2: a \to \begin{cases} \text{green} & \text{if } a = \text{frog}, \\ \text{red} & \text{if } a = \text{horse}, \\ \text{blue} & \text{if } a = \text{sheep}. \end{cases}$$

Or someone might draw the function pictorially, as in Figure III.2.

3. Let A be the set of students at this university and B the set of symbol strings. The function $f_3: A \to B$ assigns to every student their (preferred) first name.

4. Let A, B as in example 3. The function $f_4: A \to B$ assigns to each student their email password. (It is a function, even though we cannot determine the password used by a particular student.)

5. Let D be the set of days of the current year and T the set of minutes in a day. We define the "sunrise time in Denver" function $f_5: D \to T$ as assigning to each date the time of sunrise (in Denver, CO).

6. Let S the set of students in this class and G the set of grades. Define the function $f_6: S \to G$ that assigns to each student the grade they will obtain in the class. (To be a function, it is only required that the value is unique and unambiguous, not that it is easily computable or known at the moment.)

7. For $A = \mathbb{N}$ the set of nonnegative integers, let $f_7 : A \to A$, $x \mapsto x+1$. Other examples would be $f_8 : A \to A$, $x \mapsto x$, or $f_9 : A \to A$, assigning to x the number of digits x requires in decimal representation.[2]

8. Let $f_{10} : \mathbb{R} \to \mathbb{R}$, $x \mapsto \frac{x^3+x+1}{x^2+1}$.

9. Left $f_{11} : \mathbb{R} \to \mathbb{R}$, $x \mapsto \begin{cases} 1 & x < 0 \\ -2 & x = 0 \\ x+17 & 0 < x < 1 \\ \cos(x) & x \geq 1 \end{cases}$.

10. Let $g : \mathbb{N} \to \mathbb{Q}$ to give for each n the first n decimal places of π. Thus $g(0) = 3$, $g(1) = 3.1$, $g(2) = 3.14$ and so on. If we go back to the notation of relations, we get

$$\begin{aligned} g &= \{(n, \pi \text{ to } n \text{ decimal digits}\} \\ &= \{(0,3), (1,3.1), (2,3.14), (3,3.141), (4,3.1415), \ldots\}. \end{aligned}$$

On the other hand, the following attempts do not define functions (for reasons indicated).

1. The relation $R = \{(a^2, a) \mid a \in \mathbb{Q}\} \subset \mathbb{Q} \times \mathbb{Q}\}$ is not a function, *since there are multiple elements in R – e.g. $(4,2)$ and $(4,-2)$ with the same first entry, so images are not uniquely defined.*

2. Let $f : \mathbb{Q} \to \mathbb{Z}$ assigning to every rational number the closest integer. *Not a function, as the closest integer is ambiguous, say for $1/2$.* We can fix it, by defining an explicit tie-break rule for ambiguous cases. (What is often used in practice is to take the largest integer $\leq x + \frac{1}{2}$.)

3. Let $f : \mathbb{Q} \to \mathbb{Q}$, $x \mapsto \frac{x^2-1}{x-1}$. *This is not a function, as the denominator at $x = 1$ becomes zero.* One could fix it by changing the domain to $\mathbb{Q} \setminus \{1\}$, or by replacing it with the function $x \mapsto x+1$ (which for all $x \neq 1$ returns the same value).

4. Let $f : \mathbb{Q} \to \mathbb{Q}$, $x \mapsto \sqrt{x}$. *Not a function, as values, such as $\sqrt{2}$ are not rational, and for negative x are not defined in $\mathbb{Q}$.* We can fix this by changing the domain to $\mathbb{Q}_{\geq 0} = \{a \in \mathbb{Q} \mid a \geq 0\}$ and the target to $\mathbb{R}$.

5. Assigning to every subset of the rational numbers its smallest element. *Not a function, since e.g. the set $\{a \in \mathbb{Q} \mid a > 0\}$ has no smallest element.* (We could fix this using the concept of limits↝ IV.4.)

[2] You might note that we could alternatively write $f_9(x) = \lfloor \log_{10}(x) \rfloor + 1$, using the $\lfloor \cdot \rfloor$ notation for truncation.

Algorithms as functions

Many computer algorithms can be considered as functions, that take an input and based on this produce a definite output. For example, consider the algorithm which takes as input a natural number n, and outputs the first letter of n^2 written in English words. In pseudocode:

```
define: func(n)
input: a number n
output: a letter of the English alphabet
procedure:
n = n*n \\ square the input
n = toWords(n) \\ write n in English words
n = firstLetter(n) \\ choose the first letter of n
return n
```

For $func(n)$ to truly describe a function, we must assume some unique naming convention for numbers in this program to guarantee that the function has a well-defined value for each n. That is, we need the subroutine `toWords`(n) to have a determined output for all n. When this is the case, our algorithm defines a proper mathematical function `func` $: \mathbb{N} \to \mathcal{A}$ where $\mathcal{A}$ is the set of English letters. For example, to calculate `func`(9) we square 9 to 81, write 81 as *eighty-one*, and finally return the first letter e as our output: `func`$(9) = e$.

Investigating the properties of a function that is given as an algorithm is often harder than if we have a conceptual description or a formula.

However not every algorithm is a function. If the algorithm uses side-effects (accessing information that is not part of the input), or uses random numbers, it is not a function, even if input and output are otherwise well defined:

```
define: notafunc(n)
input: a positive integer n
output: a random number between 1 and n
procedure:
1. r = randomInteger(1,n) \\ random integer r between 1 and n
2. return r
```

A key difference between `func`(n) and `notafunc`(n) is the introduction of randomness in the former, which even in the case of pseudo-random numbers depends on information that is not part of the function's argument. This causes `notafunc`(n) to not have a well-defined output. For example, a call of `notafunc`(3) might give the output 2, while a second call of `notafunc`(3) could ght give 1, depending on the result of calling $randomInteger(1, 3)$. In general, as long as an algorithm does not refer to global variables that are not part of the arguments, we can think of it as a mathematical function and ask whether it is onto or one-to-one.

III.2 Some Basic Functions

It will be helpful to have a number of examples of functions at hand:

First, take a set A and let $B = A$. The *identity function* on A is $\mathrm{id}\colon A \to A$, $a \mapsto a$ the function that maps every element to the same image.

Let A, B be sets and $b \in B$ some element. A *constant function* (with value b) is the function $A \to B$, $a \mapsto b$. If $A = B = \mathbb{R}$, the graph of a constant function is a horizontal line.

Let A be a set and $S \subset A$. We let $B = \{0,1\} \subset \mathbb{Q}$. The *characteristic function* for S is

$$\chi_S\colon A \to B, a \mapsto \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases}$$

A use of such a function is to "crop" another function, by multiplying with χ_S, for the product to be zero outside S.

Polynomials and Rational Functions

Many interesting function on numbers (i.e. $A \subset \mathbb{R}$) are given by prescribing the image of $x \in A$ by a formula. The easiest version are formulas that only involve addition and multiplication (including powers of x). Such a function is called a polynomial function.

For example $f\colon \mathbb{Q} \to \mathbb{Q}$, $x \mapsto x^5 - 3x^2 + 17x + 3$ is such a function. The highest power of x that arises (in the example: 5) is called the *degree* of the polynomial, and the factors in front of powers of x are called *coefficients* of the polynomial. We can describe the general case of a polynomial as

$$c_d \cdot x^d + c_{d-1} \cdot x^{d-1} + \cdots + c_2 \cdot x^2 + c_1 \cdot x + c_0 = \sum_{i=0}^{d} c_i x^i$$

where the $c_i \in \mathbb{Q}$ are the coefficients.

If f and g are both polynomials, and $A \subset \mathbb{R}$ such that $g(a) \neq 0$ for all $a \in A$, we can also take the quotient $f(x)/g(x)$ as a new function. It is called a *rational function*.

III.3 Function Arithmetic and Shifts

We want to investigate what happens to a function (and the graph of a function) under arithmetic operations. This also can be useful to modify a given function in a desired way.

We illustrate this, using an example (Figure III.3, we transform the function given under a)): If we add a constant to f, i.e. replace $f(x)$ by $f(x) + k$ we add k to

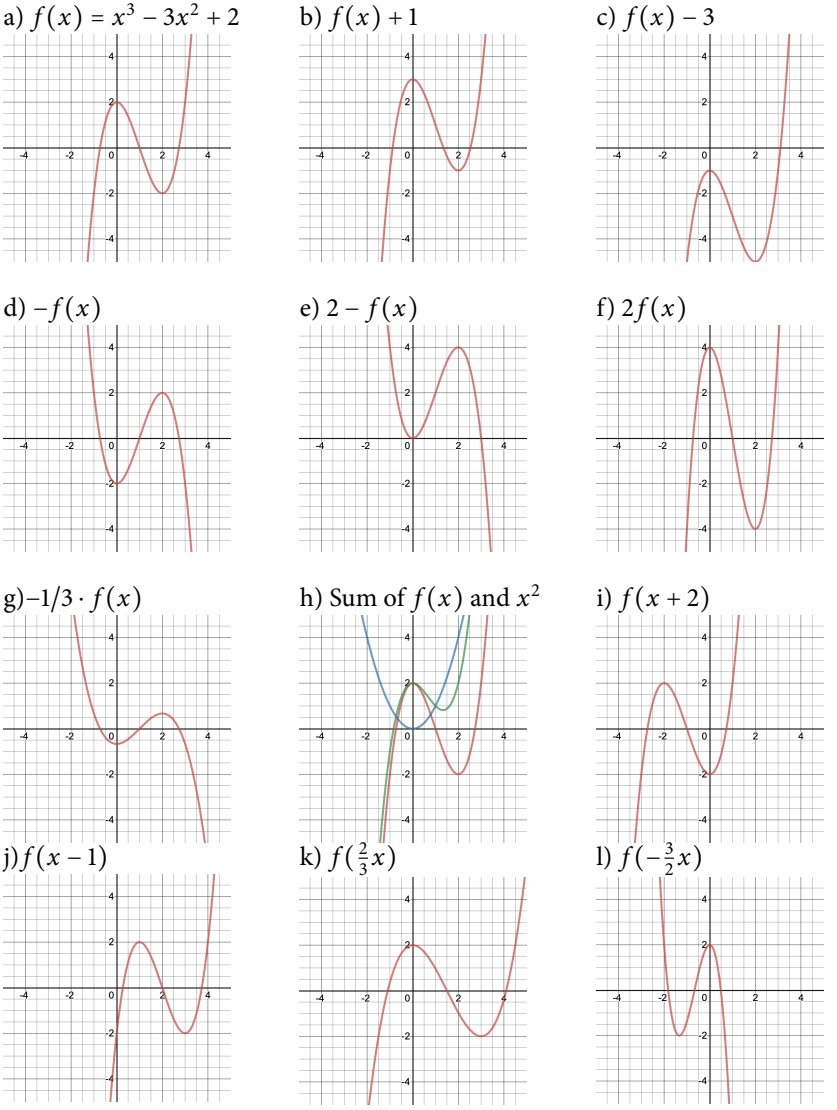


Figure III.3: Transformations of a function

all function values. This means the graph of f shifts up by k units b), respectively down c) if k is negative.

If instead we subtract f from a constant d) and e)) we get the graph of f, flipped upside down and shifted vertically.

Multiplying by a constant will stretch f) the graph vertically by this factor, respectively compress if the factor is (of absolute value) < 1 and flip if negative g).

More generally, we can take the sum of two functions, which is the function $f + g$ that assigns to each x the value $f(x) + g(x)$. See figure h) for $g(x) = x^2$ in blue and $f(x) + x^2$ in green.

So far, all transformations were vertical. To transform horizontally, we need to modify the argument x instead of the value $f(x)$: First consider if we replace x by $x + k$. Then $f(x + k)$ assigns to a point a the value $f(a + k)$ that f has at $a + k$, that is k units *to the right* of a. The effect is to shift the graph *to the left* by k units (i) for positive k, respectively to the right for negative k (j). Similarly, multiplying x by a factor k will assign to a the value $f(k \cdot a)$ that f has at k times a. Thus for $0 < |k| < 1$ the graph will be stretched horizontally (k), for $|k| > 1$ it will be compressed horizontally. And a negative k flips left and right (l).

In summary, adding to $f(x)$ or multiplying by a number will transform the graph vertically as one would expect. Changing the argument transforms the graph horizontally, but *in reverse* of what the same transformation would do vertically.

Composition

When, in the last section, we modified the argument x of a function f we actually used a special case of the composition of functions, composing f with a function $x \mapsto x + k$ or $x \mapsto kx$.

Since functions are relations, the composition of functions is just a special case of the composition of relations. Suppose $f\colon A \to B$ and $g\colon B \to C$ are functions. Then, as relations, we have that

$$f = \{(x, f(x) \mid x \in A\} \qquad \text{and} \qquad g = \{(y, g(y) \mid y \in B\}.$$

By the definition of composition of relations, we thus get the composition $g \circ f \subset A \times C$ as

$$\begin{aligned} g \circ f &= \{(a, c) \mid \exists b \in B : (a, b) \in f \text{ and } (b, c) \in g\} \\ &= \{(a, c) \mid (b, c) \in g \text{ for } (a, b) \in f\} \\ &= \{(a, c) \mid c = g(b) \text{ for } b = f(a)\} \\ &= \{(a, g(f(a))) \mid a \in A\}. \end{aligned}$$

This means that $g \circ f$ is again a function (exercise: verify this) from A to C, mapping $x \mapsto g(f(x))$. This is also the justification for the notation $g \circ f$, as we have that $(g \circ f)(x) = g(f(x))$.

Note that usually (even if $A = B = C$) $g \circ f \neq f \circ g$. For example, if $f\colon \mathbb{Q} \to \mathbb{Q}$, $x \mapsto x + 1$ and $g\colon \mathbb{Q} \to \mathbb{Q}$, $x \mapsto 2x$, then $g \circ f\colon x \mapsto 2(x + 1) = 2x + 2$, while $f \circ g\colon x \mapsto (2x) + 1$.

Unless one of the functions is relatively basic, that is only adding constants or multiplying with a scaling factor as we did in the previous section [3], the effect of composition on the graph of a function is more complicated than can be described by easy rules.

III.4 Properties of Functions

When we looked at relations, we also had two other operations. Since functions are a special case of relations, they also apply here, but are not always of interest: Taking the complement of a function will almost never be a function. We will study the question of whether the converse of a function is a function (and not just a relation) in this section.

The definition of a function had two requirements. We will define two properties that a function $f\colon A \to B$ might have. These properties correspond to the two requirements holding for the converse relation.

Fundamentally, the relation form of f is $\{(a, f(a)) \mid a \in A\}$ and the converse thus is $\{(f(a), a) \mid a \in A\}$. We want to write this instead in the form $\{(b, \text{something}) \mid b \in B\}$. This means that testing for the properties is related to solving $b = f(a)$ for a.

Onto

A function is onto if it takes every possible value, that is if its image equals its target. For example, we could take a function that maps the visitors of a hotel to the hotel's guest rooms. This function is onto whenever the hotel is booked out.

Formally:

Definition III.2: A function $f\colon A \to B$ is called *onto* (some books instead use the posher term *surjective*), if every element of B is obtained as an image, that is for every $b \in B$ there is an $a \in A$ such that $f(a) = b$.

The left two images in Figure III.4 illustrate the concept.

If a function is given by a graph, it is onto if the projection of the graph on the y-axis is all of B, equivalently that every horizontal line intersects the graph. For example (see Figure III.5), the function $f_1\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto x^3$ is onto, while the function $f_2\colon x \mapsto x^2$ is not onto (it takes only positive values).

Similarly, $f_3\colon x \mapsto x^3 - 3x^2 + 2$ is onto, but $f_4 = \frac{1}{10}\exp(x)$ is not.

Note that the target defined for the function is crucial for a decision on whether it is onto. If we change f_2 to a function $g_2\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^2$ (i.e. same rule, just changed the target), then g_2 is onto.

[3] these are called *linear* functions

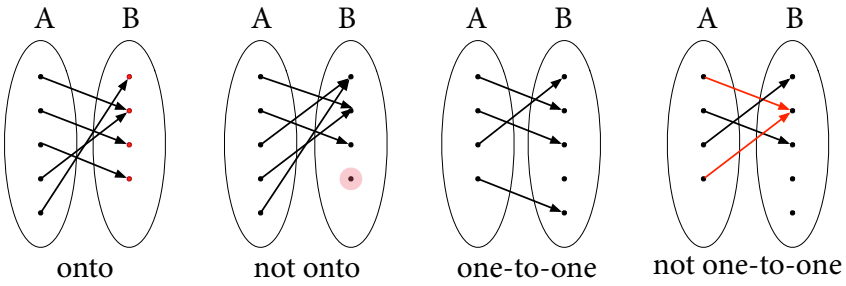


Figure III.4: Onto and One-To-One

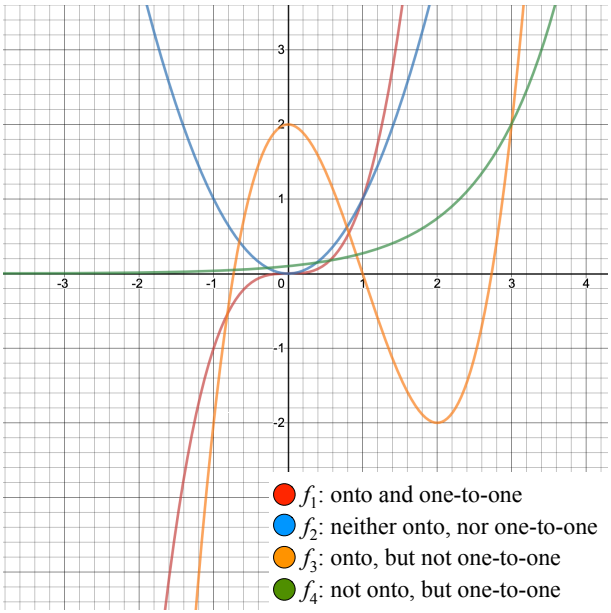


Figure III.5: Onto and One-To-One for functions $\mathbb{R} \to \mathbb{R}$

To show algebraically that a function is onto, we need to show that every element $b \in B$ in the target is obtained as an image. The easiest way of doing so is to find an explicit $a \in A$ such that $f(a) = b$. We might be able to do so (but are not guaranteed) by solving for a.

For example, for function f_1 we have that $a = \sqrt[3]{b}$ is mapped to b. On the other

hand f_3 is also onto, but it is hard to solve for a, and for $-2 \leq b \leq 2$ there are multiple possible a.

A function being onto guarantees that the source of the converse relation equals its domain, i.e. it has pairs $(b, *)$ for every $b \in B$.

One-to-One

In many situations of real-world functions $f: A \to B$, i.e. when assigning objects in B to objects in A, the intention is to use the assigned object in B in place of the original object in A. That is, there is an assumption that for a given $b \in B$ we can look-up the unique $a \in A$ for which $f(a) = b$. Examples of this are Social Security numbers, student ID numbers, car number plates[4], or bar codes on supermarket items.

But this is not always guaranteed. If we map people to names, or dates of birth there will be typically multiple matches, even in a limited population[5]. It is thus an interesting property, and in fact the one that ensures the second property for functions for a converse.

Definition III.3: A function $f: A \to B$ is called *one-to-one* (some books instead use the posher term *injective*), if there are no two different elements $a_1, a_2 \in A$, $a_1 \neq a_2$ such that $f(a_1) = f(a_2)$. In other words, if $f(a_1) = f(a_2)$ then $a_1 = a_2$.

The right two images in Figure III.4 illustrate the concept. In the above example of mapping hotel visitors to bedrooms, the function is one-to-one whenever every bedroom is single-occupancy (or empty).

If a function is given by a graph, being one-to-one means that every *horizontal* line intersects the graph in at most one point. Thus, in the functions in Figure III.5, we have that f_1 and f_4 are one-to-one, but the other two are not.

Again, being one-to-one does not only depend on the rule of the function, but its declaration. It might be possible to restrict the domain A to a subset $A_1 \subset A$ and get a new function that is one-to-one. In the examples, both f_2 and f_3 would become one-to-one, if we restricted their domain to $\{x \in \mathbb{R} \mid x > 2\}$.

To test algebraically for a function being one-to-one requires that one can solve uniquely for a with $f(a) = b$. This is for example the case for f_1 (as $\sqrt[3]{b}$ is unique in the real numbers), but it is not true for f_2 (since $f_2(-1) = f_2(1)$).

Doing so can be hard, we will see that calculus will provide better tools for testing one-to-one ⇝ VI.1.

An important application of the concept of one-to-one in computer science is that of *hash functions*. In short, a hash function takes a digital object (stored information, or a binary file) and maps it to a large integer. The hope is that the

[4] That is the reason for a number plate such as `COOLGUY14` – the number 14 is not special, but 1–13 were used already.

[5] This is of course the reason why ID numbers were invented in the first place.

function is one-to-one[6], that is the hash value allows to identify the data/file and for example to use the hash to verify that it has not been tampered with but is the original file.

While a one-to-one function allows in principle to identify the original argument a from a function value $f(a)$ this may be hard (or impossible in practice) to do for hash functions. Indeed part of the mechanism underlying bitcoin and other digital currencies is that such a look-up seems to be hard in practice and can only be done by trying out different values of a and checking when the desired value $f(a)$ is reached.

We close with a connection between the concepts of onto and one-to-one in the case of finite sets:

LEMMA III.4: If A and B are finite sets with $|A| = |B|$, then a function $f: A \to B$ is one-to-one, if and only if it is onto.

Proof: We set $n = |A| = |B|$. There are two directions to show. First assume that $f: A \to B$ is onto. If f was not one-to-one, there would be two different elements, $a_1, a_2 \in A$, such that $f(a_1) = f(a_2)$. But then f would have to have strictly fewer than n values in B and thus could not be onto.
Vice versa, assume that $f: A \to B$ is one-to-one. If it was not onto, it would have to have strictly fewer than n values, which means that not all n elements of A could have different images. □

Note that this lemma is false if the sets are infinite. The functions f_3 and f_4 in Figure III.5 are counterexamples.

III.5 Bijections and Inverse Functions

A function $f: A \to B$ that is both one-to-one and onto is called *bijective* or a *bijection*. Being bijective means that for every $b \in B$ there is a unique $a \in A$ such that $f(a) = b$.

The converse property, i.e. that for every $a \in A$ there is a unique $b \in B$, is the definition of a function. This means that the converse relation to f, namely

$$\{(f(a), a) \mid a \in A\}$$

is a function. We call it the inverse function to f, and call it f^{-1}. It is a function $B \to A$ that maps b to the unique a such that $f(a) = b$. Note that this $\cdot^{-1}$ in the name is a pure formalism due to a limited set of symbols available. It should **not** be confused with $1/f$, which is another function entirely[7]. In the example of assigning

[6] on the set of plausible inputs. That is not all possible binary files, but – say – binary files that would be valid programs. Often it is impossible to prove that such a hash function is one-to-one, but the property is only observed through examples or in approximation.

[7] Formally, $1/f$ is the inverse under pointwise multiplication $(f \cdot g)(x) = f(x) \cdot g(x)$, while f^{-1} is the inverse under composition of functions.

people to hotel rooms, if the function is bijective, its the inverse function assigns to a room its occupant.

By the definition, the relation form of the inverse function,

$$\{(b, f^{-1}(b)) \mid b \in B\},$$

is the converse relation to f. We also note that $f^{-1} \circ f: A \to A$, $a \mapsto f^{-1}(f(a)) = a$ is the identity on A, and that $f \circ f^{-1}: B \to B$, $b = f(a) \mapsto f(f^{-1}(b)) = f(a) = b$ is the identity on B.

If f is given by a formula for $f(a)$, one might be able to get a formula from solving $f(a)$ for a. For example, we have seen that the function $f_1: \mathbb{R} \to \mathbb{R}$, $x \to x^3$ is onto and one-to-one, and its inverse will be the function $f_1^{-1}: \mathbb{R} \to \mathbb{R}$, $x \mapsto \sqrt[3]{x}$. But the function $f_5: \mathbb{R} \to \mathbb{R}$, $x \mapsto x^3 + x$ is also one-to-one and onto, but we will be hard pressed to solve $a^3 + a = b$ for a.

III.6 Counting and Cardinality

One use of bijective functions is how the cardinality of a set is actually defined. Formally, we define the cardinality of a set A as the unique n, such that there is a bijective function from A to the set $\{1, 2, \ldots, n\}$. (Formally, one has to show that this n will be always the same, regardless of the function chosen, but we will skip this step here.) This function can be interpreted as numbering the elements when counting.

We give formulas for the cardinality of some of the set constructions. In the following, let A be a set with $|A| = n$ and B a set with $|B| = m$:

Cartesian Product Once we numbered the elements of A and B, we can consider the pairs in the Cartesian product as coordinates in an $n \times m$ rectangle. Thus $|A \times B| = n \cdot m$.

Power Set Every subset of A can be described by bit-string (a string of 0's and 1's) of length n that indicates for each element of A whether it is contained in the subset. These bit strings lie in the n-fold Cartesian product $\{0,1\} \times \{0,1\} \times \cdots \times \{0,1\}$, and thus there are 2^n such strings. Thus $|\mathcal{P}(A)| = 2^n$.

We note, without proof, that the number of subsets of fixed size k is given by the *binomial coefficient*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

(where $n! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot n$ is called *n factorial*)

Relations Every relation is a subset of $A \times B$. Thus the number of relations between A and B is $|\mathcal{P}(A \times B)| = 2^{n \cdot m}$.

Functions We can consider a function from A to B as an n-tuple that gives in position i the value of the i-th element of A. The formula for the order of a Cartesian product thus gives $|B \times \cdots \times B| = m^n$ different functions.

One-to-One Functions To ensure a function $A \to B$ is one-to-one, the image of the second element cannot be chosen freely, but needs to be different than the image of the first element. Thus we do not have m choices for the image of the second element, but only $m - 1$ choices. And for the third element there are $m - 2$ possible images and so on until the n-th element has $m - n + 1$ possible images. Thus there are

$$m \cdot (m-1) \cdot (m-2) \cdot \cdots \cdot (m-n+1) = m!/(m-n)!$$

one-to-one functions. Note that this product has a factor 0 (and thus becomes zero itself) if $n > m$.

Bijective Functions Note that such functions can only exist if $n = m$, but then, by Lemma III.4, it is sufficient for the function to be one-to-one. Thus there are[8] $n!/(n-n)! = n!/1 = n!$ bijective function if $m = n$ (and zero otherwise).

Onto Functions The formula for the number of onto functions is more complicated than could plausibly be explained here. It is:

$$\sum_{i=1}^{m} (-1)^{m-i} \binom{m}{i} i^n.$$

[8] We have that $0! = 1$ as product over an empty set, in the same way a sum over an empty set is 0.

Chapter

IV Sequences and Series

We now start our study of functions, and how functions change and behave "in the big picture". For this, we first consider functions on integers.

IV.1 Sequences

Definition IV.1: A *sequence* is a function $a\colon \mathbb{N}_0 \to \mathbb{R}$, defined on the set $\mathbb{N}_0 = \{z \in Z \mid z \geq 0\}$. If $i \in \mathbb{N}_0$ and $a(i)$ is a value, we also call i the *index* or the *position* in the sequence. Sometimes we ignore index 0 and only consider sequences indexed by strictly positive integers.

The word "index" (plural: *indices*)is the reason for denoting the argument by i (and thus ultimately the reason why "i" is the standard name for a loop variable when programming).

It often is convenient to write indices as subscripts and to write a_i in place of $a(i)$. We will write (a_i) to designate a sequence (that is the set of all values) in contrast to a single value a_i.

Since sequences are a special case of functions, we can describe sequences in the same way as functions. Often this is done by prescribing a rule, for example: $a_i = i + 1$ for the sequence

$$a_0 = 1, \qquad a_1 = 2, \qquad a_2 = 3, \ldots$$

Other examples would be the sequences $b_i = i^2 + 5$, $c_i = \sum_{j=0}^{i} j$ =*the sum of the numbers from* 1 *to* i, or d_i =*the* i*-th prime number*.

With indices arranged easily as $1, 2, 3, \ldots$, it also can be tempting to write a sequence by giving its first few values, and expecting the reader to identify and

continue the pattern. For example

$$\begin{aligned} a_i &: 1, 3, 5, 7, 9, 1, \ldots \\ b_i &: 10, 9, 8, 7, \ldots \\ c_i &: 3, 3.1, 3.14, 3.141, 3.1415, \ldots \end{aligned}$$

While this seems convenient, guessing the right pattern can be difficult, and is ambiguous. Indeed, despite its use in popular intelligence tests that ask the reader to identify the rule for a given set of values, it is impossible to decide what the "correct" rule should be. For example, consider the sequence

$$2, 4, 6, 8, 10, \ldots$$

which one might guess is given (assume we start indexing at 1) by the rule $a_i = 2i$ and has the next value 12. But one can also argue that the next value is −17, based on the rule

$$a_i = -\frac{29}{120}x^5 + \frac{29}{8}x^4 - \frac{493}{24}x^3 + \frac{435}{8}x^2 - \frac{3853}{60}x + 29.$$

Such problems therefore only make sense if we ask for the *simplest possible* answer (however we might want to measure it), which immediately makes it a far more difficult problem than we would want to consider here.

As for functions, we also could consider plotting the values of a sequence, but since they are only defined for integral arguments the result will be disconnected dots.

Finally, a sequence might be a sequence of data values (e.g. over time), which we know in part and would like to predict in the future.

IV.2 Recursion

Sequences are important in computer science, in that they can arise when indicating the cost (that is number of calculation steps required) of an algorithm, for an input of a given size i. In this context, sequences often are given through a *recursion*. That is, we give one (or several) initial values of the sequence, and then give a formula that determines further values from the previous ones. For example:

$$a_1 = 3, \qquad a_i = a_{i-1} + 2 \text{ for } i > 1$$

gives us the sequence $a_i = 1 + 2i$, while

$$s_1 = 1, \qquad s_i = s_{i-1} + i \text{ for } i > 1$$

gives the sequence, whose i-th entry is the sum of integers from 1 to i. It is also possible to instead give values for new indices bigger than i, as in the example

$$s_{i+1} = s_i + (i + 1).$$

Recursions might refer to multiple prior values (and then one might need to define multiple start values). The prototypical example of this is the *Fibonacci numbers*, defined as:

$$f_0 = 0, \qquad f_1 = 1, \qquad f_{n+2} = f_{n+1} + f_n.$$

We can determine arbitrary values f_i by following through the recursion and evaluate terms one by one. (This also shows that the values of the sequence are uniquely determined.) Here for example:

$$f_2 = f_0 + f_1 = 0+1 = 1, \quad f_3 = f_1 + f_2 = 1+1 = 2, \quad f_4 = 1+2 = 3, \quad f_5 = 2+3 = 5, \ldots.$$

It still can be desirable, for a sequence given by a recursion, to determine a closed (i.e. a direct value that requires no loop when evaluating) formula for its values. Doing so, however often requires tools that are significantly beyond this course (but see ↝ VII.3). Instead we shall illustrate how one can verify a closed formula for a recursively defined sequence. We do so with the above example of the "sum of the first i integers" and the recursion $s_1 = 1$, and $s_i = s_{i-1} + i$. We claim that the formula $s_i = \frac{i(i+1)}{2}$ satisfies the recursion. To show that this indeed is the case, we first check the base cases:

$$s_i = \frac{1(1+1)}{2} = 1.$$

We then assume that the index is large enough for the recursive formula to hold (in this example: $i > 1$) and verify that the formula satisfies the recursion. That is we evaluate the right hand side of the recursion when plugging in the formula:

$$\begin{aligned} s_{i-1} + i &= \frac{(i-1)((i-1)+1)}{2} + i = \frac{(i-1)i}{2} + i \\ &= \frac{i^2 - i + 2i}{2} = \frac{i^2 + i}{2} = \frac{i(i+1)}{2} \end{aligned}$$

and compare with the formula for the left hand side $s_i = \frac{i(i+1)}{2}$. (One could analogously use the alternative recursion formula $s_{i+1} = s_i + (i+1)$ instead, and then would have to simplify the left hand side when evaluating $s_{i+1} = \frac{(i+1)(i+1+1)}{2}$.)

Application: estimating the cost of a program

An important application of sequences and recursion is in estimating the number of steps a recursive algorithm will take to completion. Here the sequence value at position i is defined to be the number (or a bound thereof) of steps taken by the algorithm for an input of size n.

Consider, for example, the following (very naive) insertion-sort algorithm. To sort a list L consisting of n numbers, we first sort (by a recursive call) the first $n-1$ numbers and then insert the n-th element in the right position in the list.

```
define: sort(L,n)
input: L, a list of integers, to be sorted
       n, up to which position to sort
procedure:
if (n > 1) then # Otherwise no need to sort
  sort(L,n-1); # recursively sort the first n-1 elements
  a:=L[n];
  p:=1; # find position of first entry larger than a
  while p<n and L[p]<=a do
    p:=p+1;
  od;
  for j from n-1 downto p do # shift entries up
    L[j+1]:=L[j];            # to make space at p
  od;
  L[p]:=a;
fi;
```

For example, if $L = [5, 13, 12]$, how does `sort(L,3)` get processed? We call `sort(L,2)`, which in turn calls `sort(L,1)`, which does nothing. Exiting back to $n = 2$ we set $a = L[2] = 13$ and (since $L[1] = 5 < 13$) get $p = 2$, keeping 13 in the second position. Exiting the recursion and getting back to $n = 3$, we find (again) that $p = 2$, and shift the 13 into position 3, inserting 12 into position 2.

We want to know how many comparisons of elements our algorithm might have to do. (Such comparisons only happen in the `while` loop condition.) For this, we define a sequence cost_n, that is an upper bound on the number of comparisons required by this algorithm to sort a list of length n.

Recall that a recursive formula for a sequence is given by a collection of initial values and a formula which determines the value of the sequence at n based on previous terms of the sequence. Our initial value will be cost_1. This is the number of comparisons needed to sort a list of length one, which is 0, since we have nothing to check. Thus $\text{cost}_1 = 0$.

For $n > 1$, the algorithm calls itself recursively for length $n - 1$ (at cost cost_{n-1}, and then compares the last element a to potentially all[1] $n - 1$ elements in the list. The cost thus satisfies the recursion

$$\text{cost}_n = \text{cost}_{n-1} + (n - 1).$$

For example, $\text{cost}_2 = \text{cost}_1 + 1 = 0 + 1 = 1$ and $\text{cost}_3 = \text{cost}_2 + 2 = 1 + 2 = 3$.

By 'peeling' back the layers of the recursion, we could get a formula (this is a

[1] The clever reader might already think on how to improve the algorithm by using *binary search* to find the correct position in fewer steps.

little bit beyond the expectations for this course) for cost_n as:

$$\begin{aligned}
\text{cost}_n &= \text{cost}_{n-1} + (n-1) \\
&= \text{cost}_{n-2} + (n-2) + (n-1) \\
&= \text{cost}_{n-3} + (n-3) + (n-2) + (n-1) \\
&= \ldots \\
&= 1 + 2 + \ldots + (n-3) + (n-2) + (n-1) \\
&= \sum_{k=1}^{n-1} k \\
&= \frac{n(n-1)}{2}.
\end{aligned}$$

In any case, (even without the arithmetic done in the previous paragraph), the reader should be able to use the same tools as above to verify that the recursion

$$\text{cost}_1 = 0; \text{cost}_n = \text{cost}_{n-1} + (n-1).$$

is satisfied by the formula

$$\text{cost}_n = \frac{n(n-1)}{2}.$$

We have thus *proven* that the (not very good) algorithm given has a worst case requirement of $\frac{n(n-1)}{2}$ comparisons to sort a list of length n.

IV.3 Monotonous and Bounded Sequences

We are interested in investigating the long-term (that is, values for large indices) behavior of sequences. For this we define the following properties:

Definition IV.2: Let (a_i) be a sequence. Then this sequence is

monotonically increasing , if $a_{i+1} \geq a_i$ for all i.

strictly monotonically increasing , if $a_{i+1} > a_i$ ($a_{i+1} \neq a_i$) for all i.

bounded from above , if there exists a number $B \in \mathbb{R}$, such that $a_i \leq B$ for all i.

We define *monotonically decreasing*, *strictly monotonically decreasing*, and *bounded from below* in the obvious way by reversing the inequalities.

For example, consider the sequences

$$\begin{aligned}
a_i &= 5 - \frac{1}{i}, \\
b_i &= 5 + i, \\
c_i &= 5 + (-1)^i.
\end{aligned}$$

Then the sequences (a_i) and (b_i) are (strictly) monotonically increasing, since

$$a_{i+1} = 5 - \frac{1}{i+1} > 5 - \frac{1}{i} = a_i \text{ and } b_{i+1} = 5 + (i+1) > 5 + i = b_i.$$

The sequence (c_i) is not monotonically increasing, as

$$c_2 = 5 + (-1)^2 = 6 \not\leq 4 = 5 + (-1)^3 = c_3.$$

Similarly, we have that (a_i) and (c_i) are both bounded from above, since we have that $a_i \leq 10$, $c_i \leq 10$ for all i (i.e. we can set $B = 10$). Note that we do not need to pick the bound as tight as possible. (It is a separate question of what a tighter possible bound is, but we shall not investigate that.)

The sequence (b_i) is not bounded from above. Suppose that $B \in \mathbb{R}$ was a bound and let $i = \lceil B \rceil$ the smallest integer not smaller than B (i.e. we round up). Then $i \geq B$ and thus

$$b_i = 5 + i \geq 5 + B > B,$$

in contradiction to the assumption that B is an upper bound.

All three sequences (a_i), (b_i), (c_i) are bounded from below (e.g. by 1). However it is possible that a sequence is neither monotonic, nor bounded; for example $d_i = 5 + (-1)^i \cdot i$.

IV.4 Convergence

The best case for describing the long-term behavior of a sequence is if its values "settle down" as the index increases. As with a ball rolling in a bowl, this might not mean to stand still, but simply getting closer and closer to some value L (which we shall call the limit of the sequence). Such a number must be unique, if it exists. In this case we call the sequence *convergent* (otherwise: *divergent*) and L the *limit* of the sequence and write $L = \lim_{i \to \infty} a_i$. (We will give a more formal definition below.)

In the examples of the last section, this should be the case for the sequence $\{a_i\}$ which gets closer and closer to 5. On the other hand the sequences $\{b_i\}$and $\{d_i\}$ "run away" and the sequence $\{c_i\}$ jumps around but never settles down.

Note that one can also talk about the limit of a sequence being infinity (say for the sequence $a_i = i$), but we will not consider it in this section, as the formal statement of it needs to be different.

Finding limits

Since sequences get close to the limit, one way one can attempt to determine a limit (and whether a sequence converges) is by evaluating the sequence for large values of the index and to see whether the values seem to converge. This of course is not a proof (since we do not know whether we have chosen large enough indices), but it often gives a good idea in practice.

For sequences given by quotients of polynomials – say

$$a_i = \frac{6i^5 + i^2 - i - 1}{5i^5 - 8i^2 + i + 1}$$

there actually is a rule (we will see it later in ⇝ VI.5):

- If the degree of the denominator is larger than the degree of the numerator, the sequence converges to zero.
- If the degree of the denominator is smaller than the degree of the numerator, the sequence values will get arbitrary large in absolute value, so the sequence does not converge (or converges to $\pm\infty$).
- If the degree of the numerator and the denominator are equal, the sequence converges to the quotient of the leading coefficients – in the example $6/5$.

We also note (without proof) that limits behave reasonably with respect to arithmetic, that is if we define a new sequence whose entries are given by simple arithmetic operations on existing sequences (e.g. if (a_i) and (b_i) are sequences given by $a_i = (1/i)$ and $(b_i) = i+1$, then $(a_i + b_i)$ is the sequence whose i-th entry is $1/i + 1 + i$).

LEMMA IV.3: Let $\{a_i\}, \{b_i\}$ be sequences such that $\lim\limits_{i\to\infty} a_i = A$, $\lim\limits_{i\to\infty} b_i = B$. Then

- $\lim\limits_{i\to\infty} (a_i + b_i) = A + B$.
- $\lim\limits_{i\to\infty} (a_i - b_i) = A - B$.
- $\lim\limits_{i\to\infty} (a_i \cdot b_i) = A \cdot B$.
- If $B \neq 0$ then $\lim\limits_{i\to\infty} (a_i / b_i) = A/B$.

These laws explicitly require finite limits and do not necessarily generalize to limits of infinity.

Proving Limits

Our next task is to give a formal criterion for convergence. After all, it is not good enough to simply say "I know it if I see it". This criterion, as we will use it, might seem somewhat complicated, but ultimately evolved (over decades of failed attempts in the 19th century) as being unambiguous and formally testable.

This definition also will serve as a model for the formalization of concepts in calculus.

The basic idea is that we want to be able to corral-in the terms of the sequence at a given time, and then (again at a later point in time) make that corral smaller

(without moving it around), and smaller again. As time goes on, it will become arbitrary small (and not become larger again). In other words: If someone tells us how small the corral should be, we must be able to point to a time (that is: to an index!) from when on the corral will be that small.

Formally, what we just called a corral, will be the set of numbers around (at a given distance) a fixed, unmovable point L, which we will call the *limit* of the sequence – the number that the sequence converges towards.

Traditionally, mathematicians use the Greek letter "epsilon" (ε) to denote the distance, and our corral thus is the set

$$\{a \in \mathbb{R} \mid |a - L| \leq \varepsilon\}$$

We thus have established criterion for convergence:

- There is a number L so that
- Given an (arbitrary small) corral size ε
- We can find an index (we shall call N. This N will depend on ε)
- from when on (i.e. for $i \geq N$) all entries a_i of the sequence will be at most ε away from L.

Or, more formally:

Definition IV.4: A sequence $\{a_i\}$ *converges*, if there exists a number L, such that for an arbitrary $\varepsilon > 0$ we can find an N such that $|a_i - L| \leq \varepsilon$ once $i \geq N$. In symbols:

$$\forall \varepsilon > 0 \exists N : \forall i \geq N : |a_i - L| \leq \varepsilon.$$

We write $L = \lim_{i \to \infty} a_i$ and call it the *limit* of the sequence.

Note IV.5: Some books write the inequalities strictly ($i > N$ etc.). This will not change the property, since the statement is for all ε.

For example, consider the sequence given by $a_i = 5 - (-1)^i \frac{10}{i}$, depicted in Figure IV.1. We set $L = 5$ and choose $\varepsilon = 1$. Then for $N = 10$ we have that $|a_i - L| \leq \varepsilon$ when $i \geq N$. Similarly, we determine the corresponding N-values for $\varepsilon = 0.4$, $\varepsilon = 0.2$, and $\varepsilon = 0.125$.

The condition however requires this statement to hold for *any arbitrary* $\varepsilon > 0$, not just for the four values we illustrated. We thus need to give a general proof for an arbitrary ε. For that, we start with a scratch calculation to find when $|a_i - L| < \varepsilon$ and solve for i:

$$\varepsilon \geq |a_i - L| = \left|5 - (-1)^i \frac{10}{i} - g\right| = \left|-(-1)^i \frac{10}{i}\right| = \frac{10}{i}$$

Since i and ε are both positive, we can multiply by i and divide by ε and get

$$i \geq \frac{10}{\varepsilon}.$$

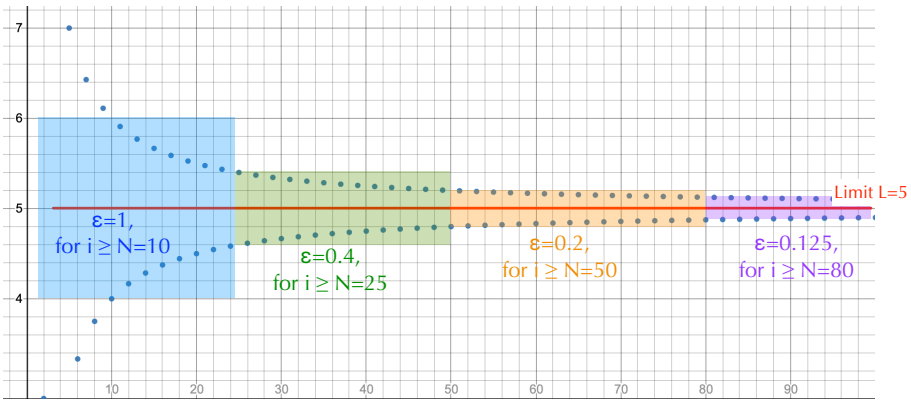


Figure IV.1: A convergent sequence $a_i = 5 - (-1)^i \frac{10}{i}$

The right hand side is the value we chose for N in the actual proof that uses the work of this scratch calculation:

Let $\varepsilon > 0$ and set $N = \frac{10}{\varepsilon}$. Then, for $i \geq N$ we have that

$$\begin{aligned} |a_i - L| &= \left|5 - (-1)^i \frac{10}{i} - 5\right| = \left|-(-1)^i \frac{10}{i}\right| = \frac{10}{i} \\ &\leq \frac{10}{N} = \frac{10}{10/\varepsilon} = \varepsilon. \end{aligned}$$

and thus $\{a_i\}$ converges to L.

With the scratch calculation beforehand, and the choice of L, it is of course predetermined that we will get $|a_i - L| \leq \varepsilon$. The argument just presents the result of the scratch calculation in a format corresponding to the definition of convergence.

Note that some of the text in the proof was underlined. This text will essentially be always the same in any proof for convergence of a sequence, while the remainder is basically built from the scratch calculation.

Note IV.6: You might find this criterion for convergence somewhat unsatisfactory, in that it requires the limit L to be given. Is is clear that we can always calculate, or represent this limit exactly? Consider for example the recursive sequence, defined by

$$a_1 = 1, \qquad a_{i+1} = 1 + \frac{1}{a_i}.$$

If we write a program to calculate values a_i for increasingly large indices, we find that the sequence seems to converge to a number 1.618033988749894848 . . . . But

what is this limit[2], and how can we express it and use it in a limit proof? To work around such problems, mathematicians often use a somewhat different criterion, called the *Cauchy criterion* (see Wikipedia), that uses not the distance from sequence values to the limit, but of sequence values never straying too far apart from each other. It however is a bit more complicated to work with, which is the reason for the choice of a more basic criterion here.

The reason for looking at formal proofs is that this is a prototype for the way formal proofs work in Calculus[3]. One argues about differences, typically denoted by ε, in function values becoming arbitrary small, but still being nonzero, as sequence indices go to infinity, or x-arguments becoming arbitrarily close (the latter then being denoted by another Greek letter, δ denoting a small entity).

IV.5 Limits, Bounds and the Real numbers

If a sequence converges, its values cannot become arbitrary large, since at some point they need to start getting close to the limit. This implies that a convergent sequence must be bounded. On the other hand, being bounded does not make a sequence convergent, as the example $(-1)^i$ shows.

Things however become interesting if a sequence is monotonic and bounded (that is increasing and bounded from above, or decreasing and bounded from below). We have the following theorem:

THEOREM IV.7: Every monotonically increasing sequence that is bounded from above must converge to a limit $L \in \mathbb{R}$. (As must every decreasing sequence bounded from below.)

Conceptually this seems clear – if we can never go back, nor cross a line, we eventually must become stationary. A formal proof however is far beyond the scope of this course.

Note that the limit in such a case is not necessarily the bound used, but it is the "best possible" bound (formally called the *supremum*).

This statement might seem more of a curiosity and have little relevance for computational sciences. The reason we mention this fact is in that it ultimately gives a justification for and explanation of what the real numbers are. So far, we defined them, somewhat sloppily, as the x-coordinates of all points on the number line, or as numbers with infinite decimal expansions. But that is only half the truth. Formally:

The real numbers are the limits of bounded, monotonic sequences.

This construction is also the reason for why we use the real numbers (and not just rational numbers of arbitrarily large denominator and arbitrarily good expan-

[2] In this case, one can actually calculate it as a solution to the quadratic equation $x = 1 + 1/x$

[3] When doing proofs, the topic is often called "Analysis"

sions): real numbers are required to get limits of sequences that should converge. Such limits do not necessarily exist in the rational numbers – take for example the sequence recursively defined in Note IV.6, whose limit is the irrational number $\frac{1+\sqrt{5}}{2}$. (There is a formal construction of the real numbers that defines an equivalence relation on sequences – multiple sequences might have the same limit – and defines the real numbers as equivalence classes of sequences under this relation.)

But the real numbers are not only what one can obtain by arithmetic involving roots. Actually there are infinitely (in an overwhelming sense) more real numbers, than numbers that can be expressed as roots of polynomials. (Such numbers are called *transcendental*.) You will have seen some examples in school: π, e. But these numbers are like dark matter. While they are overly abundant and around everywhere, it is hard to get hold of them (as one can only give a numerical approximation or a sequence that has them as limit).

Of course anything we can measure in the real world, anything we can represent as number on the computer, is a rational number. The reason we use the real numbers is so that limits exist!

IV.6 Series

A series is a special sequence, in which we sum up the terms of another sequence, that is it is the sequence of *partial sums*. If the sequence we are summing over is (a_j), we have the i-th partial sum as

$$p_i = a_0 + a_1 + a_2 + \cdots + a_{i-1} + a_i = \sum_{j=0}^{i} a_j,$$

where the $\sum$ notation is basically a mathematician's version of a for-loop. (Note that we need to use a different variable for the loop than for the loop end.)

If this sequence of partial sums converges we write the limit as an infinite sum.

$$\sum_{j=0}^{\infty} a_j = \lim_{i\to\infty} p_i = \lim_{i\to\infty} \sum_{j=0}^{i} a_j.$$

For example, we could sum over the sequence $a_j = 1/2^j$ and get the sequence of partial sums:

$$p_0 = 1, p_1 = 1 + \frac{1}{2} = \frac{3}{2}, p_2 = 1 + \frac{1}{2} + \frac{1}{4} = \frac{7}{4}, \ldots$$

Here, it is not hard to see that the partial sums are

$$p_i = \frac{2^{i+1} - 1}{2^i}$$

and this sequence thus has the limit

$$\sum_{j=0}^{\infty} \frac{1}{2^j} = 2,$$

that is, the infinite summation yields a finite number.

Clearly, the sequence over which we are summing must go to zero for the series to converge. But that is not sufficient. For example, one can show that the series

$$\sum_{i=1}^{\infty} \frac{1}{i}$$

(called the *harmonic series*) will not converge but go to infinity.

As an interesting aside, the program

```
s=0
old=-1
i=1
while s>old :
    old=s
    s=s+1/i
    print(i, ": ",s)
    i=i+1
```

(which calculates the partial sums $\sum_{i=1}^{n} \frac{1}{i}$ for increasing values of n and terminates when they stop increasing) will actually terminate (due to rounding errors) after a long time[4] and thus make it look as if the the series converges. But it does not!

There are a number of criteria and tests to see whether a particular series converges, and you will find examples and applications in any Engineering Calculus book. In this course however we are not investigating this question, but will look at two particular kinds of series (where it is relatively easy to indicate convergence).

NOTE IV.8: There is some subtlety with infinite summations: Usually we have that $a + b = b + a$, but it is possible that a series changes its value (i.e. the limit of the partial sums), or even whether it converges, if we change the order of summation. This cannot happen if the sequence converges and all terms are positive – the issue is basically that we can unbalance positive and negative terms by moving some more and more towards infinity.

IV.7 Geometric Series

A sequence is called *geometric* if its terms grow by a constant factor r, that is, if the starting term is a, the sequence is

$$a, ar, ar^2, ar^3, ar^4, \ldots$$

We will assume that $r \neq 1$, as the sequence is otherwise boring.

[4] You can replace the line incrementing `s` by `s=round(s+1/i,6)` to reduce accuracy and make it happen sooner.

A geometric series now is a series whose terms show geometric growth, i.e. $a_i = r \cdot a_{i-1}$ for a constant r, independent of i. This type of series has applications in several fields including physics, biology, economics and finance.

Definition IV.9: Let r be a ratio and a any nonzero constant. A *geometric series* is a series of the form

$$a + ar + ar^2 + \ldots + ar^{n-1} + \ldots = \sum_{n=1}^{\infty} ar^{n-1} = a \sum_{n=1}^{\infty} r^{n-1} = a \sum_{n=0}^{\infty} r^n$$

The k^{th} partial sum p_k of a Geometric series $\sum_{n=0}^{\infty} ar^n$ is the (finite) sum of the first $k+1$ terms:

$$p_k = a + ar + ar^2 + \ldots + ar^k = \sum_{i=0}^{k} ar^i$$

A simple example of geometric growth is something that grows by a factor each time unit. Suppose that a museum starts with $a = 100$ visitors per year and each year increases its visitor numbers by 3%. That is the visitor number after i years is $100 \cdot 1.03^i = a \cdot r^i$ for $r = 1.03$. If we ask for the total number of visitors after n years we get:

$$\sum_{i=0}^{n} a \cdot r^i.$$

Now consider

$$p_k = a + ar + ar^2 + \ldots + ar^k$$

and

$$rp_k = ar + ar^2 + ar^3 + \ldots + ar^{k+1}.$$

Notice that these two expressions share several terms in common, in fact

$$p_k - rp_k = (a + ar + ar^2 + \ldots + ar^k) - (ar + ar^2 + ar^3 + \ldots + ar^{k+1}) = a - ar^{k+1}$$

Now if we factor both sides of this equation, we get

$$p_k(1 - r) = a(1 - r^{k+1}).$$

As long as $r \neq 1$ we can divide both sides by $(1 - r)$ to get

$$p_k = \sum_{n=0}^{k} ar^n = a\frac{1 - r^{k+1}}{1 - r}.$$

(What happens in the case where r=1? We get $p_k = a + a(1) + a(1)^2 + \ldots + a(1)^k = (k+1)a$.)

Before we go on to get a formula for the infinite sum, we see how the formula for p_k might be used on its own: Suppose we are given a chess board and place one

grain of rice on the first square, two grains of rice on the second square, four grains of rice on the third square continuing on so that there are 2^{n-1} grains of rice on the n^{th} square. How many grains of rice are on the chess board? (There are $8 \cdot 8 = 64$ squares on a chess board.)

Since there are 64 squares on the chess board (from 0 to 63), we want to compute p_{63}. Our ratio is 2 and a is one, hence we are asked to find

$$p_{63} = 1 + 2 + 4 + \ldots + 2^{63} = \frac{1(1-2^{64})}{1-2} \approx 1.84467 \times 10^{19}$$

A similar calculation underlies the repayment of loans or mortgages:

Suppose we take out a loan of L dollars which is paid back periodically (typically monthly). The periodic payment is a dollars, the fixed interest rate *per period* is i. If b_k is the loan sum outstanding after k time periods, we have that

$$b_{k+1} = b_k \cdot (1+i) - a.$$

Using $b_0 = L$ and setting $r = 1 + i$, in the first step we have

$$b_1 = Lr - a$$

we then continue on to the next step where we get

$$\begin{aligned} b_2 &= b_1(r) - a = (Lr - a)(r) - a \\ &= Lr^2 - (a + ar) \end{aligned}$$

We might be starting to notice a pattern, however it becomes obvious in the next step

$$\begin{aligned} b_3 &= b_2(r) - a = (Lr^2 - (a + ar))(r) - a \\ &= Lr^3 - (a + ar + ar^2) \end{aligned}$$

At this point we can solve this recursion to

$$b_k = L{\cdot}r^k - \left(a + ar + ar^2 + \cdots + ar^{k-1}\right) = L{\cdot}r^k - \sum_{n=0}^{k-1} ar^n \quad \overset{\text{Sum formula}}{=} \quad L{\cdot}r^k - a\frac{1-r^k}{1-r}.$$

A bank now would set $b_m = 0$ (where m is the time after which the loan should be paid off, e.g. $m = 12 \cdot 30 = 360$ for a 30 year mortgage) and solve for a to determine the necessary monthly repayment, given the loan sum and interest rate.

For example, if we have a mortgage of $L = \$200,000$, an annual interest rate of 6% (leading to a monthly rate of $i = .06/12 = 0.005$, i.e. $r = 1.005$) and a monthly repayment sum[5] of $a = \$1,200$, we find that the outstanding sum after k months is

$$b_k = 200,000 \cdot 1.005^k - 1,200\frac{1.005^k - 1}{0.005} = 200,000 \cdot 1.005^k - 240,000\left(1.005^k - 1\right).$$

[5] Slightly moralistic remark: Incidentally, initial interest amounts to \$1000 per month at the start of the loan. An *interest only* loan thus does not save much and is a very bad deal!

After 10 years (120 months) this leaves an outstanding amount of \$167, 224.13, after 20 years \$107, 591.82, roughly half[6], after 30 years \$−903.00 (i.e. the loan is paid off after 30 years less one month[7]).

If the interest rate instead was 7% annually ($i = .07/12 = 0.005833$ monthly), we get with same repayment sum a remaining loan amount of \$159, 256.18 after 30 years, which is not even halfway paid off. A monthly repayment sum of \$1330 would be needed[8] to have the loan paid off after 30 years.

Since the sequence $p_0, p_1, p_2, \ldots, p_k, \ldots$ of partial sums converges to $\sum_{n=1}^{\infty} ar^{n-1}$, we can use the formula just derived to compute a value for the limit of a geometric series:

If $|r| < 1$, we have that $\lim_{k\to\infty} r^{k+1} = 0$ and thus

$$\sum_{i=0}^{\infty} ar^i = \lim_{k\to\infty} p_k = \lim_{k\to\infty} \frac{a(1-r^{k+1})}{1-r} = \frac{a(1-0)}{1-r} = \frac{a}{1-r}.$$

This incidentally verifies the above example

$$\sum_{i=0}^{\infty} \frac{1}{2^i} = 2$$

On the other hand, if $|r| \geq 1$, the absolute value of the numerator $1 - r^{k+1}$ will go towards infinity, and the series diverges.

As an example application of this formula, we look at how one can express the repeating decimal $0.\overline{08}$ as a fraction of two integers.

We first notice that 08 is the repeated value, so we want to break up $0.\overline{08}$ into a sum of the repeated values. Therefore we have $0.\overline{08} = 0.\underline{08} + 0.00\underline{08} + 0.000\underline{08} + \ldots$. Once we have this written as a sum we can now write each term in the sum as a fraction, hence we have $0.\underline{08} = \frac{8}{100}$, $0.00\underline{08} = \frac{8}{10{,}000} = \frac{8}{100^2}$, and $0.000\underline{08} = \frac{8}{100^3}$. Using this information we should recognize a pattern

$$0.\overline{08} = \frac{8}{100} + \frac{8}{100^2} + \frac{8}{100^3} + \ldots = \sum_{i=1}^{\infty} \frac{8}{100^i}$$

However at this point we notice that we have a geometric series which is not written in the correct form for us to apply our formula. If we factor out $(\frac{8}{100})$ we get $\sum_{i=0}^{\infty} (\frac{8}{100})(\frac{1}{100})^i$ which is now in the correct form. This gives $r = \frac{1}{100} < 1$ so our

[6] In general, this means that a loan is paid off half after roughly $\frac{2}{3}$ of its planned life time

[7] The total cost of the loan then will have been \$430, 800, more than double the loan amount. (Though inflation means that the actual value will be less.)

[8] I.e. a change of one percentage point in the interest rate increased the monthly payment (and thus the total loan cost) by 10%! No wonder people go bonkers about interest rates.

geometric series converges and since $a = \frac{8}{100}$ the series converges to

$$\frac{\frac{8}{100}}{1-\frac{1}{100}} = \frac{\frac{8}{100}}{\frac{99}{100}} = \frac{8}{99} = 0.\overline{08}.$$

For another example, consider the following paradox of Zeno[9]:

Achilles, the fastest runner in ancient Greece, runs 100 times as fast as the Tortoise. But — so the paradox claims — if the Tortoise is given an advance, Achilles will never be able to pass the Tortoise:

Suppose that Achilles runs $10m/s$ and the Tortoise only $0.1m/s$, furthermore the Tortoise is given a head start of $100m$. After 10 seconds, Achilles has reached the place where the Tortoise started. But in this time, the Tortoise has run $1m$ ahead. Achilles will reach this distance in 0.1 seconds. But then the Tortoise has moved another $0.01m$. Achilles will take 0.001 seconds to reach this, and so on. He will never reach the Tortoise. Where is the error in this argument?

The paradox is resolved, if we realize what time period is considered: All events take place within

$$10 + 0.1 + 0.001 + \cdots = 10\sum_{n=0}^{\infty}\frac{1}{100^n} = \frac{10}{1-\frac{1}{100}} = \frac{1000}{99} = 10.\overline{10}$$

seconds. This is exactly the time *when Achilles reaches (and overtakes!) the Tortoise*. The paradox arises from the implicit (and wrong, as we have calculated!) suggestion, that it describes the events for all time.

IV.8 Arithmetic Series

While not directly related to the Geometric Series, this seems to be an appropriate place to also look at summations over other sequences.

The summation over a constant sequence with fixed value a,

$$\underbrace{a + a + \cdots + a}_{n \text{ terms}} = n \cdot a$$

is basically just the definition of multiplication. But what happens if the terms change by a constant sum, that is we have that $a_{i+1} - a_i = c$ constant?

Let us first observe this for the basic case of

$$S_n = \sum_{i=1}^{n} i = 1 + 2 + \cdots + n.$$

We get a formula by writing this sum down a second time in reversed order and add both up (thus getting *twice* the value S_n):

$$\begin{array}{rccccccc} & 1 & +2 & +3 & +\cdots & +(n-1) & +n \\ + & n & +(n-1) & +(n-2) & +\cdots & +2 & +1 \end{array}$$

[9] Zeno of Elea, Greek philosopher, about 490 BC - 430 BC

We can get the value in another way by first adding down in each column. Note that each column adds up to $n+1$, and we have n columns, and this sum thus is $n(n+1)$. This gives us the formula:

$$\sum_{i=1}^{n} i = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

For example, we get that

$$1 + 2 + \cdots + 100 = \frac{100 \cdot 101}{2} = \frac{10100}{2} = 5050.$$

If subsequent terms differ by a different constant, we can handle this with a factor (and possibly a starting term). Suppose we start at 5 and differ by 3 in each step:

$$5 + 8 + 11 + 14 + \cdots + 32$$

we subtract 2 from each summand, so that they all are a multiple of 3, and then separate the two operations:

$$= (2+3) + (2+6) + (2+9) + \cdots (2+30) = \underbrace{2 + \cdots + 2}_{10 \text{ terms}} + 3\,(1 + 2 + \cdots + 10)$$

We know that there are 10 terms, as we have $(32-5)/3 = 9$ steps from the first. Now we use the two kinds of formulas and get

$$5 + 8 + 11 + 14 + \cdots + 32 = 10 \cdot 2 + 3 \cdot \left(\frac{10 \cdot 11}{2}\right) = 20 + 3 \cdot 55 = 185.$$

We note, without proof, the similar formulas:

$$\begin{aligned} \sum_{i=1}^{n} i^2 &= \frac{n(n+1)(2n+1)}{6} \\ \sum_{i=1}^{n} i^3 &= \frac{n^2(n+1)^2}{4} \end{aligned}$$

Chapter

V Differentiation

We are now almost ready to start looking at the concepts of Calculus. The underlying idea is that we should be able to describe the values of a function, or its long term behavior, based on how it changes locally.

V.1 Function limits and Continuity

Before defining the derivative of a function on the real numbers, there are some, somewhat technical concepts, we need to touch upon:

First, we want to extend the concept of the limit of a sequence (at ∞) to the limit of a function at a point.

Definition V.1: Let $f: \mathbb{R} \to \mathbb{R}$ be a function and $a \in \mathbb{R}$. If, for every sequence $\{a_n\}$ with $\lim_{n\to\infty} a_n = a$, the sequence of function values $\{f(a_n\}$ also converges to $L = \lim_{n\to\infty} f(a_n)$, and the limit L does not depend on the choice of the sequence, we call this the limit of f at a, written

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

(Calculating such a limit can potentially be hard, as one has to consider an infinitude of possible sequences.)

For many functions occurring in the real world, this limit is equal to the function value $f(a)$, since nature does not jump, but this is not guaranteed for every function.

Our definition of functions however allows for *arbitrary* functions, such as the

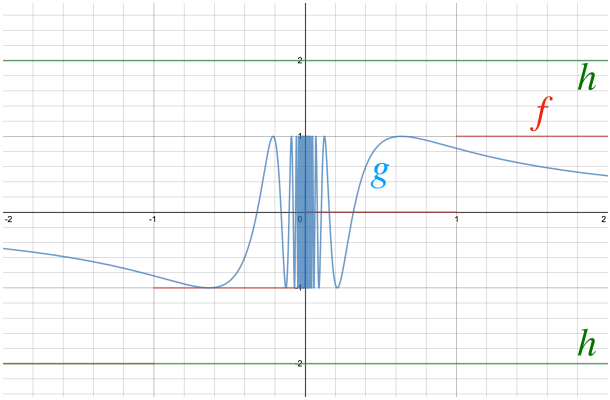


Figure V.1: Some discontinuous functions

following ones (Figure V.1):

$$f\colon \mathbb{R} \to \mathbb{R} \quad , \quad x \mapsto \lfloor x \rfloor \qquad \text{Round down to integer}$$

$$g\colon \mathbb{R} \to \mathbb{R} \quad , \quad x \mapsto \begin{cases} \sin(1/x) & x \neq 0 \\ 1 & x = 0 \end{cases}$$

$$h\colon \mathbb{R} \to \mathbb{R} \quad , \quad x \mapsto \begin{cases} 3 & x \in \mathbb{Q} \\ -3 & x \notin \mathbb{Q} \end{cases}$$

If we investigate the function f around $1/2$, there is no way to see that it jumps at 0 and at 1. Similarly, why is the value of g at 0 one (and not zero or -1, or something in between). And we can't even look fine enough to decide from the graph that h is a function.

Instead, we want that the graph of the function is smooth, that is that we get close to a point x, we can predict the function value $f(x)$. Getting closer will ultimately give a better approximation.

Formally, we define a function $f\colon \mathbb{R} \to \mathbb{R}$ to be *continuous* at a point x_0 (otherwise: *discontinuous*) if

$$\text{The limit } \lim_{x \to x_0} f(x) \text{ exists and is equal to the function value } f(x_0).$$

If the function is continuous at every point x_0, we simply call it continuous (without "at").

We see for example that the "rounding down" function $f(x) = \lfloor x \rfloor$ is not continuous at 1, since the sequence $a_i = 1 - 1/i$ has $\lim_{i\to\infty} a_i = 1$, but $f(a_i) = f(1 - 1/i) = \lfloor 1 - 1/i \rfloor = 0$, and thus $\lim_{i\to\infty} f(a_i) = 0i \neq 1 = f(1)$.

Informally, a function is continuous, if small changes in the argument imply small changes in the value – that is we can approximate the function value at x_0 by

the function values at numbers close to x_0. For the graph of the function this means that it may not have any jumps, nor start wild oscillations, but should be followed easily with a pencil. Most functions you will encounter (that do not jump) are likely to be continuous.

To show formally that a function is continuous, using the definition, can be hard. We therefore do not investigate this further in this class[1]. Instead we study what continuity is good for.

Some continuous functions

Many functions you know from school are continuous. And due to the laws of limits it is not just these functions, but also functions composed by arithmetic operations (as long as a denominator does not become zero), and also compositions of such functions:

- Nonnegative powers of x: x^a for $a \in \mathbb{R}$. This includes the constant function x^0.
- Thus also polynomials and (as long as the denominator is nonzero) rational functions.
- Trigonometric functions $\sin(x), \cos(x)$, and $\tan(x) = \sin(x)/\cos(x)$ (when the denominator is nonzero).
- Inverse trigonometric functions (Note that arcsin and arccos are only defined on the domain $\{-1 \dots 1\}$.
- The exponential function $\exp(x) = e^x$ and (for positive x) its inverse the natural logarithm $\log(x) = \ln(x)$. (All logarithms in this course without a specific basis are natural logarithms.)

The definition of some of these functions (such as $\sin(x)$ as measurement on a triangle in a circle) given in school is not amenable to easy computation, we will see later ⇝ VII.2 how this can be done.

V.2 Why Care About Continuity

The first use of continuity is that it makes approximation meaningful. We can work, for example, with numerical approximations of x values, and trust that the function value $f(x)$ will not differ too much from $f(x_0)$ if x is an approximation of x_0.

A consequence of this is that we can "control" the function values. If we want to find x to achieve a particular function value $f(x)$, we can do so by iterative approximation. On the other hand, the situation that small changes in the input

[1] There is a criterion that is used in the Calculus class for mathematicians that formalizes the "close approximation" idea, but we do not need it here

could produce (arbitrary) large changes in the output is close to the definition of chaos.

A consequence of this property is the following statement:

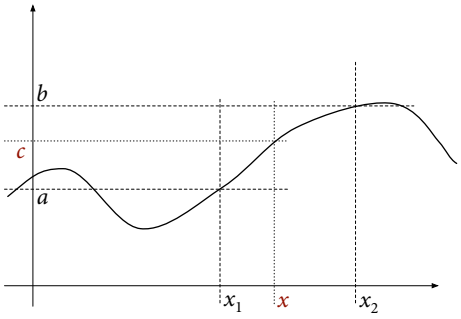


Figure V.2: The intermediate Value Theorem

THEOREM V.2 (Intermediate Value Theorem): Let $f: \mathbb{R} \to \mathbb{R}$ be a continuous function and $x_1 < x_2 \in \mathbb{R}$ with $f(x_1) = a$ and $f(x_2) = b$. Let c be a number between a and b. Then there exists $x_1 \leq x \leq x_2$ such that $f(x) = c$ (Figure V.2).

We can use this theorem to find x for a particular $f(x)$-value. Most frequently this is done to find x, such that $f(x) = 0$. The method is called *halving intervals* or *bisection* method.

1. Start with $s, t \in \mathbb{R}$ such that $f(s) < 0$ and $f(t) > 0$. (Or vice versa: $f(s) > 0$ and $f(t) < 0$.)
2. Let $m = \frac{s+t}{2}$ be the midpoint of the interval. If $f(m) = 0$ and stop. Otherwise:
3. If $f(s)$ and $f(m)$ have the same sign (positive or negative) replace s by m. Otherwise replace t by m (since $f(m)$ and $f(t)$ have the same sign).
4. If $|t - s|$ is not small enough (to the approximation quality we want), go to step 2.

In each step of the algorithm, the length of the interval from s to t halves, and the desired x value must lie in the interval. We can repeat until s and t approximate this x sufficiently well.

For example, consider the function $f(x) = \sin(x)$. We want to approximate the number π, knowing[2] that $f(\pi) = 0$. We start with $s = 2$ and $t = 4$, since we know that $2 < \pi < 4$. We now iterate, setting $m = (2 + 4)/2 = 3$ and calculate $f(3)$, which is positive. Thus we replace s by 3 and iterate. The following table shows the

[2] We measure the angle of a full circle as 2π

further iterations (with underlined numbers indicating the end point that was replaced by m in each step). After 20 iterations, the length of the interval has become $2/2^{20} \sim 10^{-6}$. This agrees with the fact that we gain a further correct digit every $\log_2(10) \sim 3.3219$ steps and thus should expect $20/3.3219 = 6$ correct digits in the result, approximating π as 3.141593 (versus the correct[3] 3.1415926 . . .).

#	s	t	$\lvert t-s\rvert$	$f(s)$	$f(t)$	$m = \frac{s+t}{2}$	$f(m)$
0	2	4	2	0.909297	-0.756802	3	0.141120
1	3	4	1	0.141120	-0.756802	3.5	-0.350783
2	3	3.5	0.5	0.141120	-0.350783	3.25	-0.108195
3	3	3.25	0.25	0.141120	-0.108195	3.125	0.016592
4	3.125	3.25	0.125	0.016592	-0.108195	3.1875	-0.045891
5	3.125	3.1875	0.0625	0.016592	-0.045891	3.156250	-0.014657
6	3.125	3.15625	0.03125	0.016592	-0.014657	3.140625	0.000968
7	3.140625	3.156250	0.015625	0.000968	-0.014657	3.148438	-0.006845
8	3.140625	3.148438	0.007813	0.000968	-0.006845	3.144531	-0.002939
9	3.140625	3.144531	0.003906	0.000968	-0.002939	3.142578	-0.000985
10	3.140625	3.142578	0.001953	0.000968	-0.000985	3.141602	-0.000009
11	3.140625	3.141602	0.000977	0.000968	-0.000009	3.141113	0.000479
12	3.141113	3.141602	0.000488	0.000479	-0.000009	3.141357	0.000235
13	3.141357	3.141602	0.000244	0.000235	-0.000009	3.141479	0.000113
14	3.141479	3.141602	0.000122	0.000113	-0.000009	3.141541	0.000052
15	3.141541	3.141602	0.000061	0.000052	-0.000009	3.141571	0.000022
16	3.141571	3.141602	0.000031	0.000022	-0.000009	3.141586	0.000006
17	3.141586	3.141602	0.000015	0.000006	-0.000009	3.141594	-0.000001
18	3.141586	3.141594	0.000008	0.000006	-0.000001	3.141590	0.000003
19	3.141590	3.141594	0.000004	0.000003	-0.000001	3.141592	0.000001
20	3.141592	3.141594	0.000002	0.000001	-0.000001	3.141593	-0.000000

Such a process of halving intervals can be implemented easily. It however takes a while to get a good approximation, which is why we will see a better method in a later chapter ⇝ VI.4.

V.3 Partial Sums and Derived Sequences

Before defining derivatives properly, let us look at an example of change that happens at discrete intervals, and how function values and change are related.

Imagine the ledger of a business that lists every day the sum of income minus expenses (let's call it the *flow*), and the total money held by the business:

[3] Count the digits in each word in the sentence *How I want a drink, alcoholic of course, after the heavy chapters involving quantum mechanics.*

Day	Income-Expenses	Money held
0	-	0
1	893	893
2	70	963
3	992	1955
4	682	2637
5	115	2752
6	215	2967
7	-497	2470
8	-246	2224
9	252	2476
10	-301	2175

Going through the columns, we get two sequences, both indexed by the first column. The first sequence, which we shall call a_i is the daily flow. The second sequence, let's call it b_i, is the money held by the business. Is there a relation between the two columns?

Of course. Assuming we started with 0 money held, the money held on the end of day i is the sum of the flow of days 1 to i. We write this as

$$b_i = a_1 + a_2 + a_3 + \cdots + a_{i-1} + a_i = \sum_{j=1}^{i} a_j$$

and have that the b_i are the *partial sums* over the sequence a_i.

Can we do the same thing backwards? Surely – solving for a_i finds that

$$a_i = \sum_{j=1}^{i} a_i - \sum_{j=1}^{i-1} a_i = b_i - b_{i-1}.$$

The two sequences are thus "related" in that each sequence completely determines the other one and vice versa. We shall call the sequence $\{a_i\}$ the *derivative sequence* of the sequence $\{b_i\}$ and the sequence $\{b_i\}$ an *antiderivative* or an *indefinite integral* of $\{a_i\}$.

Calculus is about studying such a correspondence between functions defined on (e.g.) the real numbers, while sequences are functions defined on the positive integers.

Before going there, let us look at a few ways how this correspondence plays out and helps us with determining information. First, imagine we would have started not at 0 but with some money in the bank, say 1000 currency units. Then the ledger would have looked almost the same, but for the last column being increased by 1000:

Day	Income-Expenses	Money held
0	-	1000
1	893	1893
2	70	1963
3	992	2955
4	682	3637
5	115	3752
6	215	3967
7	-497	3470
8	-246	3224
9	252	3476
10	-301	3175

Denote the sequence given by the third column in this ledger by c_i. We note that c_i is built from the *same changes* as b_i is. Thus a_i is also the derivative of c_i and c_i is an integral of a_i. (That is why we said "an integral" and not "the integral".) It is not hard to see that there are many more antiderivatives (namely different starting values in the bank), but that any two antiderivatives simply differ by a constant (namely the difference of their starting account values).

If we care about the flow over a number of days, say from day 4 to day 7 (inclusively), we need to add up the flows of these three days:

$$a_4 + a_5 + a_6 + a_7 = \sum_{j=4}^{7} a_j.$$

Using the sum notation, we see that this multi-day flow difference also can be expressed as a difference of values of an antiderivative over multiple days, we have that

$$\sum_{j=4}^{7} a_j = (\sum_{j=1}^{7} a_j) - (\sum_{j=1}^{3} a_j) = b_7 - b_3 = c_7 - c_3.$$

Here 7 is the time when the counting ends (the evening of day 7) and 3 the time when it starts (the evening of day 3 as giving the same amount as on the morning of day 4 which is not listed separately in the ledger.)

This difference formula will hold for whatever antiderivative we are choosing. This holds, because the starting account value has no impact on the flow over the four days we are measuring. Thus antiderivatives (more specifically the *difference* of the values of antiderivatives between a start and an end point, which will be called a *definite integral*.) We will encounter this easy idea later again under the name of "Fundamental Theorem of Calculus".

We have seen that an antiderivative helps with summing up changes over a period. But what can derivatives be used for? To illustrate this, look at a different sequence $\{b_i\}$ (starting with b_0) whose changes are smaller, but which we tabulate

over a longer range:

$$0, 8, 15, 21, 24, 26, 28, 26, 25, 23, 21, 19, 18, 16, 15, 14, 13, 14, 15, 16, 18, 20, 22, 25, 28, 31, 34, 37, 40, 43, 46, 48, 50, 51, 52, 53, 52, 51, 49, 48, 45, 42, 40, 36, 33, 31, 28, 27, 26, 27, 30$$

We get a sequence of changes $\{a_i\}$ (starting with a_1) as $a_i = b_i - b_{i-1}$:

$$7, 6, 3, 2, 2, -2, -1, -2, -2, -2, -1, -2, -1, -1, -1, 1, 1, 1, 2$$

The values of both sequences are depicted (with points connected) in Figure V.3.

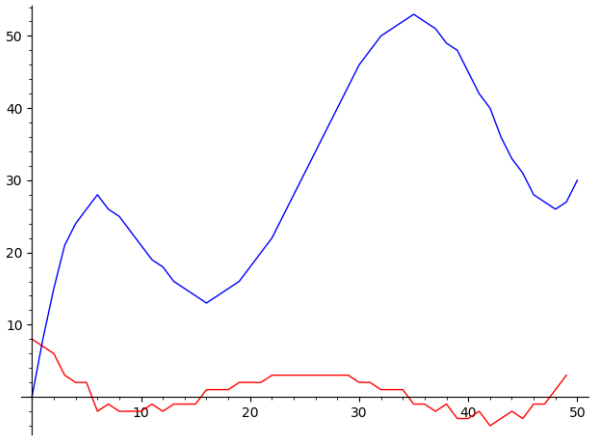


Figure V.3: A sequence and its derivative

An obvious question one can ask for such a sequence b_i is for what the maximum and minimum (largest and smallest) values over the investigated period are. (In the previous example these would have been the lowest and the highest worth of the business.) We mark the areas where the function is (locally, that is in relation to its neighbors) maximal or minimal in Figure V.4.

We note that at these index values (for which the function bn is maximal, respectively minimal), which are aligned along the x-axis, the derivative is zero. (An eagle-eyed reader might notice that we are slightly cheating here: Since we sum up the values of the derivative, the maximum happens at the x-value plus 1, and our derivatives are only close to zero. This will be resolved later when we will decrease the step-width more and more.)

Furthermore, at an index i, where b_i has minimum value, the derivative a_i changes from being negative to being positive. And at the index i where b_i has maximum value, the derivative changes sign from positive to negative. The reason

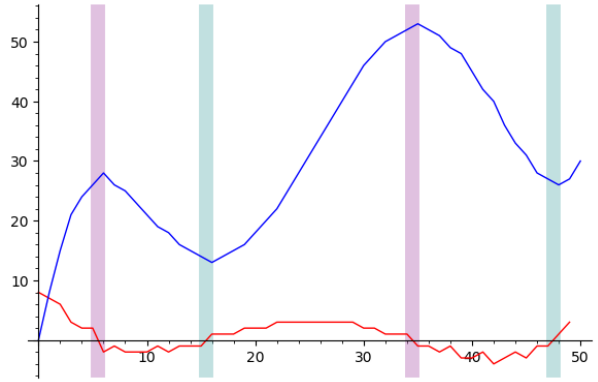


Figure V.4: Maxima and Minima

for this is easily understood. To be a maximum value (say),at index im the values must have grown in step i (that is the derivative is positive at i), as otherwise the value cannot be larger than other different ones. But if the derivative did not become negative at $i + 1$, the function would have been growing even further.

Thus, if we look at all places where the derivative changes from positive to negative we see that these are exactly the *local maxima*, that is the places where the function is larger than in the neighborhood (with the same argument about growing before but not after).

A similar kind of argument can be made for local minima.

Finding indices where the derivative is (close to) zero is easy. The derivative thus allows us to find maxima or minima (which are harder to find).

This is a further indication that the concept of a derivative is a useful concept, and we will spend most of the remainder of the class studying derivatives and antiderivatives and their consequences.

V.4 Aliasing

The concept of summing up changes is rather basic, and the reader might wonder what else there could be to Calculus. One fundamental issue is that the prior example had a natural step-width (the day), while this is not true for many other situations. Indeed, results might be fundamentally wrong if we impose an artificial step-width. This phenomenon goes under the name of *aliasing* and also can occur when digitizing pictures or sound signals. In this section (which is not required later on) we illustrate what can happen.

Imagine we have temperatures oscillating as on a typical winter day in Colorado, with a cold mornings but temperate early afternoons. The red curve in Fig-

ure V.5, left (with the x-axis indicating days) shows how temperature changes over time. Now an alien (who has no concept of an Earth day) samples the temperature, namely in intervals if 0.8 days. The blue dots indicate the measurements. Based on these measured values, one clearly would expect the temperature to fluctuate periodically in a cycle of 4 days, as indicated by the blue curve in Figure V.5, right. Different sample periods will lead to differently bad results (a sample period of ex-

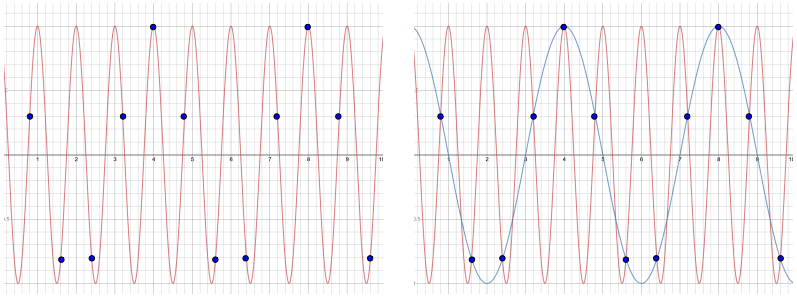


Figure V.5: Bad sampling leads to aliasing

actly one day would indicate that

The *Nyquist-Shannon* sampling theorem in fact shows, that one needs to sample a periodic signal at at least twice the maximal frequency (that is twice per period) to be able to detect these frequencies.

The same effect can be seen in so-called *Moiré* patterns when overlaying fabric meshes or when naively reducing the size of a digital image, as in Figure V.6: When reducing the image by simply sampling pixels at regular intervals, not only does an underlying regular grid (from taking a picture off a monitor) become overwhelming, even the direction of the grid changes! A more sophisticated algorithm instead will try to keep colors in local areas the same and is able to produce a much better result (the right image has the same reduced size parameters as the middle one).

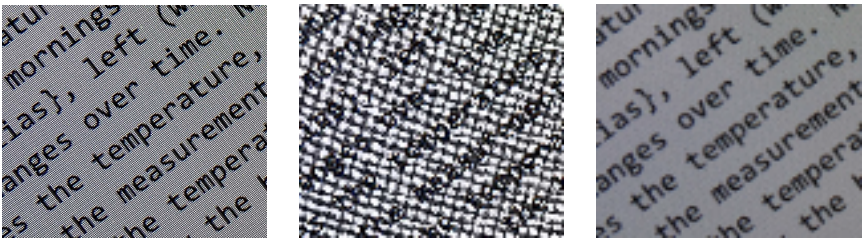


Figure V.6: Image reduced by bad and by good algorithm

V.5 The Derivative of a Function

We have seen that it can be useful to study how the values of a function change. Let us look at what this means for a function defined on the real numbers: We could, at a point x_0 look how the function changes between x_0 and $x_0 + 1$ (namely by $f(x_0 + 1) - f(x_0)$). But there is really no reason to take a step of 1. We could similarly look at the change when taking a step by 2 or by $\frac{1}{2}$. And of course one would expect the change to be larger if the step-width becomes larger.

Secant Slope

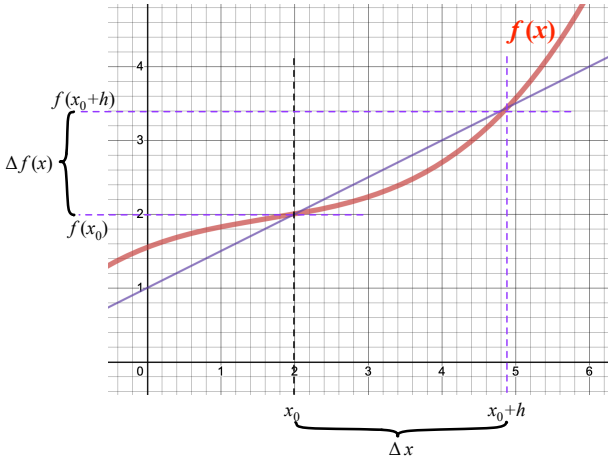


Figure V.7: A secant

We can deal with this variability by not considering the absolute change $f(x_0 + 1) - f(x_0)$ by one unit, but by considering a variable step width (which we denote by h or by Δx, Δ being used here to indicate a difference), and to consider the change relative to the step width. That is, we consider the fraction

$$\frac{f(x_0 + h) - f(x_0)}{(x_0 + h) - x_0} = \frac{f(x_0 + h) - f(x_0)}{h} \tag{V.3}$$

This fraction is the slope of the line between the points $(x_0, f(x_0))$ and $(x_0 + h, f(x_0 + h))$, Figure V.7. Such a line is called a *secant*, as it intersects the graph twice. The ratio in equation (V.3) gives the slope of this secant, and is called the *difference quotient*. Since the numerator is the change of function values, sometimes this quotient is also written as $\frac{\Delta f(x)}{\Delta x} = \frac{\Delta y}{\Delta x}$.

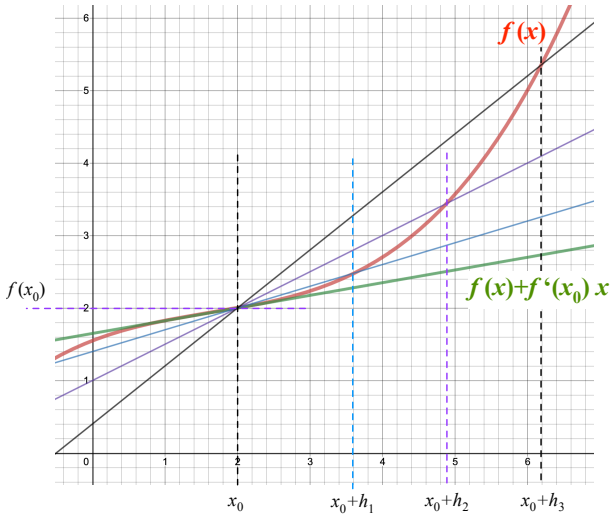


Figure V.8: Multiple secants and the tangent

But we can choose different values of h (Figure V.8), and (unless the function is a line) this will yield different secants with different slopes, depending on the value of h.

To remove this ambiguity we now make h small. That is, we consider the difference quotient as a function of h and study the limit

$$\lim_{h\to 0} \frac{f(x_0 + h_n) - f(x_0)}{h_n}$$

This limit (if it exists) is the slope of the tangent line to the graph of f at x_0.

This approach immediately raises two questions:

Will this limit always exist? No. It is possible that the limit does not exist, for example if the function is not continuous at x_0. Also the limit might not exist. For example, if the graph has a "corner" at x_0 (as the graph for $|x|$ has at $x_0 = 0$), the limit for a sequence of positive h_n will differ from the limit of a sequence of negative h_n. The formal definition thus needs to require this limit to exist:

Definition V.4: Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function that is continuous at x_0. If the limit

$$L = \lim_{h\to 0} \frac{f(x_0 + h_n) - f(x_0)}{h_n}$$

exists, then f is called *differentiable* at x_0. The value L (which is the slope of the tangent to the graph of f at x_0) is called the *derivative of f at x_0* and denoted by $f'(x_0)$.

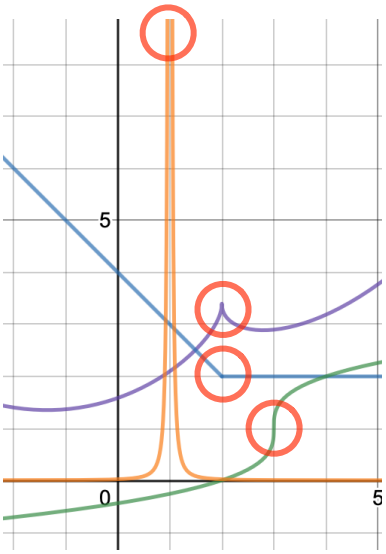


Figure V.9: Not differentiable

Note that in particular that a function that is not continuous at $x_0 \in \mathbb{R}$ is not differentiable at x_0. But there are continuous functions that are not differentiable. Figure V.9 shows some typical situations:

- The function tends to infinity (so might not even be continuous).
- The slope becomes vertical at a point.
- The graph of the function has a "kink", so there will not be a unique tangent.

How can we test for being differentiable, and calculate the value of the derivative? Testing differentiability can be technical and is not a focus in this course. We will see how this works in a few easy examples, but then introduce easier rules for finding derivatives.

V.6 Basic Derivatives, Polynomials

Let us consider a few easy cases. First, consider a constant function $f(x) = c$. In this case $f(x_0) = c = f(x)$ for every x and thus

$$\frac{f(x_0 + h_n) - f(x_0)}{h_n} = \frac{c - c}{h_n} = 0$$

and thus the limit exists, and is always equal

$$\lim_{n\to\infty} \frac{f(x_0+h_n)-f(x_0)}{h_n} = 0$$

Therefore: A constant function is differentiable at every x_0, and has derivative $f'(x_0) = 0$.

Next consider the function $f(x) = x$. Then

$$\frac{f(x_0+h_n)-f(x_0)}{h_n} = \frac{(x_0+h_n)-x_0}{h_n} = \frac{h_n}{h_n} = 1.$$

Again, the limit exists, and is always equal and we have that $f'(x_0) = 1$.

For the next cases, we will need the limit laws from Lemma IV.3: Consider the function $f(x) = x^a$ for an integer $a > 1$. Then (by the binomial theorem [4]) we have that

$$f(x_0+h_n) = (x_0+h_n)^a = \sum_{i=0}^{a} \binom{a}{i} x_0^{a-i} h_n^i = x_0^a + \sum_{i=1}^{a} \binom{a}{i} x_0^{a-i} h_n^i$$

and thus

$$\begin{aligned}
\frac{f(x_0+h_n)-f(x_0)}{h_n} &= \frac{(x_0+h_n)^a - x_0^a}{h_n} = \frac{x_0^a + \sum_{i=1}^{a}\binom{a}{i}x_0^{a-i}h_n^i - x_0^a}{h_n} \\
&= \frac{\sum_{i=1}^{a}\binom{a}{i}x_0^{a-i}h_n^i}{h_n} = \sum_{i=1}^{a}\binom{a}{i}\frac{x_0^{a-i}h_n^i}{h_n} = \sum_{i=1}^{a}\binom{a}{i}x_0^{a-i}h_n^{i-1} \\
&= a x_0^{a-1} + \sum_{i=2}^{a}\binom{a}{i}x_0^{a-i}h_n^{i-1}.
\end{aligned}$$

But every summand in the sum on the right hand side has a factor h_n and, since $\lim_{n\to\infty} h_n = 0$ we have $\lim_{n\to\infty}\sum_{i=2}^{a}\binom{a}{i}x_0^{a-i}h_n^{i-1} = 0$ and thus

$$\lim_{n\to\infty} \frac{(x_0+h_n)^a - x_0^a}{h_n} = a x_0^{a-1}.$$

Once more, the limit exists and is independent of the choice of sequence h_n. Thus the function $f(x) = x^a$ is differentiable at every x_0 and has the derivative $f'(x_0) = ax^{a-1}$. This result subsumes (for $a = 0$ and $a = 1$ the previous two, and one can show that it even holds if a is an arbitrary real number.

A similar application of the limit laws gives us that, if f and g are differentiable at x_0, so is $(f+g)(x)$ with derivative $f'(x_0)+g'(x_0)$, as is $(cf)(x)$ (for a constant $c \in \mathbb{R}$) with derivative $cf'(x_0)$.

Combining all of this, we have that a polynomial $f(x) = \sum_{i=0}^{n} a_i x^i$ is differentiable at every x_0 and has the derivative[5]

$$f'(x_0) = \sum_{i=1}^{n} a_i \cdot i \cdot x_0^{i-1}$$

[4] recall Pascal's triangle

[5] and this is all you need to remember from this section!

V.7 The Derivative as a Function

We have so far, for a given function f defined a derivative at a point x_0, as the slope of the tangent line to the graph of the function at point $(x_0, f(x_0)$ and denoted it as $f'(x_0)$. Assuming the function f is differentiable at every $x_0 \in \mathbb{R}$, we can calculate these derivative values $f'(x_0)$ for every x_0 and consider f' as a function itself (that assigns to x_0 the value $f'(x_0)$). This function is called the derivative (function) of f. Besides the name f', it also is denoted by $\frac{\mathrm{d}f}{\mathrm{d}x}$ or $\frac{\mathrm{d}}{\mathrm{d}x}f$, where the $\frac{\mathrm{d}}{\mathrm{d}x}$ is to be considered as a particular symbol (alluding to the difference quotient), not as an actual fraction. In Physics, also the notation $\dot{f}$ is sometimes used.

This derivative function assigns to every point x the slope of the tangent to the graph of f at the point $(x, f(x))$. We already know how to determine a formula for this derivative function for polynomials:

Example V.5: Let $f(x) = x^5 + x^4 - 20x^3 - 20x^2 + 64x + 64$. Figure V.10 shows tangents (green) to the graph (red)of f at some points. If we determine the slope of these tangents at all points $(x, f(x))$, we get the graph (blue) of the derivative $f'(x)$.

By the result of the previous section, we actually can calculate a formula for this derivative as $f'(x) = 5x^4 + 4x^3 - 60x^2 - 40x + 64$. Using this formula, we can calculate exact values of $f'(x)$ for arbitrary points x, as in the following table:

x	-4	-3	-2	-1	0	1	2	3	4
$f'(x)$	288	-59	-48	45	64	-27	-144	-83	480

Note that these values agree with the slopes of the tangents.

Note V.6: The notation $\frac{\mathrm{d}f}{\mathrm{d}x}$ serves another purpose, namely indicating with respect to *which* variable we take the derivative. If we look at multiple functions (or functions with a parameter), a function expression might contain other symbols, though only one will be considered the variable with respect to which we take the derivative. In the case of $\frac{\mathrm{d}f}{\mathrm{d}x}$ this variable is x,. though one also might write $\frac{\mathrm{d}f}{\mathrm{d}a}$ to indicate that the relevant variable is called a.

Ultimately, of course such functions with be considered as being of multiple variables, and the derivative with respect to one variable means we are considering only a "slice" of the function. More about this will come in multivariable calculus.

Here we just should note that e.g. $\frac{\mathrm{d}f}{\mathrm{d}a} = x^2$ for $f(x) = a \cdot x^2$, (considering a as the variable and x as constant), while $\frac{\mathrm{d}f}{\mathrm{d}x} = a$ and $\frac{\mathrm{d}f}{\mathrm{d}b} = 0$.

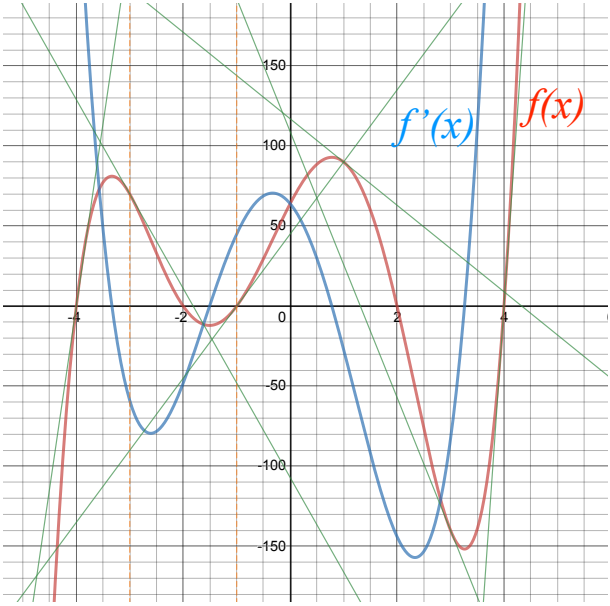


Figure V.10: Multiple tangents and the derivative

V.8 Derivatives of Elementary Functions

A function is called *elementary*[6], if it cannot be composed from other functions. These are typically functions that are useful in particular applications and that have their own key on the calculator, such as $\exp(x)$ or $\sin(x)$.

By careful inspection of the graphs of some of these functions it is possible to identify expressions for their derivatives, we give one example of this below. (The ultimate justification for some of these formulas will for us come through Taylor series VII.2.)

We note these derivative formulas in the following table:

[6] my dear Watson

Function $f(x)$	Derivative $f'(x)$
$\sin(x)$	$\cos(x)$
$\cos(x)$	$-\sin(x)$
$\exp(x) = e^x$	$\exp(x)$
$\log(x)$	$1/x$
$\arcsin(x)$	$\frac{1}{\sqrt{1-x^2}}$
$\arccos(x)$	$\frac{-1}{\sqrt{1-x^2}}$
$\arctan(x)$	$\frac{-1}{1+x^2}$
$\mathrm{bla}(x)$	$\frac{e^x-1}{x}$

This table also gives an indication what is special about the basis e in the exponential function $\exp(x) = e^x$, as for this basis the function equals its own derivative.

NOTE V.7: The reader will note a function $\mathrm{bla}(x)$, that she did not encounter before. It is a made-up function that is introduced solely to provide example problems for homework and exams that cannot be solved with standard computational tools.

EXAMPLE V.8: We have seen in Section V.6 a proof of the formula for differentiating polynomials. Here we give a similar argument for why $\dfrac{\mathrm{d}\sin(x)}{\mathrm{d}x} = \cos(x)$:

To establish this we need to show that

$$\lim_{h\to 0} \frac{\sin(x+h)-\sin(x)}{h} = \cos(x).$$

Recall from trigonometry that $\sin(x+h) = \sin(x)\cos(h) + \cos(x)\sin(h)$. Then some algebra and application of limit properties yields

$$\begin{aligned}\lim_{h\to 0} \frac{\sin(x+h)-\sin(x)}{h} &= \lim_{h\to 0} \frac{\sin(x)\cos(h)+\cos(x)\sin(h)-\sin(x)}{h}\\ &= \lim_{h\to 0}\left(\frac{\sin(x)\big(\cos(h)-1\big)}{h} + \frac{\cos(x)\sin(h)}{h}\right)\\ &= \sin(x)\lim_{h\to 0}\left(\frac{\cos(h)-1}{h}\right) + \cos(x)\left(\lim_{h\to 0}\frac{\sin(h)}{h}\right).\end{aligned}$$

It remains to evaluate the limits. Toward that end let us analyze the graphs of the two summands. Figure V.11, shows on the left the graph of $\frac{\cos(h)-1}{h}$ around $h = 0$, while the right side shows the graph of $\frac{\sin(h)}{h}$ around $h = 0$.

From these graphs we see that

$$\lim_{h\to 0}\frac{\cos(h)-1}{h} = 0, \quad \text{and} \quad \lim_{h\to 0}\frac{\sin(h)}{h} = 1$$

Putting it all together we find

$$\lim_{h\to 0} \frac{\sin(x+h)-\sin(x)}{h} = \sin(x)\cdot 0 + \cos(x)\cdot 1 = \cos(x)$$

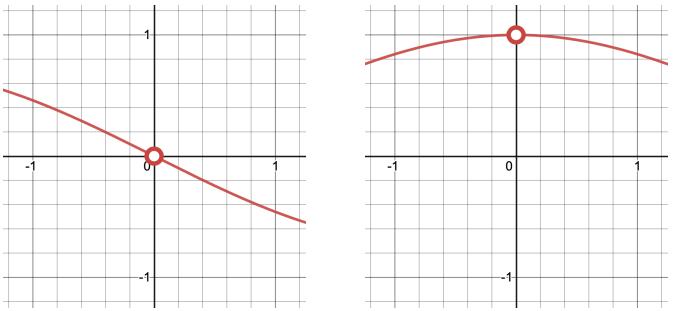


Figure V.11: Graphs of $\frac{\cos(h)-1}{h}$ and $\frac{\sin(h)}{h}$, for $h \neq 0$

which is what we wanted to show.

There is (of course) a formal, non-graphical, way to derive these limits which requires some as well as a theorem from advanced calculus, called the "squeeze theorem" (or "sandwich theorem"), which is beyond the scope of this text. Alternatively, we will see an argument involving Taylor series later in the text.

V.9 Differentiation Rules

When we studied functions, we looked at how we can build new functions from simpler ones. These constructions also allow us to write down formulas for the derivatives, as we will study in this section. We will aim to give a justification for each of these formulas, though will not always give a formal proof.

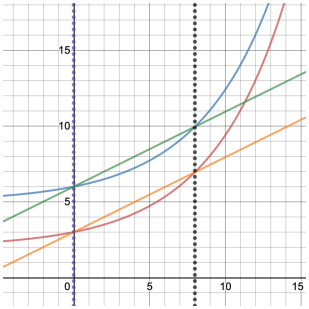
Let us start with some basic cases. (Some of these will end up later just being special cases of a more general rule, so you won't have to memorize all of this, but it is to help justifying the more general cases.)

We start with the basic transformations to a function's graph, as in section III.3. Let us look at the graph of a function under transformations, and see how this affects the slope of a secant. (Since the derivative is the slope of a tangent as limit of the secant slope, it will be transformed in the same way.)

In the examples in Figure V.12, we always transform an original function f (red) and its secant (orange) (from $x_0 = 0$ to $x_0 + h = 8$) to a new function p (blue, with corresponding green secant). Recall that the slope of the secant is given as

$$\frac{\Delta f}{\Delta x} = \frac{f(x_0+h)-f(x_0)}{x_0+h-x_0} = \frac{f(x_0+h)-f(x_0)}{h},$$

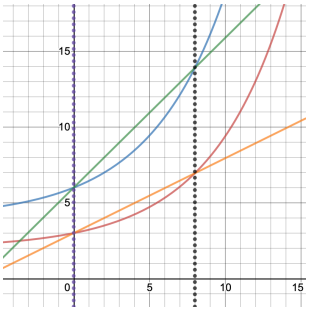
and that we thus only need to see how Δf and/or Δx transform. Note that vertical shift can be considered as a special case of addition, and that the rules for addition and for vertical scaling agree with what we've seen before in Section V.6 for polynomials.



Vertical Shift:
$p(x) = f(x)+3$. Then $\Delta p = \Delta f$ and Δx stays the same:

$$\frac{\Delta p}{\Delta x} = \frac{\Delta f}{\Delta x}$$

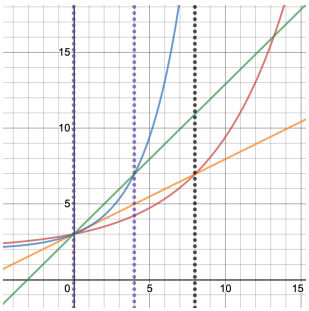
$$p'(x) = f'(x).$$



Vertical Scaling:
$p(x) = 2 \cdot f(x)$. Then $\Delta p = \Delta f$ and Δx stays the same.

$$\frac{\Delta p}{\Delta x} = 2\frac{\Delta f}{\Delta x}$$

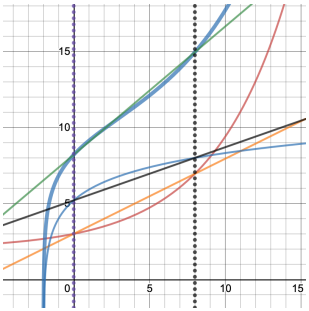
$$p'(x) = 2 \cdot f'(x).$$



Horizontal Scaling:
$p(x) = f(2 \cdot x)$. Then $\Delta p = \Delta f$, but Δx shrinks by a factor 2.

$$\frac{\Delta p}{\Delta x} = \frac{\Delta f}{\Delta x/2} = 2\frac{\Delta f}{\Delta x}$$

$$p'(x) = 2 \cdot f'(x).$$



Addition:
$p(x) = f(x) + g(x)$. Then $\Delta p = \Delta f + \Delta g$, and Δx stays the same.

$$\frac{\Delta p}{\Delta x} = \frac{\Delta f + \Delta g}{\Delta x}$$

$$p'(x) = f'(x) + g'(x).$$

Figure V.12: Basic transformations of the derivative

Product Rule

Next, lets look at the case of a product of functions, that is we have that $p(x) = f(x)\cdot g(x)$, and the function values change by $\Delta f = f(x_0+h)-f(x_0)$, respectively by $\Delta g = g(x_0+h)-g(x_0)$. Then (imagine f and g as sides of a rectangle, whose area is p and that changes as both sides change length):

$$\begin{aligned}\Delta p &= p(x_0+h)-p(x_0) = f(x_0+h)\cdot g(x_0+h) - f(x_0)g(x_0)\\ &= (\Delta f + f(x_0))\,(\Delta g + g(x_0)) - f(x_0)g(x_0)\\ &= \Delta f\Delta g + \Delta f\cdot g(x_0) + f(x_0)\cdot\Delta g + f(x_0)g(x_0) - f(x_0)g(x_0)\\ &= \Delta f\Delta g + \Delta f\cdot g(x_0) + f(x_0)\cdot\Delta g.\end{aligned}$$

If we now consider the value of the derivative as limit of the difference quotient

$$\begin{aligned}p'(x_0) &= \lim_{h\to 0}\frac{\Delta p}{\Delta x} = \lim_{h\to 0}\frac{\Delta f\Delta g + \Delta f\cdot g(x_0) + f(x_0)\cdot\Delta g}{\Delta x}\\ &= \lim_{h\to 0}\frac{\Delta f\Delta g}{\Delta x} + \lim_{h\to 0}\frac{\Delta f}{\Delta x}\cdot g(x_0) + f(x_0)\cdot\lim_{h\to 0}\frac{\Delta g}{\Delta x}\\ &= f'(x_0)g(x_0) + f(x_0)g'(x_0) + \lim_{h\to 0}\frac{\Delta f\Delta g}{\Delta x}\end{aligned}$$

and observe that in the remaining limit the numerator $\Delta f\Delta g$ shrinks *twice as fast* (because of the double-Δ) as the denominator Δx. This limit is thus equal to zero and we get (replacing x_0 by a general x) the *product rule*

$$p'(x) = f'(x)g(x)+f(x)g'(x), \qquad \frac{\mathrm{d}}{\mathrm{d}x}(f\cdot g)(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(x)\cdot g(x)+f(x)\cdot\left(\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right)$$

Chain Rule (= Composition Rule)

The case of composition of functions looks most complicated, as we have a change depending on change. Thus, let us first look at some easy examples of how such change accumulates, namely the case of polynomials of degree one.

For example, imagine that a bicyclist ascends a mountain road, at a rate of 3000 foot/hr, starting at ground level (5000 ft). Measuring in units of hours and 1000 ft, her altitude after x hours thus is $g(x) = 5 + 3x$. The temperature decreases by 2 degree per thousand foot and is, at altitude a, given as $f(a) = 80 - 2a$. The temperature at time x thus is $p(x) = f(g(x))$. But how does the temperature change per time unit, i.e. what is $p'(x)$? We can answer this in three different ways:

First, in this example, we can evaluate $p(x) = f(g(x)) = 80-2(5+3x) = 70-6x$ and calculate the derivative $p'(x) = -6$. (That is per time unit, the temperature decreases by 6 degrees.)

Secondly, we can observe that the composition consists of a horizontal scaling by a factor of 3, and a left shift by 5 units. The horizontal scaling (as in the example in the previous section) will increase the slope of secants and tangents by a factor of 3. (The horizontal shift means that the derivative values are shifted as the function is – in this example this will make no difference.) That is, we get the derivative of the function f, which is $f'(a) = -2$, evaluated at $a = g(x)$, and multiplied by a factor 3, resulting in $p'(x) = -6$ as before.

Thirdly, we can work with the definition of the derivative as limit of the difference quotient (and this will work not just in this particular example). We have that

$$\frac{\Delta p}{\Delta x} = \frac{\Delta p}{\Delta g} \cdot \frac{\Delta g}{\Delta x}$$

where we have set $\Delta g = g(x_0 + h) = g(x_0)$ and

$$\Delta p = p(x_0 + h) - p(x_0) = f(g(x_0 + h)) - f(g(x_0)).$$

Applying the limit rule for the product, we get

$$\begin{aligned} \lim_{h\to 0} \frac{\Delta p}{\Delta x} &= \lim_{h\to 0} \left(\frac{\Delta p}{\Delta g} \cdot \frac{\Delta g}{\Delta x} \right) = \left(\lim_{h\to 0} \frac{\Delta p}{\Delta g} \right) \cdot \underbrace{\left(\lim_{h\to 0} \frac{\Delta g}{\Delta x} \right)}_{=g'(x_0)} \\ &= \lim_{h\to 0} \frac{f(g(x_0 + h)) - f(g(x_0))}{g(x_0 + h) - g(x_0)} \cdot g'(x_0) \end{aligned}$$

We now replace $g(x_0 + h)$ by $g(x_0) + h$. This of course is not true in general, but one can show[7] that for small values of h this is a good enough approximation so that the value of the limit stays the same. Thus, continuing the equations we have

$$\begin{aligned} \lim_{h\to 0} \frac{\Delta p}{\Delta x} &= \lim_{h\to 0} \frac{f(g(x_0) + h) - f(g(x_0))}{g(x_0) + h - g(x_0)} \cdot g'(x) \\ &= f'(g(x_0)) \cdot g'(x_0) \end{aligned}$$

with the last equation following from substituting $g(x_0)$ in place of x_0. We thus get the *chain rule*:

$$\frac{\mathrm{d}(f \circ g)}{\mathrm{d}x}(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(g(x)) \cdot \frac{\mathrm{d}g}{\mathrm{d}x}(x)$$

In our example we have that $f'(x) = -2$ and $g'(x) = 3$, thus $(f \circ g)'(x) = -6$. But this chain rule can do many other functions.

[7] This is done in Junior level mathematics classes, called "Analysis"

If we have that $p = f \circ g$, we can also write (mirroring the way we did prove the result, and making it look as if it was basic arithmetic with differentials:

$$\frac{\mathrm{d}p}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}g} \cdot \frac{\mathrm{d}f}{\mathrm{d}x}$$

In the context of machine learning, the (multi-variable) version of the chain rule goes under the name of *back propagation*.

Another application of the chain rule is in estimating the error of the value of a function composition $f \circ g$, at a point x, given an error Δx in x. We get from the difference quotient that the error in the value $f(g(x))$ is

$$\Delta(f \circ g) \sim (f'(g(x))g'(x))\,\Delta x.$$

Examples of using the Chain Rule

Since the chain rule is so important we give a number of examples where it is applied:

Example V.9: Consider the function $f(x) = (3x^2+5)^{1014}$. In principle could compute $f'(x)$ by expanding but that would be impossible to do by hand. Instead we apply the chain rule:

$$f'(x) = 1014(3x^2+5)^{1013} \cdot \frac{d}{dx}(3x^2+5) = 6 \cdot 1014x + (3x^2+5)^{1013}.$$

Next consider the function $g(x) = \sin(\ln(x^2))$. The function g is a composition of *three* simpler functions sin, ln, and x^2. To compute $g'(x)$ we apply the chain rule first for the composition of $\ln(x^2)$ with $\sin(x)$:

$$\begin{aligned} g'(x) &= [\sin(\ln(x^2))]' \\ &= \cos(\ln(x^2)) \cdot \frac{d}{dx}\ln(x^2). \end{aligned}$$

How do we compute $\frac{d}{dx}\ln(x^2)$? Simply by applying the chain rule again.

$$\begin{aligned} \cos(\ln(x^2)) \cdot \frac{d}{dx}\ln(x^2) &= \cos(\ln(x^2)) \cdot \frac{1}{x^2} \cdot \frac{d}{dx}x^2 \\ &= \cos(\ln(x^2)) \cdot \frac{1}{x^2} \cdot 2x. \end{aligned}$$

Example V.10: Next, let's think about computing the derivative of exponentiation functions $h(x) = b^x$ where $b > 0$ is some arbitrary constant. We know how to find derivatives if $b = e$ (remember $(e^x)' = e^x$), but what is $b = 2$ or $b = \pi$ or something else? Then we simply rewrite the function as a composition with the exponential function:

$$h(x) = b^x = \left(e^{\ln(b)}\right)^x = e^{\ln(b)x}.$$

Computing the derivative $h'(x)$ now is simply the chain rule:

$$h'(x) = e^{\ln(b)x} \cdot (\ln(b)x)' = e^{\ln(b)x} \ln(b) = b^x \ln(b).$$

In words: the derivative an exponential function is just itself multiplied by the natural logarithm of its base.

Example V.11: Finally, let's compute the derivative of $j(x) = x^x$. The power rule doesn't apply since that only works for functions like x^2 or x^π, that is, x raised to some power. The rule for exponential functions we just derived doesn't apply either, because it assumed b was a constant. The trick is to take the natural logarithm of both sides of this equation to get $\ln(j(x)) = x \ln(x)$. Differentiating the right hand side yields

$$\frac{d}{dx} x \ln(x) = \ln(x) + x \cdot \frac{1}{x} = \ln(x) + 1$$

by the product rule. But if we differentiate the left hand side using the chain rule we find that

$$\frac{d}{dx} \ln(j(x)) = \frac{1}{j(x)} \cdot j'(x) = \frac{j'(x)}{x^x}.$$

This means that we have $j'(x)/x^x = \ln(x)+1$. Therefore $j'(x) = x^x(\ln(x)+1)$. This method of first taking natural logarithms, then differentiate, and finally solve for the derivative is applicable more generally and is called *logarithmic differentiation*.

Finally, the chain rule justifies the derivatives of the inverse functions of $\sin(x)$ and $\exp(x)$ that were given in V.8.

Write $\log(\exp(x)) = x$ and take the derivative on both sides. With the chain rule we get.

$$\log'(\exp(x)) \cdot \exp(x) = 1.$$

We substitute $y = \exp(x)$ for $\log'(y) \cdot y = 1$, and solve as

$$\log'(y) = \frac{1}{y}, \qquad \log'(x) = \frac{1}{x}.$$

Similarly, we derive both sides of $\arcsin(\sin(x)) = x$ to get

$$1 = \arcsin'(\sin(x)) \cdot \cos(x) = \arcsin'(\sin(x)) \cdot \sqrt{1 - \sin^2(x)}.$$

(recall that $\sin^2(x) + \cos^2(x) = 1$). Setting $y = \sin(x)$ this gives

$$1 = \arcsin'(y) \cdot \sqrt{1 - y^2}.$$

and thus

$$\arcsin'(x) = \frac{1}{\sqrt{1 - x^2}}.$$

The quotient rule

We can write the quotient of two functions $f(x)/g(x)$ as a product

$$f(x) \cdot (g(x))^{-1}.$$

The chain rule then gives us that

$$\frac{\mathrm{d}}{\mathrm{d}x}(g(x))^{-1} = -(g(x))^{-2} \cdot g'(x) = -\frac{g'(x)}{g(x)^2}$$

and thus the *quotient rule*

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{f(x)}{g(x)} = f'(x)(g(x))^{-1} + f(x) \cdot \left(-\frac{g'(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$$

The Derivative Algorithm

All the differentiation rules we have seen so far can be combined into an algorithm that provides a purely mechanical way to calculate the derivative of a function given by a formula. (Indeed, it is a no-too-hard exercise to implement this in a programming language of your choice. The most difficult part will actually be to parse a string representing a formula, as to be able to take it apart into its constituent parts.)

procedure DERIVATIVE(f) ▷ f is a function
  Find the outermost way how f is constructed
  if f is a power: $f(x) = x^n$ **then**
    ▷ n can be a fraction (root) or negative (reciprocal)
    return $n \cdot x^{n-1}$
  else if f is a ("scalar") multiple: $f(x) = c \cdot g(x)$ **then**
    return $c \cdot$ Derivative(g)
  else if f is a sum: $f(x) = g(x) + h(x)$ **then**
    ▷ Consider difference as sum $g(x) + (-1) \cdot h(x)$
    $dg :=$ Derivative(g)
    $dh :=$ Derivative(h)
    return $dg + dh$ ▷ Sum rule: $(g+h)' = g' + h'$
  else if f is a product: $f(x) = g(x) \cdot h(x)$ **then**
    $dg :=$ Derivative(g)
    $dh :=$ Derivative(h)
    return $dg \cdot h + g \cdot dh$ ▷ Product rule: $(g \cdot h)' = g' \cdot h + g \cdot h'$
  else if f is a quotient: $f(x) = g(x)/h(x)$ **then**
    $dg :=$ Derivative(g)
    $dh :=$ Derivative(h)
    return $\frac{dg \cdot h - g \cdot dh}{h^2}$ ▷ Quotient rule: $(g/h)' = (g' \cdot h - g \cdot h')/h^2$
  else if f is a composition: $f(x) = g(h(x))$ **then**
    $dg :=$ Derivative(g)

```
        dh := Derivative(h)
        return dg(h(x)) · dh(x)                          ▷ Chain Rule:
(g(h(x)))′ = g′(h(x)) · h′(x).
    else if f is an exponentiation: f(x) = g(x)^h(x) then
        Write f(x) = exp(log(g(x)^h(x))) = exp(h(x) · log(g(x))).
        return Derivative(exp(h(x) · log(g(x))))
    else
        Look f up in the table in Section V.8, and return the associated derivative.
    end if
end procedure
```

V.10 Higher Derivatives

The derivative of a function is again a function on its own, and thus (assuming the function behaves well) has its own derivative. We write

$$f''(x) = (f'(x))' = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}^2}{\mathrm{d}x^2}f(x)$$

for this derivative. Similarly we can define third derivatives etc. Since the notation of multiple dashes can get overly complicated, one also writes

$$f^{(n)}(x) = \frac{\mathrm{d}^n}{\mathrm{d}x^n}f(x)$$

for the n-th derivative.

With the derivative describing a change, the second derivative describes the change of the change. The standard example of this in the real live would be a function that gives the position of an object over time. Its derivative indicates how fast the position changes over time – that is the velocity. And the second derivative indicates how fast the velocity changes. That is the acceleration.

Example V.12: Many familiar functions like polynomials, $\sin(x)$, and e^x have the property of being *infinitely differentiable* (meaning you can take the derivatives of these functions as many times as you'd like) but not all functions have this property. Consider the piecewise function

$$f(x) = \begin{cases} x^2 & \text{if } x \geq 0 \\ -x^2 & \text{if } x < 0 \end{cases}$$

whose graph is given in Figure V.13, left.

The graph of f is formed from a regular parabola except that the left half has been mirrored across the x-axis. Since $f(x) = x^2$ for $x > 0$ we know $f'(x) = 2x$ for $x > 0$. Also since $f(x) = -x^2$ for $x < 0$ we know $f'(x) = -2x$ for $x < 0$. Using the limit definition of a derivative you can verify that $f'(0) = 0$ (or just observe from

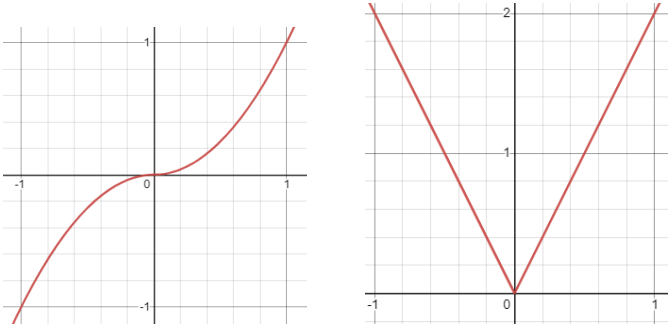


Figure V.13: A function that is only differentiable once, and its derivative

the graph that the tangent line at $x = 0$ is horizontal). Then we can express $f'(x)$ piecewise

$$f'(x) = \begin{cases} 2x & \text{if } x \geq 0 \\ -2x & \text{if } x < 0 \end{cases}.$$

Actually there's a simpler way to express $f'(x)$: $f'(x) = 2|x|$. The graph of this function, Figure V.13, right, has a "kink" at $x = 0$, which implies that it is not differentiable there. Of course, it is infinitely differentiable away from $x = 0$.

Interestingly, this can become far more complicated. There are functions $f(x)$ such that

- $f(x)$ is differentiable everywhere ($f'(x)$ exists for all x in the domain of f).
- $f'(x)$ is continuous everywhere.
- $f'(x)$ is not differentiable anywhere in it's domain.

However, such functions are extremely unlikely to arise in any practical context, but just reflect the fact that our definition of a function is rather general. Indeed, the fact that such functions exist was only established hundreds of years after the inception of calculus.

Chapter

VI Applications of Differentiation

The derivative of a function describes how the values of a function change, and this indication of change can be used to understand a function better or to find x-values at which the function behaves in a particular way (such as having maxima, minima, or zeroes).

One way to think about this is if you imagine the graph of a function to depict the track (viewed from above) along which a bicycle rode. There are places where the rider performed actions to change the way she rode, and other places where she basically continued in the same way as before. These are the places we want to identify.

In many textbooks this topic goes under the name of "curve sketching", which was a main application before the advent of easily accessible plotting tools. However, even if we do not want to plot a function by hand, understanding how a function and its derivatives are related is useful for applications.

VI.1 Increasing and Decreasing

The value of the derivative is the slope of the tangent line to the graph. That means that the graph goes up when the derivative is positive and goes down when the derivative is negative. More formally, using the same language as with sequences:

Definition VI.1: Let $D \subset \mathbb{R}$ and $f\colon D \to \mathbb{R}$ a differentiable function and $x_0 \in D$. We say that f is

increasing **at** x_0 if $f(x_0) \geq 0$.

strictly increasing **at** x_0 if $f(x_0) \gneqq 0$.

decreasing **at** x_0 if $f(x_0) \leq 0$.

strictly decreasing **at** x_0 if $f(x_0) \lneqq 0$.

We say that f is increasing (& c.) on D, if it its increasing (& c.) at every $x_0 \in D$

Note that f is increasing on D if $f(a) \geq f(b)$ (strictly, if $f(a) \gneqq f(b)$) whenever $a > b$. We have analogous definitions for decreasing. (and

This allows us to characterize when differentiable functions are one-to-one: A function that is strictly increasing must be one-to-one, since two different points in the domain must have one of them larger (a, say), and then $f(a) > f(b)$, in particular $f(a) \neq f(b)$.

Again the same holds for a strictly decreasing function. We also see easily that a (continuous) function that first increases and then decreases must reach the same value twice, and thus cannot be one-to-one.

The only further case that is permitted is if $f'(x_0) = 0$ at individual points (as for x^3 at $x = 0$). We state this (without a formal proof, that would be somewhat technical):

LEMMA VI.2: Let $f: D \to \mathbb{R}$ be a differentiable function. Then f is one-to one, if and only if f is increasing on D (or decreasing on D), and strictly increasing (respectively strictly decreasing) at all but finitely many points in D.

VI.2 The Shape of a Curve

Knowing where a function increases and decreases also helps us to describe the overall shape of the graph of a function. Consider for example the graph[1] of a function $f: \mathbb{R} \to \mathbb{R}$ in Figure VI.1:

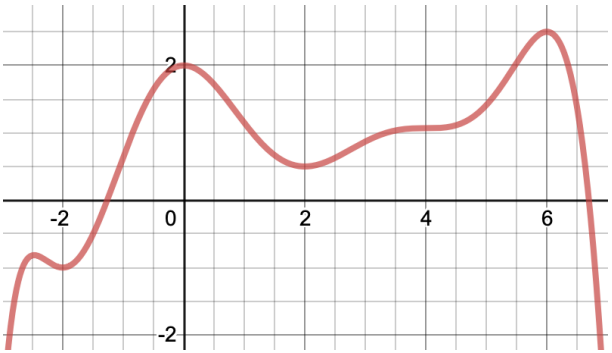


Figure VI.1: The graph of a function

[1] In case someone cares, it is the horribly looking horribly looking function $1109/31850496x^9 - 63797/79626240x^8 + 114341/19906560x^7 - 61627/9953280x^6 - 823919/9953280x^5 + 1137739/4976640x^4 + 68581/207360x^3 - 5731/4320x^2 + 2$

We calculate the derivative and look at where the derivative is negative (that is the original function is decreasing):

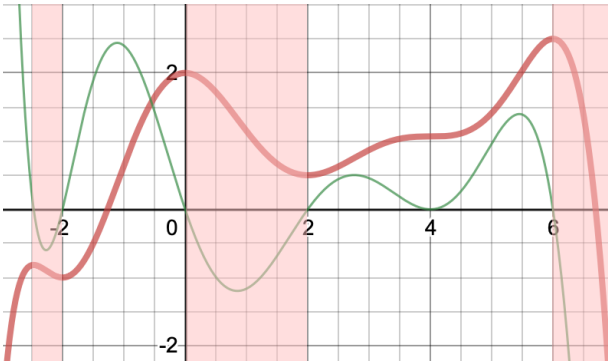


Figure VI.2: Intervals where the function is decreasing

Note that at the points where increasing/decreasing changes (that is where the derivative is zero), the function has a local maximum or minimum. "Local" here means that it is larger than at any other point in the neighborhood, though it might be even larger somewhere away[2].

DEFINITION VI.3: A point x_0, where $f'(x_0) = 0$ is called a *critical point* for f.

We have seen that (local) maxima and minima occur only at critical points, namely a maximum if the derivative changes from positive to negative, and a minimum if the derivative changes from negative to positive.

In the example we have critical points at -2.5,0,and 6 (maxima) as well as at -2 and 2 (minima).

But there is a further critical point, namely we have $f'(4) = 0$. Here the derivative becomes zero, but does not change its sign (i.e. stays nonnegative). What happens is that the function briefly stops growing[3] but then immediately grows again. Such a point is called a *saddle point*.

We can distinguish easily between the three kinds of critical points, if we also look at the second derivative. At a maximum, the derivative was positive and becomes negative, so the second derivative must be negative. At a minimum, with an analogous argument, the second derivative is positive. And at a saddle point, the derivative becomes zero but does not change its sign. That means a saddle point is a local minimum or maximum of the derivative, and thus must have the second derivative zero as well.

[2] In the same way as the tallest person in town is not guaranteed to be the tallest person in the country

[3] like a mountain climber who briefly stops to catch a deep breath

Turning Points

Let us now look at the general impact of the second derivative. Figure VI.3 now shows first (green) and second (blue) derivative, as well as an indication of the places where the derivatives become zero.

Definition VI.4: A point x_0, where the second derivative is zero, $f''(x_0) = 0$ is called a *turning point* (or *inflection point*) for f.

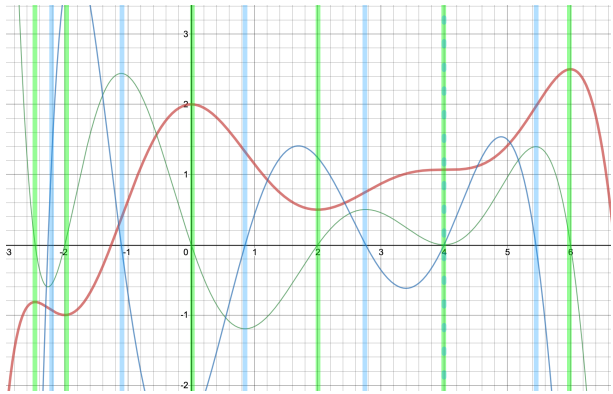


Figure VI.3: Critical and Turning points

We see that when the second derivative is positive, the derivative becomes larger and the growth of the function increases. The graph of the function is therefore *convex*, that is it turns left. (Some textbooks use "concave down" in place of "convex".) Similarly, the graph is *concave*, that is it turns right, if the second derivative is negative. At turning points, when the second derivative changes its sign, the graph of the function changes from left turn to right turn, or vice versa.

Let us summarize these observations on the shape of f in table VI.1. The middle column describes critical points, the second and fourth row turning points.

Using this table, we now can describe the shape of the graph of a function, though it does not allow us to determine absolute function values or decide on tie-breaks amongst local maximal/minima which ones are larger.

Note that there is a symmetry between a function and its negative (flipped upside down), and both will have the same critical points and turning points. We thus need to have at least one information about a derivative value being positive or negative to be able to distinguish between these two cases.

Let us do this for the function in the example. Using the zeroes of the two derivatives (the colored vertical lines in Figure VI.3), we split the domain from -3 to 7 into intervals. We indicate the type as **C**ritical or **T**urning. Note how two

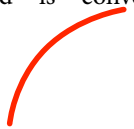
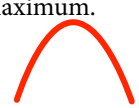
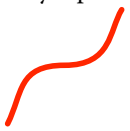
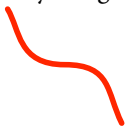
	$f'(x_0) > 0$	$f'(x_0) = 0$, *Critical Point*	$f'(x_0) < 0$
$f''(x_0) < 0$	f increases and is convex. 	f reaches a maximum. 	f decreases and is convex. 
$f''(x_0) = 0$ and f'' changes from $-$ to $+$. (I.e. $f'''(x) > 0$) *Turning Point*	f increases and has a turning point from right to left 	f has a saddle point, increasing. f' stays positive 	f decreases and has a turning point from right to left 
$f''(x_0) = 0$ constant zero	f increases straight. f' is constant. 	f is constant.	f decreases straight. f' is constant. 
$f''(x_0) = 0$ and f'' changes from $+$ to $-$. (I.e. $f'''(x) < 0$) *Turning Point*	f increases and has a turning point from left to right 	f has a saddle point, decreasing. f' stays negative 	f decreases and has a turning point from left to right 
$f''(x_0) > 0$	f increases and is concave. 	f reaches a minimum. 	f decreases and is concave. 

Table VI.1: The local shapes of a curve

critical points are separated by turning points.

x	Type	$f'(x)$	$f''(x)$
		+	−
−2.5	C	0	−
		−	−
−2.2	T	−	0
		−	+
−2	C	0	+
		+	+
−1.1	T	+	0
		+	−
0	C	0	−
		−	−
0.85	T	−	0

x	Type	$f'(x)$	$f''(x)$
		−	+
2	C	0	+
		+	+
2.75	T	+	0
		+	−
4	CT	0	0
		+	+
5.45	T	+	0
		+	−
6	C	0	−
		−	−

We now can (Figure VI.4) find the corresponding local shapes in the table and compose them to approximate the overall shape of the function graph.

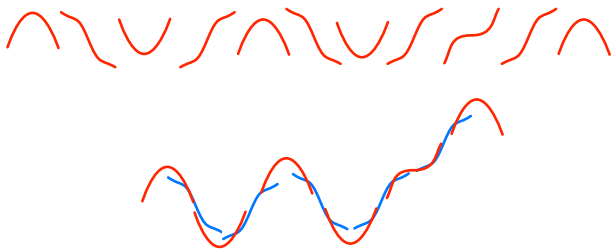


Figure VI.4: Composing local shapes

The advantages of using the derivative over other potential methods are:

- No plot is needed.
- One can find the relevant points exactly
- Identifying a point where a derivative is zero is often easier than identifying the point of a maximum or minimum[4]
- One can solve the problem, even if the definition of the function involves other variable parameters (and one thus cannot plot).

[4] Imagine a pool in which the water level is changing. It is easier to see when the water level is crossing a threshold than when it is maximal.

EXAMPLE VI.5: For a more complicated example, consider the function $f(x) = x^2e^{-x}$. Let's find the critical points, turning points, and determine where the function is increasing/decreasing and where it is concave up/down.

First we find the critical points. To do that we need to find the derivative $f'(x)$:

$$\begin{aligned} f'(x) &= \left(x^2e^{-x}\right)' \\ &= (x^2)'e^{-x} + x^2(e^{-x})' \quad \text{(product rule)} \\ &= 2xe^{-x}x^2 + x^2(-e^{-x}) \quad \text{(power rule on } x^2\text{, chain rule on } (e^{-x})') \\ &= (2x - x^2)e^{-x} \quad \text{(simplification)}. \end{aligned}$$

Now we solve $f'(x) = (2x - x^2)e^{-x} = 0$. The factor e^{-x} is never zero so we just need to find the zeros of $2x - x^2$, which are $x = 0$ and $x = 2$. These are the (only!) critical points.

Now we figure out where $f(x)$ is increasing and decreasing. This is determined by the sign of $f'(x)$, which must stay the same outside the critical points. We thus consider $f'(x)$ on the intervals $(-\infty, 0), (0, 2), (2, \infty)$. From a calculator we find $f'(-1) \approx -8.15$, so f' is negative at $x = -1$, and thus f is decreasing for $x < 0$. Similarly, we calculate $f'(1) \approx 0.3679$ and thus have f increasing for $0 < x < 2$. And $f'(3) \approx -0.1494$ so $f(x)$ must be decreasing again for $2 < x$. This together tells us that f has a local minimum at $x = 0$ and a local maximum at $x = 2$.

Now we find turning points and think about concavity. Toward that end we will need $f''(x)$:

$$\begin{aligned} f''(x) &= ((2x - x^2)e^{-x})' \\ &= (2x - x^2)'e^{-x} + (2x - x^2)(e^{-x})' \quad \text{(product rule)} \\ &= (2 - 2x)e^{-x} + (2x - x^2)(-e^{-x}) \quad \text{(power rule and chain rule)} \\ &= (x^2 - 4x + 2)e^{-x} \quad \text{(simplification)}. \end{aligned}$$

Again the factor e^{-x} is never zero and we thus only need to consider the polynomial $x^2 - 4x + 2$. Its roots are $x = 2 - \sqrt{2} \approx 0.586$ or $x = 2 + \sqrt{2} \approx 3.414$. These are the turning points. A calculator gives us that $f''(0) = 2$, $f''(2) \approx -.2707$, $f''(4) \approx 0.0366$, thus $f(x)$ is concave up for $x < 2 - \sqrt{2}$, concave down for $(2 - \sqrt{2} < x < 2 + \sqrt{2})$ and concave up for $2 + \sqrt{2} < x$. A look at the graph of $f(x)$ in Figure VI.5 illustrates this.

VI.3 Optimization

An important application for finding maxima/minima is in optimization. We are considering a problem that involves a parameter that can be chosen, and have a cost function whose value we want to minimize (or a "gain" function we want to maximize), that means we look for the maximum or minimum of a function.

As we have seen, these extrema must happen at critical points. We just need to decide which ones are maxima, which ones minima, and, if there are several, which

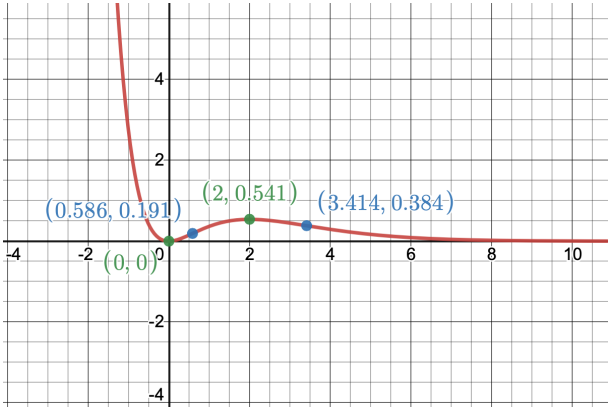


Figure VI.5: The graph of $x^2 \cdot e^{-x}$ with critical points and turning points.

ones are best. Furthermore, if the set of valid parameter values if not the whole real axis, it is possible that a maximum/minimum is reached at the maximal/minimal parameter values: The function $f(x) = x + 1$ has, for values $2 \leq x \leq 5$ the maximal value 6 at $x = 5$, though this is not even a critical point.

Let us consider this in an example:

We have a fence of 100 units length and want to use it to surround a rectangular area that is as large as possible. Assuming we use the whole fence, we denote the length of one side (and the opposite side) by x, then the other two sides[5] both have length $50 - x$.

The area of the rectangle then is $f(x) = x \cdot (50 - x) = 50x - x^2$ square units, and the permitted values for x are from 0 to 50 (in both end cases the rectangle degenerates to a line).

To find the critical points, we calculate $f'(x) = 50 - 2x$ and find $x = 25$ is the only critical point. Since $f(25) = 25^2 = 625 > 0$ and $f(0) = 0 = f(50)$ this is a maximum and the global maximum. The best fence thus has one side of length 25 (and thus the other side also of length $50 - 25 = 25$).

For a more involved problem, imagine that the manufacturer *Consolidated Widgets* decides to produce a new exquisite widget, using their trademark *magic dust*. Due to the interaction of the magic dust with the other ingredients, the manufacturing cost of a widget containing x units of dust is $f(x) = x^3 - 50x^2 + 700x - 1000$ doubloons. To be exquisite, the widget also needs to contain at least 5 units of dust, and can contain at most 50 units. For which amount of dust is the manufacturing cost per widget minimal?

[5] It is $50 - x$ and not $100 - x$ as there are two sides in each direction.

Again, we consider the derivative $f'(x) = 3x^2 - 100x + 700 = (x-10)(3x-70)$. The critical points thus are 10 and $70/3 \sim 23.33$. We now evaluate $f(x)$ at the critical points, as well as the end points:

x	5	10	23.33	50
$f(x)$	1375	2000	814.81	34000

We find that the minimal manufacturing cost (namely 814.81) is obtained when using 23.33 units of dust per widget (while the maximum cost would be for 50 units).

VI.4 Newton's Method

We go back to the problem already studied in Section V.2 on how to find (an approximation of) x_0 such that $f(x_0) = 0$. We had seen that the method of halving intervals required $\log_2(10) \sim 3.3219$ steps in average for each extra digit of accuracy in the approximation – in the example we got an error $< 10^{-6}$ (i.e. 6 decimal digits) after 20 iterations. This is far too bad for many practical applications, and we want to try better.

The problem of the interval halving method is that it always goes to the middle of the interval, while the actual point x_0 with $f(x_0) = 0$ might be far closer to the a side of the interval. Figure VI.6, for example, shows four different functions on the interval $[2,5]$, all of which have $f(2) = 1$ and $f(5) = -1$ (and thus a zero in the interval), but the placement of the zeroes is quite different. The left or right side of the interval thus might not be a much better approximation of the root.

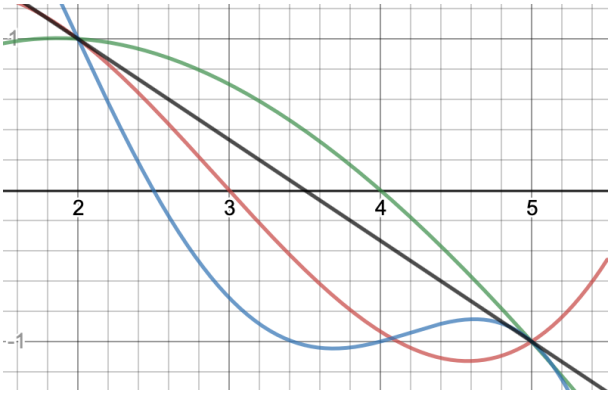


Figure VI.6: Different placement of zeroes

The idea of the Newton method is to not just consider halved intervals as a good approximation, but use the tangent line (whose slope indicates how steep the

function is) to find a better approximation[6]. That is, if we have a point x_n as approximation of the root, we follow the tangent to f at x_n (whose slope is $f'(x_n)$) to the point where it intersects the x-axis, and take that x-value as the next approximation x_{n+1}. Figure VI.7 illustrates this process.

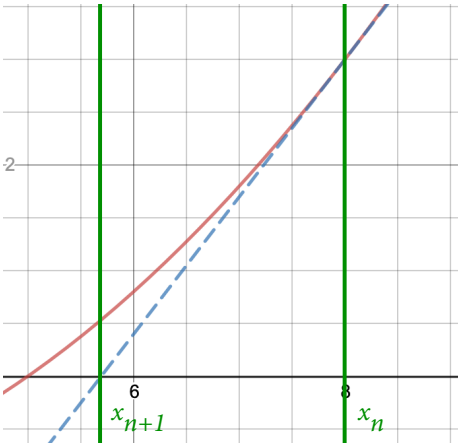


Figure VI.7: The Newton Method

To find a formula for x_{n+1} observe that the tangent line goes through the points $(x_{n+1}, 0)$ and $(x_n, f(x_n))$ and has slope $f'(x_n)$. Thus

$$f'(x_n) = \frac{f(x_n) - 0}{x_n - x_{n+1}}$$

and we solve (assuming that $f'(x_0) \neq 0$) for

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Let us look at this in the same example as we did for the halving method. We take $f(x) = \sin(x)$ and start at $x_0 = 2$. The formula gives us the recursion

$$x_{n+1} = x_n - \frac{\sin(x_n)}{\cos(x_n)}$$

and we calculate values as

$$x_0 = 2, \quad x_1 = 4.185039863, \quad x_2 = 2.467893675, \quad x_3 = 3.266186277$$
$$x_4 = 3.140943912, \quad x_5 = 3.141592654, \quad x_6 = 3.141592654$$

[6] That is, to find a valley, we go downhill

Note that after 5 iterations (instead of 20 before) we approximated the zero to 6 digits. Even more, at that point we actually have it approximated already to 9 digits!

This is not just happenstance. One can show that, as long as the starting value is not too far away for the root, the Newton method converges quadratically, that (up to a constant) the error after each step is bounded by the square of the previous error. That means that the number of correct digits after each step will increase by a factor (in the above example it seems to be roughly 2), while for the interval halving method it only increased by one digit every three steps. This shows that the Newton method is far more powerful than interval halving.

VI.5 Indefinite Limits and L'Hospital's rule

For reasons we shall see in the next section, we want to be able to calculate limits of quotients.

We have see the limit rule for quotients that (in a version for functions) gives us that for a number a (or $a = \infty$) and functions $f, g\colon \mathbb{R} \to \mathbb{R}$ we have

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)},$$

if the fraction of limits on the right hand side exists. But what if numerator and denominator on the right hand side are both 0? Or both ∞? An answer to this question is given by a test that is called[7] L'Hospital's rule (pronounced *lowpytaal's* rule).

THEOREM VI.6: Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be differentiable functions and $a \in \mathbb{R}$ or $a = \infty$ such that $\lim_{x\to a} f(x) = 0 \lim_{x\to a} g(x)$, respectively $\lim_{x\to a} f(x) = \infty \lim_{x\to a} g(x)$. Then, if $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ exists, we have that

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

The reader surely will have noted that this quotient of derivatives is *not* the derivative of the quotient!

Proof: We give a proof only for the case that $a \in \mathbb{R}$ and $f(a) = 0 = g(a)$ and that $f'(x)$ and $g'(x)$ are continuous at a. (The other cases are significantly harder.) Then

$$\begin{aligned}
\lim_{x\to a} \frac{f(x)}{g(x)} &= \lim_{x\to a} \frac{f(x)-0}{g(x)-0} = \lim_{x\to a} \frac{f(x)-f(a)}{g(x)-g(a)} \\
&= \lim_{x\to a} \frac{\frac{f(x)-f(a)}{x-a}}{\frac{g(x)-g(a)}{x-a}} = \frac{\lim_{x\to a} \frac{f(x)-f(a)}{x-a}}{\lim_{x\to a} \frac{g(x)-g(a)}{x-a}} \\
&= \frac{f'(a)}{g'(a)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.
\end{aligned}$$

[7] Not after the inventor, but after a popularizer of it

□

For example, we have that

$$\lim_{x\to\infty} \frac{x-1}{x^2-1} = \lim_{x\to\infty} \frac{1}{2x} = 0$$

Note that a situation of ∞/∞ can give us finite (nonzero) limits. For example

$$\lim_{x\to\infty} \frac{3x-5}{73-15} = \lim_{x\to\infty} \frac{3}{7} = \frac{3}{7}$$

It is possible to apply L'Hospital's rule multiple times. In the following example, the first application of L'Hospital's rule still leaves us with an ∞/∞ case, but the second one then gives the value:

$$\lim_{x\to\infty} \frac{4x^2+3x-1}{5x^2-x+1} = \lim_{x\to\infty} \frac{8x+3}{10x^2-1} = \lim_{x\to\infty} \frac{8}{10} = \frac{4}{5}.$$

More generally, if we have a quotient of polynomials, we need to apply L'Hospital's rule as many times as (the smaller of) the degrees of numerator and denominator:

THEOREM VI.7: Let $f(x), g(x)$ be nonzero polynomials. Then limit of the quotient

$$\lim_{x\to\infty} \frac{f(x)}{g(x)}$$

equals:

0 if the degree of $g(x)$ is larger than the degree of $f(x)$.

$\pm\infty$ if the degree of $f(x)$ is larger than the degree of $g(x)$

a/b if the degree of f equals the degree of g, and a is the leading coefficient of f and b the leading coefficient of g.

The same statement obviously also holds with n in place of x and explains the limits of sequences we observed earlier.

EXAMPLE VI.8: Note that even if L'Hospital's rule is applicable, there is no guarantee that the result is helpful for solving the problem. Consider the limit

$$\lim_{x\to\infty} \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

It is easy to see that $e^x - e^{-x} \to \infty$ and $e^x + e^{-x} \to \infty$ as $x \to \infty$, so it is valid to apply L'Hospital's rule here. But if we do, we find

$$\lim_{x\to\infty} \frac{e^x - e^{-x}}{e^x + e^{-x}} = \lim_{x\to\infty} \frac{\frac{d}{dx}(e^x - e^{-x})}{\frac{d}{dx}(e^x + e^{-x})} = \lim_{x\to\infty} \frac{e^x + e^{-x}}{e^x - e^{-x}}.$$

This limit really doesn't seem any simpler. Indeed, if we try L'Hospital's rule again we find

$$\lim_{x\to\infty}\frac{e^x+e^{-x}}{e^x-e^{-x}}=\lim_{x\to\infty}\frac{\frac{d}{dx}\left(e^x+e^{-x}\right)}{\frac{d}{dx}\left(e^x-e^{-x}\right)}=\lim_{x\to\infty}\frac{e^x-e^{-x}}{e^x+e^{-x}}$$

which is where we started! We can keep trying to apply L'Hospital's rule, but we'll simply alternate back and forth between these equally-difficult-to-compute limits.

VI.6 Order of growth

Definition VI.9: For two functions $f, g\colon \mathbb{R}\to\mathbb{R}$, we say that f is of *order of* g (or f is *big O* of g) if there exists $c, N\in\mathbb{R}$, such that $|f(x)|\le c\,|g(x)|$ for all $x\ge N$. We write[8] $f(x)=\mathcal{O}\left(g(x)\right)$.

If $f(x)=\mathcal{O}\left(g(x)\right)$, this means that behavior of f for large values of g is dominated by the behavior of g.

For example, we have that $x^2+5x=\mathcal{O}\left(x^2\right)$, since for $x\ge 5$ we have that $x^2\ge 5x$ and thus $\left|x^2+5x\right|\le x^2+x^2=2x^2$.

This notation is often used when comparing the performance of algorithms. Here $f(n)$ [9] is a function that, for a given algorithm indicates the cost (number or operations) required for an input that has size n. If there is a second algorithm, whose cost is given by a function g, we consider the first algorithm as not worse than the second, if $f(x)=\mathcal{O}\left(g(x)\right)$. If in addition $g(x)=\mathcal{O}\left(f(x)\right)$, the algorithms are considered equivalent.[10] We consider this more formally below.

The justification for such reasoning is that the constant c will make such a comparison independent of particular computers or programming languages used. The focus on large values of the input size allows us to ignore artifacts for small examples, or effects such as caching or results. The use of the $\mathcal{O}\left(\right)$ notation also allows us to focus an the main contribution to an algorithms runtime.

For example, when searching an object in a list of length l, a linear search has cost $\mathcal{O}\left(n\right)$ (go through the list until you find the object), while if the list is sorted a binary search has only cost $\mathcal{O}\left(\log(n)\right)$, which we shall see is much better.

The test for $f(x)=\mathcal{O}\left(g(x)\right)$ looks a little bit like the criterion for limit of a series, and proving it might be technical. The following theorem shows that this can be tested easier in many cases:

Theorem VI.10: Let $f, g\colon\mathbb{R}\to\mathbb{R}$ be two functions such that $L=\lim_{x\to\infty}\dfrac{f(x)}{g(x)}$ exists. Then

[8] The use of the equal sign is somewhat misleading and awkward, but that is the convention that is usually used, so we stick with it.

[9] It is common to give these classes as functions of a variable n, rather than x

[10] There is a plethora of variants of $\mathcal{O}\left(\right)$ that provide modified kinds of comparisons.

a) if $L = 0$ then $f(x) = \mathcal{O}(g(x))$

b) if $L = \infty$ then $g(x) = \mathcal{O}(f(x))$.

c) if $0 < L \neq \infty$ is a finite number then $f(x) = \mathcal{O}(g(x))$ **and** $g(x) = \mathcal{O}(f(x))$.

The proof of a) is basically to apply the definition of a limit, and to multiply by the denominator $g(x)$, this will yield the criterion for $\mathcal{O}()$. For b) and c) also consider the reciprocal fraction $g(x)/f(x)$.

NOTE VI.11: The criteria in this theorem are sufficient, but not necessary. For example consider $f(x) = \sin(x)$ and $g(x) = 1 \cdot x^0$. Then clearly $\sin(x) = \mathcal{O}(1)$ (choose constants $c = 1$ and $N = 1$), but the limit $\lim\limits_{x\to\infty} \dfrac{\sin(x)}{1}$ does not exist.

An important consequence of this theorem is the following observation, that allows to ignore "lower order" terms as "noise":

Suppose that $f(x) = \mathcal{O}(g(x))$ and $h(x) = \mathcal{O}(f(x))$. Then[11] $f(x) + h(x) = \mathcal{O}(g(x))$.

Thus, for example, setting $f(x) = g(x) = x^3$, $h(x) = 9x^2 + 17 + \log(x)$, we get that $x^3 + 9x^2 + 17 + \log(x) = \mathcal{O}\left(x^3\right)$.

Complexity classes

The criterion of Theorem VI.10 is often ready-made for using L'Hospital's theorem. For example let $f(x) = x + 1$ and $g(x) = x^2 - 3x + 2$. We have that

$$\lim_{x\to\infty} \frac{x+1}{x^2-3x+2} = \lim_{x\to\infty} \frac{1}{2x-3} = 0$$

Thus $x + 1 = \mathcal{O}\left(x^2 - 3x + 2\right)$.

The same argument will work for *any* pair of polynomials of degree 1 and 2.

Indeed, by applying L'Hospital's theorem several times, we find that if $f(x)$ is a polynomial of degree m, and $g(x)$ a polynomial of degree $n \geq m$ then $f(x) = \mathcal{O}(g(x))$. In particular (setting $g(x) = x^m$) we have that $f(x) = \mathcal{O}(x^m)$.

More generally, we get that for $a, b > 0$ with $a < b$ that $x^a = \mathcal{O}\left(x^b\right)$ but *not* vice versa.

We similarly compare a power x^a (for $a > 0$) to $\exp(x)$ and to $\log(x)$ (we give the argument for $a \in \mathbb{Z}$, the argument for non-integral a is very similar):

$$\begin{aligned}\lim_{x\to\infty} \frac{x^a}{\exp(x)} &= \lim_{x\to\infty} \frac{a\cdot x^{a-1}}{\exp(x)} = \cdots = \lim_{x\to\infty} \frac{a\cdot(a-1)\cdots 1\cdot x^{a-a}}{\exp(x)} \\ &= \lim_{x\to\infty} \frac{a!}{\exp(x)} = 0,\end{aligned}$$

[11] Since, by the limit rules, $\dfrac{f(x)+h(x)}{g(x)} = \dfrac{f(x)}{g(x)}\left(1 + \dfrac{h(x)}{f(x)}\right)$ and $\lim\limits_{x\to\infty}\left(1 + \dfrac{h(x)}{f(x)}\right) = 1$.

so $x^a = \mathcal{O}(\exp(x))$. Similarly

$$\lim_{x\to\infty} \frac{\log(x)}{x^a} = \lim_{x\to\infty} \frac{1/x}{a\cdot x^{a-1}} = \lim_{x\to\infty} \frac{7}{a\cdot x^a} = 0$$

and thus $\log(x) = \mathcal{O}(x^a)$.

We thus get a hierarchy of function classes that all have different growth (and thus would describe different algorithmic performance). We summarize this in Table VI.2 and Figure VI.8. The most prominent class distinctions are sub-linear,

	Name	Example algorithm in class
$\mathcal{O}(1)$	constant, or bounded	Append to a list
$\mathcal{O}(\log(n))$	logarithmic	Binary search in sorted list
$\mathcal{O}(n^c), 0 < c < 1$	sublinear, or fractional power	Testing n for primality by trial division
$\mathcal{O}(n)$	linear	Searching through an unsorted list
$\mathcal{O}(n\log(n))$	superlinear	Merge sort
$\mathcal{O}(n^c), 1 < c < 2$		
$\mathcal{O}(n^2)$	quadratic	Bubble sort
$\mathcal{O}(n^c), 2 < c < 3$		
$\mathcal{O}(n^3)$	cubic	Solving a system of n linear equations (standard method)
$\mathcal{O}(n^c), 3 < c$		
	All up to here are called "polynomial time"	
$\mathcal{O}(\exp(n^c)), 0 < c < 1$	subexponential	Best known factorization of n-digit number.
$\mathcal{O}(\exp(n))$	exponential	Brute-force breaking of a password of length n
$\mathcal{O}(n!)$	factorial	Traveling salesman by trying out all tours
$\mathcal{O}(\exp(\exp(n)))$	doubly exponential	Solving a system of n polynomial equations (using Gröbner bases).

Table VI.2: Common complexity classes and examples

linear, polynomial and exponential+beyond, since the composition of functions in these classes (which corresponds one algorithm calling another one) again lies in the same class.

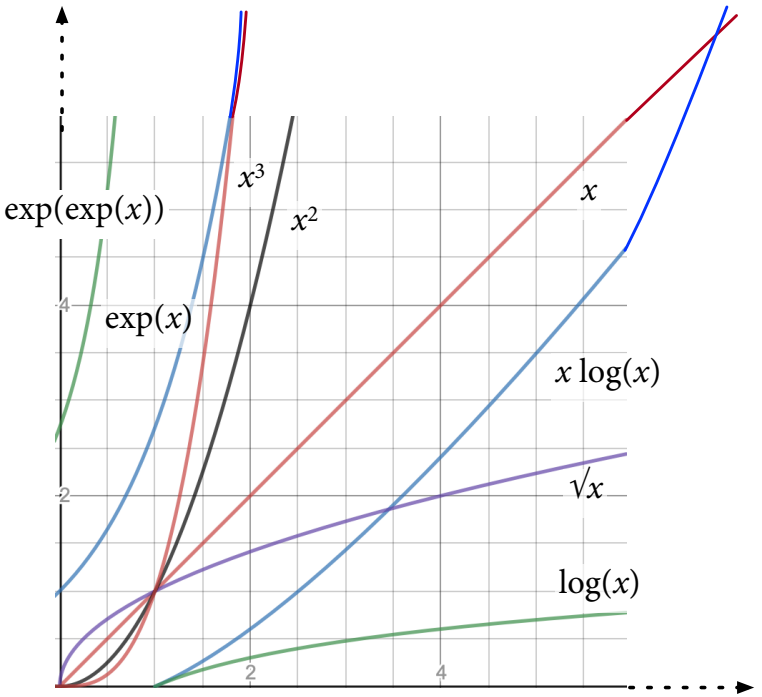


Figure VI.8: Some common growth functions

Complexity Equivalence

We claimed above that two functions could be considered equivalent if either is $\mathcal{O}$ of the other. What we shall show here is that this indeed defines an equivalence relation on functions, with this the objects $\mathcal{O}(f(x))$ can be simply considered as equivalence classes of functions.

As in Section II.4, we start with a relation. Define a relation $\sim$ on the set of functions $\mathbb{R} \to \mathbb{R}$ as $f(x) \sim g(x)$ if and only if $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$. Now we need to show that $\sim$ is reflexive, symmetric, and transitive, this will establish that $\sim$ is an equivalence relation,

Reflexivity: Let $f(x)$ be a function. We must show that $f(x) \sim f(x)$, that is, $f(x) = \mathcal{O}(f(x))$ (and the identical a second time, formally swapping the two $f(x)$). Take $c = 1$ and $N = 0$, then for $x > N$ it follows immediately that $|f(x)| \leq |f(x)|$.

Symmetry: Let $f(x)$ and $g(x)$ be functions. We must show that if $f(x) \sim g(x)$, then $g(x) \sim f(x)$. This follows immediately from the symmetry in the defi-

nition of the equivalence relation. If $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$ then we get the same statements if we swap the roles of $f(x)$ and $g(x)$.

Transitivity: Let $f(x)$, $g(x)$, and $h(x)$ be functions. We must show that if $f(x) \sim g(x)$ and $g(x) \sim h(x)$, then $f(x) \sim h(x)$. Suppose that $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(h(x))$, then we have some c_1, c_2, N_1, and N_2 such that

$$|f(x)| \leq c_1|g(x)| \text{ if } x \geq N_1 \text{ and } |g(x)| \leq c_2|h(x)| \text{ if } x \geq N_2.$$

Note that both statements hold simultaneously if we take x to be greater than the maximum of N_1 and N_2. Denote this maximum by N. Then we can combine the inequalities to obtain

$$|f(x)| \leq c_1c_2|h(x)| \text{ if } x \geq N.$$

Thus we have shown that $f(x) = \mathcal{O}(h(x))$. Proceeding in the same way, we can find that $h(x) = \mathcal{O}(f(x))$ as well.

We have thus shown that the relation $\sim$ satisfies all the requirements to be an equivalence relation.

The relation $\sim$ thus creates a partition of the set of function onto equivalence classes (which we shall call *complexity classes*). The notation $\mathcal{O}(f(x))$ then can be interpreted as the complexity class containing $f(x)$.

The reader might be wondering about why we defined the relation in such a complicated way. What if we defined a similar relation $\vdash$ by having $f(x) \vdash g(x)$ if $f(x) = \mathcal{O}(()\, g(x))$. Is this still an equivalence relation? It turns out that it is not. Notice that the proof of reflexivity and transitivity still hold just as well in the case of $\vdash$, but in contrast to $\sim$, symmetry is not built into the definition of our new relation. Consider the functions $f(x) = x$ and $g(x) = x^2$. For $c = 1$ and $N = 0$ it follows that $|x| \leq 1 \cdot |x^2|$ for $x \geq 0$. On the other hand, it is impossible to find a c and N so that $|x^2 \leq c|x|$ for $x \geq N$.

Chapter

VII

Taylor Series

Polynomials are in many ways the easiest, most convenient, functions to work with, and it thus would be nice if every function was a polynomial. The formula for the geometric series shows that this might be possible, if we allow for something like a polynomial, but of infinite degree: Consider the function $f(x) = 1/(1-x)$, that is not a polynomial. Then, as long as $|x| < 1$ we have that

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{i \geq 0} x^i$$

(by the summation formula for the geometric series).

In this chapter we will work with three ideas: The first is how we can approximate any function by a polynomial (called a Taylor polynomial), and to see that such approximations become better as the degree increases. The second topic is what a polynomial of infinite degree should be (it will be called a power series). Finally we will extend Taylor polynomials to infinite degree to see how we can represent most of the functions that we encounter as power series.

VII.1 Taylor Polynomials

The basic idea for finding a polynomial approximation of a function is that we want the polynomial and the function to have the same values at a point, as well as the same values of the derivatives.[1]

We start by considering values at 0 and consider a polynomial

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = \sum_{i=0}^{n} a_i x^i$$

[1] We shall see that this turns out to be a useful approach. In fact it is better than the alternative approach of having the same values at a number of different points.

Its derivatives are

$$\begin{aligned} p'(x) &= a_1 + 2a_2x + 3a_3x^2 + \cdots + n \cdot a_n x^{n-1} = \sum_{i=1}^{n} i \cdot a_i x^{i-1}, \\ p''(x) &= 1 \cdot 2a_2 + 2 \cdot 3a_3x + \cdots + (n-1) \cdot n \cdot a_n x^{n-2} = \sum_{i=2}^{n} (i-1) i \cdot a_i x^{i-2}, \end{aligned}$$

and generally, for the k-th derivative

$$p^{(k)}(x) = \sum_{i=k}^{n} (i-k+1)\cdots(i-1) i \cdot a_i x^{i-k}.$$

This means that the value of this derivative at 0 is the constant term (for $i = k$):

$$p^{(k)}(0) = 1 \cdot 2 \cdots (k-1) \cdot k a_k = k! a_k,$$

where $k!$ (called *k-factorial*) is the product over all numbers from 1 to k.

To ensure that the derivatives of $p(x)$ are equal to those of $f(x)$, we thus get the conditions $f^{(k)}(0) = k! a_k$ which we can solve for

$$a_k = \frac{f^{(k)}(0)}{k!}.$$

Definition VII.1: Let $f: \mathbb{R} \to \mathbb{R}$ be a function that is repeatedly differentiable. The *Taylor polynomial* for f around $a = 0$ of degree n is the polynomial

$$\sum_{i=0}^{n} \frac{f^{(i)}(0)}{i!} x^i$$

(where $f^{(k)}(0)$ is the value of the k-th derivative at 0).

For example, if $f(x) = \exp(x) = e^x$ (we pick this example for the easy pattern of derivatives), we have that

$$f(0) = e^{(}0) = 1, \quad f'(0) = e^{(}0) = 1, \quad f''(0) = e^{(}0) = 1, \quad \ldots, f^{(k)}(0) = e^{(}0) = 1$$

and we thus get the Taylor polynomials of degree $0, 1, 2, 3, \ldots$ as

$$\begin{aligned}
p_0(x) &= \frac{1}{0!}x^0 = 1 \\
p_1(x) &= \frac{1}{0!}x^0 + \frac{1}{1!}x = 1 + x \\
p_2(x) &= 1 + x + \frac{1}{2!}x^2 = 1 + x + \frac{x^2}{2} \\
p_3(x) &= 1 + x + \frac{x^2}{2} + \frac{1}{3!}x^3 = 1 + x + x^2 + \frac{x^3}{6} \\
p_4(x) &= 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} \\
&\vdots \\
p_n(x) &= \sum_{i=0}^{n} \frac{x^i}{i!}
\end{aligned}$$

The first few polynomials, together with the function $f(x) = \exp(x)$ are shown in Figure VII.1.

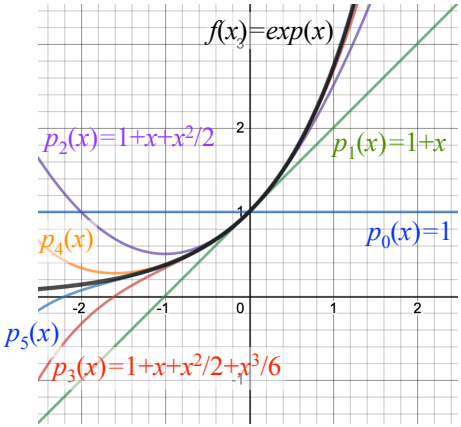


Figure VII.1: Taylor approximations to $\exp(x)$

If we choose instead $f(x) = \frac{1}{x+2}$, we get

$$\begin{aligned} f'(x) &= -\frac{1}{(x+2)^2} \\ f''(x) &= \frac{2}{(x+2)^3} \\ f'''(x) &= -\frac{6}{(x+2)^4} \\ f^{(k)}(x) &= (-1)^k \frac{k!}{(x+2)^{k+1}} \end{aligned}$$

and thus Taylor polynomials

$$\begin{aligned} p_0(x) &= \frac{1}{2} \\ p_1(x) &= \frac{1}{2} - \frac{x}{4} \\ p_2(x) &= \frac{1}{2} - \frac{x}{4} + \frac{x^2}{8} \\ &\vdots \\ p_5(x) &= \frac{1}{2} - \frac{x}{4} + \frac{x^2}{8} - \frac{x^3}{16} + \frac{x^4}{32} - \frac{x^5}{64} \end{aligned}$$

These polynomials are shown in Figure VII.2.

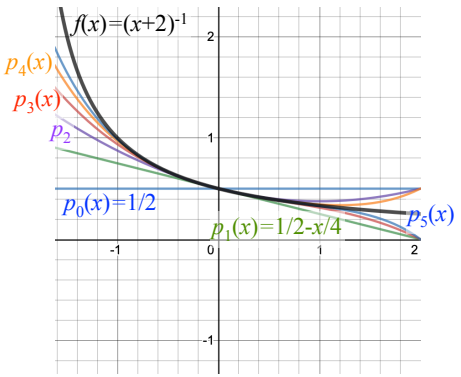


Figure VII.2: Taylor approximations to $(x+2)^{-1}$

Finally, take $f(x) = \sin(x)$, and we get

$$f(0) = \sin(0) = 0, f'(0) = \cos(0) = 1, f''(0) = -\sin(0) = 0, f'''(0) = -\cos(0) = -1, \ldots$$

Thus every second coefficient is zero and we get new Taylor polynomials only for odd indices:

$$\begin{aligned} p_1(x) = p_2(x) &= x \\ p_3(x) = p_4(x) &= x - x^3/6 \\ p_5(x) = p_6(x) &= x - x^3/6 + x^5/120 \end{aligned}$$

as shown in Figure VII.3. These calculations allow us to make the following obser-

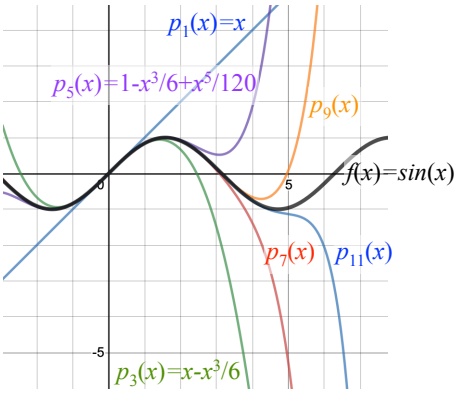


Figure VII.3: Taylor approximations to $\sin(x)$

vations:

1. As the degrees increase, Taylor polynomials just accumulate further terms.

2. The Taylor polynomials approximate well around $a = 0$, but the approximation becomes worse if we go away from $a = 0$.

3. The approximation gets better, the higher the degree of the Taylor polynomial is.

We shall give a justification for this (and show that this true in general) below.

So far we have formed Taylor polynomials around $a = 0$. The rules we know about shifting functions horizontally allow us to define a polynomial around arbitrary real numbers a by shifting accordingly. The corresponding (more general) definition is unsurprising.

DEFINITION VII.2: Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function that is repeatedly differentiable and $a \in \mathbb{R}$. The Taylor polynomial for f around a of degree n is the polynomial

$$\sum_{i=0}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i$$

(where $f^{(k)}(a)$ is the value of the k-th derivative at a). We call a the *center* of the Taylor polynomial.

Note that for $a = 0$ this just gives the prior definition.

For example, using the calculations of the derivatives we already did, we find the Taylor polynomial of degree 5 for $f(x) = \frac{1}{x+2}$ around $x = -1$ as

$$1 - (x+1) + (x+1)^2 - (x+1)^3 + (x+1)^4 - (x+1)^5.$$

Figure VII.4 compares this Taylor polynomial with the one around $a = 0$ we had computed above. Unsurprisingly the polynomial around $a = -1$ approximates better for x-values closer to -1, while the one around $a = 0$ approximates better for x-values closer to 0.

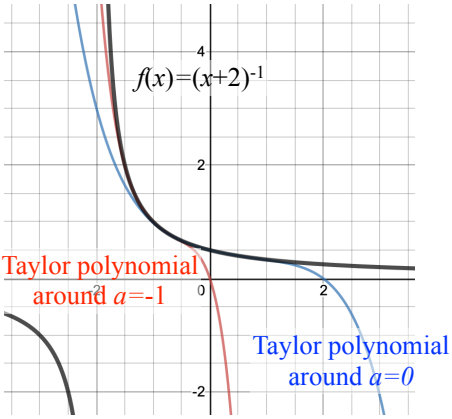


Figure VII.4: Taylor approximations to $(x+2)^{-1}$ around $a = -1$ and $a = 0$

Approximation Error

The core to understanding the way a Taylor polynomial approximates a function is the *error term*. One can show:

THEOREM VII.3: Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function that can be differentiated n times, $a \in \mathbb{R}$, and $l > 0$. Then the approximation error of the Taylor polynomial of degree

n for x in in an interval of length $2l$, centered at a (that is, $|x-a| \leq l$, respectively $a-l \leq x \leq a+l$) is:

$$|f(x) - p_n(x)| \leq \frac{M}{(n+1)!}|x-a|^{n+1} \leq \frac{M}{(n+1)!}l^{n+1}$$

where M is an upper bound for the value of the $n+1$-st derivative of f on the interval $a-l \leq x \leq a+l$.

We shall not attempt to prove this theorem, but let's explain what is says:

1. Taylor polynomials can be used to approximate a function, and the quality of the approximation can be quantified.

   In sciences, it is common to replace a complicated function in approximation by a Taylor polynomial of small degree (often degree 1, also called *first order approximation*). A *nonlinear* phenomenon is one that cannot be explained by using a Taylor polynomial approximation of degree 1, but which requires a higher degree.

2. The factor $|x-a|^{n+1}$ in the error term tells us that the approximation is the better the closer x is to the center a.

3. With $(n+1)!$ in the denominator and a numerator l^{n+1}, the approximation usually gets better, if the degree of the polynomial gets larger. (This is basically the fact that $n!$ is in a higher complexity class than 2^n.)

4. Finally for the somewhat mysterious parameter M: It is a number, namely a bound for the values of the $n+1$-st derivative of f on the interval. For most functions (basically all functions that we shall encounter in this course, maybe all you will ever encounter in your professional life) these derivatives are bounded, independent from n. For if this was not the case, higher and higher derivatives would need to be larger and larger, which means that the function is in some way "strange"[2] You are unlikely to encounter them in this course.

   As long as this number M is bounded (and again, this is something we shall assume), the Taylor polynomials provide increasingly good approximations.

In other words, this theorem is a justification of the approximation properties we observed in the examples above.

Indeed, if you press a key for "sine" or "exp" on your calculator, or if you call the built-in functions `sin` or `exp` in your favorite programming language, what is happening internally, that the resulting value is obtained fundamentally (there are many other practical tricks involved) through a Taylor polynomial of suitably high degree.

[2] The standard example is the function $\exp(-x^{-2})$ around $a = 0$, which is, for small values of x, practically indistinguishable from the constant zero function.

Using approximations

To further illustrate how Taylor polynomial approximations work, consider the following four functions:

$$1+\sin(x), \quad \exp(x), \quad \frac{1}{\sqrt{1-2x}}, \quad \frac{1}{1-x}.$$

All four have value 1 at 0 and increase, in a plot (Figure VII.5, deliberately not

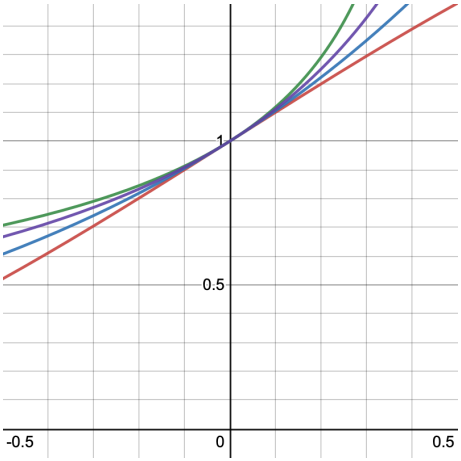


Figure VII.5: Four close functions

labeled) they are close for small values of $|x|$. It seems hard to determine by hand which graph belongs to which function.

To help with this decision, consider Taylor approximations of the four functions:

$$\text{a)} \quad 1+\sin(x) \sim 1+x-\frac{x^3}{6}+\frac{x^5}{120}-\frac{x^7}{5040}+\cdots$$

$$\text{b)} \quad \exp(x) \sim 1+x+\frac{x^2}{2}+\frac{x^3}{6}+\frac{x^4}{24}+\frac{x^5}{120}+\cdots$$

$$\text{c)} \quad \frac{1}{\sqrt{1-2x}} \sim 1+x+\frac{3x^2}{2}+\frac{5x^3}{2}+\frac{35x^4}{8}+\frac{63x^5}{8}+\cdots$$

$$\text{d)} \quad \frac{1}{1-x} \sim 1+x+x^2+x^3+x^4+x^5+\cdots$$

All start with $1+x$, which explains why the functions are so close for small $|x|$. But the coefficients of the x^2 terms differ, they are (in increasing order) 0 for a), $\frac{1}{2}$ for

b), 1 for d) and $\frac{3}{2}$ for c). This means that for small values of x, we expect $1+\sin(x)$ to be the smallest and $\frac{1}{\sqrt{1-2x}}$ to be the largest. Since $x^2 = (-x)^2$ this will hold for both positive and negative x, as the plot shows. And for $|x|$ sufficiently small the graph of $1+x+\frac{3}{4}x^2$ would[3] lie in the middle between the top and bottom two lines.

Using the error estimate

We briefly illustrate how the error term formula can be used to get a concrete estimate of the approximation error. Suppose we want to approximate $\sin(x)$ on the interval from -2 to 2 by a Taylor polynomial, centered at $a = 0$. Then (because of the interval) $|x-a| \le 2$. Next we want to estimate the derivatives. Since $\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) = \cos(x)$ and $\frac{\mathrm{d}}{\mathrm{d}x}\cos x = -\sin(x)$ we choose $M = 1$ as the largest value of these functions. (We could have chosen also $M = 5$ as a (not that tight) bound. In other cases it can be harder to give very tight estimates, but usually even a rough estimate will usually do well. The error estimate then gives us an error

$$|f(x) - p_3(x)| \le \frac{1}{4!}2^4 \sim 0.66$$

for a polynomial of degree 3. [4] For a degree 8 approximation [5]

$$|f(x) - p_8(x)| \le \frac{1}{9!}2^9 = \frac{512}{362880} \sim 0.0014.$$

If we wanted to guarantee an error of less than 10^{-6} we can try out increasing values for n [6] and find that this is the case for $n \ge 14$.

An somewhat obvious improvement of the approximation is by cranking up the degree. Basically, we shall show in Section VII.2 that a Taylor polynomial "of degree infinity" can be used in place of the function (with exact values, no approximation).

Fast inverse square root

An application of Taylor approximations together with Newton's Method is found in a famous example of an optimized routine, namely the evaluation of the function $a \mapsto \frac{1}{\sqrt{(a)}}$.

This function is used in digital signal processing to re-scale a vector to length 1, which is required for example when calculating the illumination level and shading of a surface in 3D graphics.

[3] not depicted in the plot as this requires serious zooming in for negative x.

[4] Note that this is an estimate and a promise that the error is not worse. Actually the error typically will be much smaller than what the estimate.

[5] For $f(x) = \sin(x)$ the degree 8 Taylor polynomial is actually the same as the degree 7 Taylor polynomial.

[6] since solving the equation for n is not possible

When rendering a scene, this value has to be calculated separately for each triangle, its execution therefore is highly time critical.

While the newer x86 SSE instruction set includes a dedicated operation `rsqrtss` for it, historically this functionality had to be coded from more basic operations. However both square root and division are expensive, which explains why other approaches have been considered.

The following method became prominent in the early 2000s as being used in the video game `Quake III`, though it dates back to uses in computer graphics in the 1980s, for example on `SGI Indigo` graphics workstations. What we shall describe is a somewhat simplified version, that leaves out subtleties in how numbers are coded on the computer[7], and how error is minimized.

The method works by first computing one rough approximation of the value, and then using one step of the Newton method to refine the result. (It turns out that the initial approximation is good enough that one step of the Newton method suffices, but we shall not show this.)

EXAMPLE VII.4: The input number a (when computing $1/\sqrt{a}$) is given as a floating point number in the form[8]

$$a = 2^e \cdot (1 + m)$$

with $0 \le m < 1$ (if we had $m > 1$ one could instead increase e). That means that

$$\frac{1}{\sqrt{a}} = 2^{-\frac{1}{2}e} \cdot \frac{1}{\sqrt{1+m}}.$$

The term $2^{-\frac{1}{2}e}$ can immediately be taken as a factor of a power of 2. For the second factor we look at the Taylor polynomial (with respect to the variable m) for $\frac{1}{\sqrt{1+m}}$, which is

$$1 - \frac{m}{2} + \frac{3m^2}{8} + \cdots$$

and use the degree 1 approximation $1 - \frac{m}{2}$. (Here in fact, the code uses a number different from 1 to minimize the error, but how this is done is beyond the scope of this text.) Note that $m/2$ can be computed cheaply as a shift of a binary number.

We thus have the initial approximation

$$a_0 = 2^{-\frac{1}{2}e} \cdot \left(1 - \frac{m}{2}\right).$$

Now for the Newton method. The function we want to find a zero of is

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

[7] this incidentally is also the reason for the name "`0x5F3759DF` method" that is sometimes used

[8] This is dictated by the CPU of the computer

whose value will be zero exactly when $x = \frac{1}{\sqrt{x}}$. Newton's method (Section VI.4) states that the next iteration must be

$$a_{n+1} = a_n - \frac{f(a_n)}{f'(a_n)}).$$

We calculate $f'(x) = -\frac{2}{x^3}$ and thus

$$\frac{f(x)}{f'(x)} = \frac{\frac{1-x^2a}{x^2}}{-\frac{2}{x^3}} = \frac{x(1-ax^2)}{2}.$$

The first Newton iterate is thus

$$a_1 = a_0 + \frac{a_0(1-a_0x^2)}{2} = \frac{a_0(3-a_0x^2)}{2} = a_0\left(\frac{3}{2} - \frac{a}{2}a_0^2\right).$$

This gives us the algorithm (as published on the web, e.g. https://en.wikipedia.org/wiki/Fast_inverse_square_root):

```
1   const float threehalfs = 1.5F;

2   x2= number * 0.5F;
3   y = number;
4   i  = *(long*) &y;       // floating point hacking
5   i = 0x5f3759df-(i>>1);
6   y = * ( float * ) &i;
7   y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
```

Lines 4-6 calculate $2^{-\frac{1}{2}e} \cdot \left(c - \frac{m}{2}\right)$ with some trickery that uses bit operations on floating point numbers by reinterpreting their bit pattern as an integer – the `(i>>1)` is the $\frac{m}{2}$ and the hexadecimal constant is the optimized value for c. Line 7 then is the Newton iteration. (The code in fact includes a second iteration that has been commented out, as the first one turned out to be good enough in practice.)

VII.2 Taylor Series

If we form Taylor polynomials of increasing degree for a function $f(x)$ (that is infinitely often differentiable), we get a sequence of partial sums of the infinite *Taylor Series*[9]

$$t(x) = \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} x^i.$$

While it might look strange to have the variable x involved, one could imagine setting x to some number, then this becomes just an ordinary series. We call such series, that involve powers of a variable x, *power series*.

[9] For reasons of time we now focus only on the case of series centered around $a = 0$. The same theory would hold if we center around another a and replace x by $x - a$.

For example, we get for $f(x) = \exp(x)$ the Taylor series

$$t(x) = \sum_{i=0}^{\infty} \frac{1}{i!} x^i.$$

One can show that for such a series there exists $b \geq 0$ such that:

a) The series converges, if $-b < x < b$.

b) If $-b < x < b$ one has that $f(x) = t(x)$ (where $t(x)$ is the limit of the infinite series for the chosen value x.

c) If there is *any* power series with properties a) and b) that produces the values of f, it must be identical to $t(x)$.

The actual value of b will depend on the coefficients (and thus on the function f), or one can describe it using the error term from the previous section. In bad cases it can be just 0, but there are many important functions for which b can be chosen arbitrary large. (Such functions are called *analytic*.)

These facts have a number of important consequences, which we shall briefly explore:

Treating Functions as if they are Polynomials: We shall see below that such power series can be treated, for arithmetic, as well as for calculating derivatives, as if they were polynomials (albeit of infinite degree). This can make some arguments about such functions easier.

Finding Taylor polynomials and Taylor series: Taylor polynomials are just partial sums of Taylor series. We shall see that we can manipulate Taylor series to obtain series for functions for which it would be harder to find Taylor polynomials by direct means (that is calculating derivatives).

Defining new functions: A very convenient way of defining functions with particular properties is as Taylor series. Indeed, this is how functions such as $\exp(x)$ and $\sin(x)$ get properly defined: The definition in school, as powers e^x or ratios of sides of triangles, cause difficulties[10] in how one would evaluate these for irrational values of x.

Instead, mathematicians *define* these functions through power series

$$\begin{aligned} \exp(x) &= \sum_{i=0}^{\infty} \frac{1}{i!} x^i \\ \sin(x) &= \sum_{i=0}^{\infty} \frac{(-1)^i}{(2i+1)!} x^{2i+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \end{aligned}$$

[10] Not to say they actually never address the question!

and then show that these functions agree with (and have the same properties) as the functions with the same names known from school.

There are many other functions (for example so-called "Bessel functions") that arise in applications of mathematics, and that are only defined as such a Taylor series.

Complex Numbers

Taylor series thus also are the key of extending the definition of functions to the complex numbers. Being based on polynomials, one can evaluate a Taylor series at a complex number as well as one could do at a real number. This makes it possible to define (say) $\sin(x)$ for a complex value of x, even though the geometric interpretation then does not make sense.

We illustrate this in a classical example, that is called *Euler's formula*, which connects the exponential function with sine and cosine: Substitute ix (where $i^2 = -1$) for x in the exponential function, and we get

$$\begin{aligned} e^{ix} &= 1 + ix + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \frac{(ix)^5}{5!} + \frac{(ix)^6}{6!} + \frac{(ix)^7}{7!} + \frac{(ix)^8}{8!} + \cdots \\ &= 1 + ix - \frac{x^2}{2!} - \frac{ix^3}{3!} + \frac{x^4}{4!} + \frac{ix^5}{5!} - \frac{x^6}{6!} - \frac{ix^7}{7!} + \frac{x^8}{8!} + \cdots \\ &= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right) \\ &= \cos(x) + i\sin(x), \end{aligned}$$

and thus in particular that

$$1 + e^{i\cdot\pi} = 0.$$

Evaluating also e^{-ix} allows us to solve for $\sin(x)$ and $\cos(x)$ and express both functions in terms of the exponential function as

$$\begin{aligned} \sin x &= \operatorname{Im}\left(e^{ix}\right) = \tfrac{e^{ix} - e^{-ix}}{2i}, \\ \cos x &= \operatorname{Re}\left(e^{ix}\right) = \tfrac{e^{ix} + e^{-ix}}{2}. \end{aligned}$$

Taylor Series Operations

Before looking at (and constructing) more examples, we briefly state some of the properties for arithmetic with Taylor series. The basic rule is that in the interval of convergence, Taylor series behave like polynomials:

THEOREM VII.5: Let $f, g: \mathbb{R} to \mathbb{R}$ be functions that are infinitely often differentiable and that are (for values of x in a given interval) equal to their respective Taylor series

$$f(x) = \sum_{i=0}^{\infty} a_i x^i, \quad g(x) = \sum_{i=0}^{\infty} b_i x^i.$$

Then (for values of x in the interval) we have that

1. For $c \in R$, $(cf)(x) \sum_{i=0}^{\infty}(c \cdot a_i)x^i$.
2. $(f+g)(x) = f(x) + g(x) = \sum_{i=0}^{\infty}(a_i + b_i)x^i$.
3. $(f \cdot g)(x) = f(x) \cdot g(x) = \sum_{i=0}^{\infty}\left(\sum_{j=0}^{i} a_j \cdot b_{i-j}\right)x^i$.
4. $f'(x) = \sum_{i=0}^{\infty}(i+1)a_{i+1}x^i = \sum_{i=1}^{\infty} i \cdot a_i x^{i-1}$.
5. The power series $F(x) = \sum_{i=0}^{\infty} \frac{a_i}{i+1}x^{i+1}$ has the derivative $F'(x) = f(x)$. (I.e. F is an *antiderivative* of f.)

In summary, Taylor series can be treated like polynomials.

Note that – since Taylor series are unique – these equations give *the* Taylor series for the respective functions, and any operation (such as substituting x^k for x) on a Taylor series must be the Taylor series for the same operation of a function.

As with polynomials, if a function is even (that is $f(x) = f(-x)$) then all powers of x in the Taylor series have even exponent. Similarly all powers have odd exponent, if the function is odd ($f(-x) = -f(x)$).

Examples and Applications

This theorem has interesting consequences for finding derivatives and Taylor series.

Let us start with the Taylor series for $\exp(x)$ which is (for any $x \in \mathbb{R}$)

$$\exp(x) = \sum_{i=0}^{\infty} \frac{1}{i!}x^i.$$

Its derivative thus is

$$\frac{\mathrm{d}}{\mathrm{d}x}\exp(x) = \sum_{i=0}^{\infty} \frac{i+1}{(i+1)!}x^i = \sum_{i=0}^{\infty} \frac{i}{i!}x^i = \exp(x).$$

This proves that $\frac{\mathrm{d}}{\mathrm{d}x}\exp(x) = \exp(x)$.

Similarly, consider the Taylor series for $\sin(x)$ which is

$$\sin(x) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(2i+1)!}x^{2i+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

We thus have the derivative

$$\frac{\mathrm{d}}{\mathrm{d}x}\sin(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{i=0}^{\infty} \frac{(-1)^{i+1}}{(2i)!}x^{2i} = \cos(x)$$

(and similarly for the derivative of $\cos(x)$.)

These observations are ultimately the justifications for the derivatives of elementary functions we stated in Section V.8.

Next, if $|x| < 1$ the formula for the geometric series gives us that

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{i=0}^{\infty} x^i$$

Substituting $-x$ for x gives us

$$\frac{1}{1+x} = \sum_{i=0}^{\infty} (-1)^i x^i$$

We know that $F(x) = \log(1+x)$ is a function with $F'(x) = 1/(1+x)$ and $F(0) = 0$. We thus get the new Taylor series

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} x^i.$$

Returning to the use of the geometric series, we can write

$$\frac{x}{x-3} = -\frac{x}{3}\frac{1}{1-x/3}.$$

Substituting $x/3$ into the series for $\frac{1}{1-x}$ gives us

$$\frac{1}{1-x/3} = \sum_{i=0}^{\infty} \left(\frac{x}{3}\right)^i,$$

and thus

$$\frac{x}{x-3} = -\frac{x}{3}\sum_{i=0}^{\infty} \left(\frac{x}{3}\right)^i = \sum_{i=1}^{\infty} \left(\frac{x}{3}\right)^i.$$

One can use this idea more generally (this goes under the name of *partial fractions*, but we won't go into details here) for rational functions, that is fractions of polynomials. If we can write these as sums of fractions with easier denominators (that is, factors of the original denominator), it is possible to combine Taylor series of the summands to a series for the original function.

For example (You should be able to follow this calculation, but it is not expected that you would be able to come up with it on your own), consider

$$f(x) = \frac{21x - 15}{x^3 - 7x^2 + 5x - 35} = \frac{-3x}{x^2+5} + \frac{3}{x-7}.$$

As above, we calculate

$$\frac{3}{x-7} = -\frac{3}{7}\frac{1}{1-x/7} = -\frac{3}{7}\sum_{i=0}^{\infty} \left(\frac{x}{7}\right)^i = \sum_{i=0}^{\infty} \frac{-3x^i}{7^{i+1}}.$$

and

$$\frac{-3x}{x^2+5} = \frac{3x}{5}\frac{1}{1-(-x^2/5)} = \frac{3x}{5}\sum_{i=0}^{\infty}\left(\frac{-x^2}{5}\right)^i = -\frac{3x}{5}+\frac{3x^3}{25}-\frac{3x^5}{125}+\cdots,$$

combining to a Taylor series for $f(x)$:

$$\begin{aligned}\frac{21x-15}{x^3-7x^2+5x-35} &= \frac{3}{7}-\frac{132x}{245}-\frac{279x^2}{1715}+\frac{5808x^3}{60025}+\frac{13011x^4}{420175}-\frac{287892x^5}{14706125}\\ &\quad -\frac{640839x^6}{102942875}+\frac{14090208x^7}{3603000625}+\frac{31384611x^8}{25221004375}\\ &\quad -\frac{690502692x^9}{882735153125}-x\frac{1537928439x^{10}}{6179146071875}+\cdots\end{aligned}$$

VII.3 Outlook: Solving Recursions

A powerful mathematical tool, called *generating functions* uses Taylor series manipulations to find explicit expressions for recursively defined series.

We illustrate this with the example of the Fibonacci numbers, defined as:

$$f_0 = 0, \qquad f_1 = 1, \qquad f_{n+2} = f_{n+1} + f_n.$$

We now define a function from a Taylor series, whose coefficients are the Fibonacci numbers:

$$F(x) = \sum_{i=0}^{\infty} f_i x^i = x + x^2 + 2x^3 + 3x^4 + 5x^5 + 8x^6 + 13x^7 + \cdots$$

Note that

$$xF(x) = \sum_{i=0}^{\infty} f_i x^{i+1} = \sum_{i=1}^{\infty} f_{i-1} x^i$$

and

$$x^2F(x) = \sum_{i=0}^{\infty} f_i x^{i+2} = \sum_{i=2}^{\infty} f_{i-2} x^i$$

and thus

$$\begin{aligned}xF(x)+x^2F(x) &= f_0x + \sum_{i=2}^{\infty}\underbrace{(f_{i-1}+f_{i-2})}_{=f_i}x^i = \sum_{i=2}^{\infty} f_i x^i\\ &= \sum_{i=0}^{\infty} f_i x^i - f_0 - f_1 x = F(x) - x.\end{aligned}$$

We can solve this equation for $F(x)$, and get

$$F(x) = \frac{x}{1-x-x^2} = x\cdot\left(\frac{a}{x-\alpha}+\frac{b}{x-\beta}\right)$$

for

$$\alpha = \frac{-1+\sqrt{5}}{2}, \quad \beta = \frac{-1-\sqrt{5}}{2}, \quad a = \frac{-1}{\sqrt{5}}, \quad b = \frac{1}{\sqrt{5}}.$$

The geometric series gives us (as above) that

$$\frac{a}{x-\alpha} = \sum_{i=0}^{\infty} \frac{-a}{\alpha^{i+1}} x^i, \quad \frac{b}{x-\beta} = \sum_{i=0}^{\infty} \frac{-b}{\beta^{i+1}} x^i,$$

and thus the Taylor series

$$\begin{aligned} F(x) &= x \cdot \sum_{i=0}^{\infty} \left(\frac{-a}{\alpha^{i+1}} + \frac{-b}{\beta^{i+1}} \right) x^i \\ &= \sum_{i=0}^{\infty} \left(\frac{-a}{\alpha^{i+1}} + \frac{-b}{\beta^{i+1}} \right) x^{i+1} = \sum_{i=1}^{\infty} \left(\frac{-a}{\alpha^i} + \frac{-b}{\beta^i} \right) x^i. \end{aligned}$$

Comparing the coefficients of x^i (which we can do because Taylor series are unique) gives us that for $i > 0$ we have

$$\begin{aligned} f_i &= \frac{-a}{\alpha^i} + \frac{-b}{\beta^i} = \frac{1}{\sqrt{5} \cdot \alpha^i} + \frac{-1}{\sqrt{5} \cdot \beta^i} \\ &= \frac{1}{\sqrt{5} \cdot \left(\frac{-1+\sqrt{5}}{2} \right)^i} + \frac{-1}{\sqrt{5} \cdot \left(\frac{-1-\sqrt{5}}{2} \right)^i} \\ &= \frac{1}{\sqrt{5}} \left(\left(\frac{2}{\sqrt{5}-1} \right)^i + \left(\frac{2}{\sqrt{5}+1} \right)^i \right), \end{aligned}$$

which is an explicit formula for the Fibonacci numbers.

Chapter

VIII

Antiderivatives

VIII.1 Reverting Integration

So far we have looked at derivatives, that is takes a function and considered how it changes.

Clearly one can also look at the reverse process, that is take a function $f(x)$ and find a new function $g(x)$ such that $g'(x) = f(x)$, i.e. whose derivative is $f(x)$. We then call $g(x)$ an *antiderivative* of $f(x)$.

Note VIII.1: We are careful in talking about **an** antiderivative and not **the** antiderivative. This is that if we define $h(x) = g(x) + c$ for some $c \in \mathbb{R}$, then $g(x)$ and $h(x)$ are **both** antiderivatives of $f(x)$.

Definition VIII.2: The set of all antiderivatives of a function $f(x)$ is called the *indefinite integral* of $f(x)$ and written as

$$\int f(x)\mathrm{d}x$$

In this integral, we call $f(x)$ the *integrand*[1].

If we have one antiderivative, all other antiderivatives will differ by an additive constant.[2] We therefore often will express indefinite integrals by giving one representative function, and add a $+C$ to indicate that an arbitrary constant could be added. For example:

$$\int \cos(x)\mathrm{d}x = \sin(x) + C,$$

since $\sin'(x) = \cos(x)$.

[1] That is the function that is to be integrated

[2] Formally, we can define an equivalence relation on functions as having the same derivative. The equivalence classes then are the indefinite integrals.

We can verify such statements about antiderivatives easily by calculating a derivative. For example, the claim

$$\int exp(x) \cdot (x+1)\mathrm{d}x = x \cdot \exp(x) + C$$

is easily verified by computing the derivative:

$$\frac{\mathrm{d}}{\mathrm{d}x} x \cdot \exp(x) = \exp(x)(x+1).$$

Antiderivatives typically are not used to investigate functions (as we have done with derivatives), but to express a summation operation on function values and to calculate areas ⇝ VIII.5.

VIII.2 Basic Antiderivative Rules

While calculating derivatives is a straightforward process that can easily be automated, calculating antiderivatives is much harder. In particular, there are many situations whether it is not possible to give an easy formula for an antiderivative of a function given by a formula.

Engineering Calculus courses spend a large amount of time to teach strategies that can be used to find antiderivatives in many practical cases. However much of this work can be done nowadays by computer algebra systems that implement such strategies as well as a more generic algorithm (called the Risch-algorithm, and being far more complicated than the derivative algorithm given in an earlier chapter, thus we will not study it here).

Our goal thus will be just to describe some of the basic methods for finding antiderivatives, without aiming to make the reader an expert in these techniques. Rather the goal is to describe the basic techniques that are used so that it would be possible to follow a calculation of an antiderivative.

Sums and multiples The two most basic derivative rules are

$$\frac{\mathrm{d}}{\mathrm{d}x}(f(x)+g(x)) = f'(x)+g'(x), \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}x}(c \cdot f(x)) = c \cdot f'(x)$$

(for c being a constant, independent of x). This immediately gives the following antiderivative rules:

$$\int f(x)+g(x)\mathrm{d}x = \int f(x)\mathrm{d}x + \int g(x)\mathrm{d}x, \qquad \text{and} \qquad \int c \cdot f(x)\mathrm{d}x = c \cdot \int f(x)\mathrm{d}x.$$

Polynomials and Power series The derivative rule for powers of x:

$$\frac{\mathrm{d}}{\mathrm{d}x}x^a = a \cdot x^{a-1}$$

reverts to the antiderivative rule for $a \neq -1$:

$$\int x^a = \frac{1}{a+1}x^{a+1} + C.$$

that is verified by taking the derivative on the right. (If $a = -1$, the integral will be $\log(x) + C$.)

Together with the rules of the previous paragraph this gives us a general rule for the antiderivative of a polynomial:

$$\int \left(\sum_{i=0}^{n} a_i x^i\right) \mathrm{d}x = \sum_{i=0}^{n} \frac{a_i}{i+1}x^{i+1} + C.$$

This rule also extends to power series:

$$\int \left(\sum_{i\geq 0} a_i x^i\right) \mathrm{d}x = \sum_{i\geq 0} \frac{a_i}{i+1}x^{i+1} + C.$$

Looking up the Result We can take a table for known derivatives and swap its columns to get a table of antiderivatives.[3] For us, the table in Section V.8 immediately gives the following list of antiderivatives:

Function $f(x)$	Antiderivative $\int f(x)\mathrm{d}x$
$\cos(x)$	$\sin(x) + C$
$\sin(x)$	$-\cos(x) + C$
$\exp(x) = e^x$	$\exp(x) + C$
$1/x$	$\log(x) + C$
$1/\cos^2(x)$	$\tan(x) + C$
$\frac{1}{\sqrt{1-x^2}}$	$\arcsin(x) + C$
$\frac{-1}{1+x^2}$	$\arctan(x) + C$
$\frac{e^x-1}{x}$	$\mathrm{bla}(x) + C$

Example VIII.3: The rules for sums and multiples already allow us to find antiderivatives for more complicated functions, even if they look scary at first glance. Suppose we are given $f(x) = 7\sin(x) + 4x^3 - \frac{\pi}{1+x^2} - \frac{2e^x}{x} + \frac{2}{x}$ and want to compute $\int f(x)\,dx$. We can do this by applying the sum and multiple rules along with our

[3] Such *integral tables* can be weighty and prominent tomes.

table of common integrals. Then

$$\begin{aligned}\int f(x)\,dx &= \int 7\sin(x) + 4x^3 - \frac{\pi}{1+x^2} - \frac{2e^x}{x} + \frac{2}{x}\,dx \\ &= \int 7\sin(x)\,dx + \int 4x^3\,dx + \int \frac{-\pi}{1+x^2} + \int -2\frac{e^x-1}{x}\,dx \\ &= 7\int \sin(x)\,dx + 4\int x^3\,dx - \pi\int \frac{1}{1+x^2} - 2\int \frac{e^x-1}{x}\,dx \\ &= -7\cos(x) + x^4 - \pi\arctan(x) - 2\mathrm{bla}(x) + C.\end{aligned}$$

Of course, this problem was contrived to appear complicated but actually be straightforward to solve. However, in practice, a surprising number of antidifferentiation problems really boil down to something like the above. In general if you have some functions $f_1(x), f_2(x), \ldots, f_n(x)$ whose antiderivatives you know and a function $g(x) = \sum_{i=1}^{n} c_i f_i(x)$, where $c_1, c_2, \ldots, c_n$ are some constants then the sum and multiple rules together tell us

$$\int g(x)\,dx = \sum_{i=1}^{n} c_i \int f_i(x)\,dx.$$

But since we already know each $\int f_i(x)\,dx$ the problem is solved. Problems like this are a matter of simple bookkeeping.

VIII.3 Integration by parts

For products of functions the situation is more complicated. The product rule for derivatives tells us that

$$\frac{\mathrm{d}}{\mathrm{d}x}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x)),$$

that is, a product of function gets transformed into the sum of two products. We therefore cannot revert this rule directly as a rule for products. But we can rewrite it as

$$f'(x)g(x) = \frac{\mathrm{d}}{\mathrm{d}x}(f(x)g(x)) - f(x)g'(x)$$

and thus (taking antiderivatives on both sides) get:

$$\int f'(x)g(x)\mathrm{d}x = f(x)g(x) - \int f(x)g'(x)\mathrm{d}x.$$

(There is no $+C$ on the right hand side, since an integral remains.) That is, we can transform the integral over a product $f'(x)g(x)$ into an expression involving an integral over another product $f(x)g'(x)$ that requires us to find the antiderivative of one factor and replace the other factor by its derivative. If this second integral

is easier than the first one (e.g. if the derivative of g makes some factor disappear), this can be of use. This process is called *integration by parts*

For example, consider the integral $\int \cos(x) \cdot x \mathrm{d}x$. We set $f(x) = \sin(x)$ and observe that $f'(x) = \cos(x)$ is one of the factors. Similarly we set $g(x) = x$ which gives us $g'(x) = 1$. Thus

$$\int \cos(x) \cdot x \mathrm{d}x = \sin(x) \cdot x - \int \sin(x) \cdot 1 \mathrm{d}x.$$

But we know that $\int \sin(x) \mathrm{d}x = -\cos(x) + C$. Thus

$$\int \cos(x) \cdot x \mathrm{d}x = \sin(x) \cdot x - (-\cos(x)) + C = \sin(x) \cdot x + \cos(x) + C.$$

Similarly, we can calculate $\int \exp(x) \cdot x \mathrm{d}x$: We set $f(x) = \exp(x)$ and $g(x) = x$. Then $f'(x) = \exp(x)$ and $g'(x) = 1$ and

$$\int \exp(x) \cdot x \mathrm{d}x = \exp(x) \cdot x - \int \exp(x) \cdot 1 \mathrm{d}x = \exp(x) \cdot x - \exp(x) + C = \exp(x)(x-1) + C.$$

Note that it is important, which of the factors we call $f'(x)$ and which one $g(x)$. If we had switched the roles, we would have gotten

$$\int \exp(x) \cdot x \mathrm{d}x = \frac{1}{2} x^2 \cdot \exp(x) - \int \frac{1}{2} x^2 \exp(x) \mathrm{d}x,$$

and thus a more complicated integral. And of course, there is no guarantee that it is always possible to use integration by parts in a suitable way and to obtain a "good" integral on the right hand side. It is a game of trial and error.

A final, somewhat sneaky, application can be found by setting $f'(x) = 1$ a factor that is not actually written down. In some cases this allows to integrate functions that have an "easier" derivative. For example,

$$\int \log(x) \mathrm{d}x = \int 1 \cdot \log(x) \mathrm{d}x = x \log(x) - \int \frac{1}{x} \cdot x \mathrm{d}x = x \log(x) - \int 1 \mathrm{d}x,$$

and the integral on the right hand side easily evaluates as $x + C$, thus giving a new antiderivative

$$\int \log(x) \mathrm{d}x = x \log(x) - x + C.$$

VIII.4 Substitution

A similar problem arises with compositions of functions. The chain rule

$$\frac{\mathrm{d}}{\mathrm{d}x} f(g(x)) = f'(g(x)) \cdot g'(x) \tag{VIII.4}$$

introduces an extra factor $g'(x)$ in addition to the composition. That basically means that we can take the antiderivative of a composition only if such an extra factor is there. One way around this would be to create such a factor and simultaneously to divide by it to keep the expression the same. The process of substitution is a way to process such cases systematically.

The process of doing this is called *substitution* since we replace an expression in x with a new variable (often called u).

However, in practice, it can be an art to find the "correct" (which is sometimes not obvious) expression to substitute. A regular calculus course will cover many strategies and heuristics for this, while this course will only consider basic cases (or situations in which the substitution is provided).

We first pick a sub-expression we want to substitute. Call this expression u. Since it will be a function in x we can write $u = g(x)$. We can write (While it looks as if we are working with fractions, it is a formal symbol manipulation, not really proper arithmetic.)

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}g(x) = g'(x)$$

which we can write as

$$g'(x)\mathrm{d}x = \mathrm{d}u, \text{ respectively } \mathrm{d}x = \frac{1}{g'(x)}\mathrm{d}u$$

Now suppose we want to integrate the derivative of a composition (as given in equation (VIII.4),

$$\int f'(g(x)) \cdot g'(x)\mathrm{d}x.$$

Say we decide to substitute the subexpression $g(x)$. We Identify this expression, as well as the expression $g'(x)\mathrm{d}x$, and substitute

$$\int \frac{\mathrm{d}}{\mathrm{d}x} f(\underbrace{g(x)}_{=u}) = f'(g(x)) \cdot \underbrace{g'(x)\mathrm{d}x}_{=\mathrm{d}u} = \int f'(u)\mathrm{d}u$$

We now calculate the antiderivative (in the new variable u), getting

$$= f(u) + C$$

and finally substitute back for u to get the antiderivative

$$= f(g(x)) + C$$

as desired.

For example, to calculate

$$\int (\sin(x))^2 \cos(x)\mathrm{d}x,$$

we could set $u = \sin(x)$ and thus $\cos(x)\mathrm{d}x = \mathrm{d}u$ and substitute

$$= \int (u)^2 \mathrm{d}u = \frac{1}{2}u^3 + C = \frac{1}{2}(\sin(x))^3 + C.$$

Alternatively, with the same result, we could have replaced only the $\mathrm{d}x$ as

$$\int (\sin(x))^2 \cos(x)\mathrm{d}x = \int (u)^2 \cos(x)\frac{1}{\cos(x)}\mathrm{d}u = \int u^2 \mathrm{d}u.$$

For another example take

$$\int \cos(x^2)x\mathrm{d}x$$

and substitute $u = x^2$ and $\mathrm{d}x = \frac{1}{2x}\mathrm{d}u$:

$$= \int \cos(u)\frac{x}{2x}\mathrm{d}u = \int \cos(u)\frac{1}{2}\mathrm{d}u = \frac{1}{2}\sin(u) + C = \frac{1}{2}\sin(x^2) + C.$$

What is important is that after the substitution process the old variable will have vanished completely in favor of the new variable. This might require us to use the inverse function of the substitution $g(x)$ to rewrite an x-expression in terms of u.

Say we change the last example slightly to

$$\int \cos(x^2)x^3\mathrm{d}x,$$

and substitute $u = x^2$, we get

$$= \int \cos(x^2)x^2 \cdot x\mathrm{d}x = \int \frac{1}{2}\cos(u)u\mathrm{d}u.$$

(This integral now can be solved using integration by parts.)

This resolving of remaining expressions can cause problems, and not every substitution will ultimately lead to success. Say we change the example once more to

$$\int \cos(x^2)x^2\mathrm{d}x,$$

and again substitute $u = x^2$, we get (using $x = \sqrt{u}$)

$$\int \cos(x^2)x^2\mathrm{d}x = \int \cos(u)u\frac{1}{2x}\mathrm{d}u = \int \frac{\cos(u)u}{2\sqrt{u}}\mathrm{d}u,$$

which is an integral we cannot solve either.

Substitutions often are of sub-expressions of the integrand, but can be more subtle. For example if we want to integrate

$$\int \frac{1}{\sqrt{1-x^2}^3}\mathrm{d}x = \int \frac{1}{(1-x^2)^{3/2}}\mathrm{d}x$$

we can[4] substitute $u = \sin^{-1}(x)$ (that is $x = \sin(u)$) and get $\mathrm{d}u = \frac{1}{\sqrt{1-x^2}}\mathrm{d}x$ and thus

$$= \int \frac{1}{(1-\sin^2(u))^{3/2}}\sqrt{1-\sin^2(u)}\mathrm{d}u = \int \frac{1}{(1-\sin^2(u))}\mathrm{d}u$$

which, using the fact that $\sin^2(u) + \cos^2(u) = 1$ and thus $1 - \sin^2(u) = \cos^2(u)$, gives us

$$= \int \frac{1}{\cos^2(u)}\mathrm{d}u = \tan(u) + C = \tan(\sin^{-1}(x)) + C$$

VIII.5 Definite Integrals

An important application of antiderivatives is in calculations of areas (and volumes), or more generally in summations of infinitely many terms that are infinitesimally small. In investigating this, we shall start with geometric aspects, and the connection to antiderivatives will emerge later.

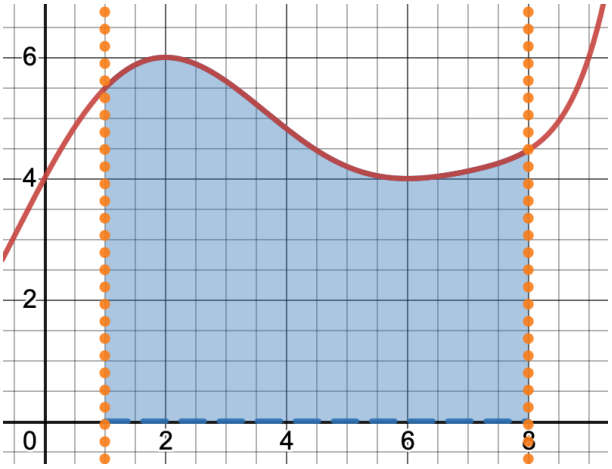


Figure VIII.1: Area below a graph

Definition VIII.5: Consider the graph of a function $f\colon \mathbb{R} \to \mathbb{R}$ for x-values in an interval $a \le x \le b$, as depicted in Figure VIII.1 for $a = 1$ and $b = 8$. The value of the area enclosed by

- The x-axis,
- the line $x = a$,

[4] This is not something you would be expected to come up with, just to be able to follow

- the line $x = b$, and
- the graph of f.

is called the *definite integral* of f from a to b and denoted by

$$\int_a^b f(x)\mathrm{d}x.$$

Obviously there is no $+C$ in the definite integral – the area is not subject to choice.

Note VIII.6: For completeness, a few extra rules and observations on the definite integral are in order:

- If $a = b$ the area is zero: $\int_a^a f(x)\mathrm{d}x = 0$.
- If $b < a$, we count the area negatively.
- Similarly, if the function $f(x) \leq 0$. The area above the x-axis is counted positively, the area below the x-axis negatively.
- We can split an area by a vertical line $x = c$ into two parts, for $a \leq c \leq b$. Then
$$\int_a^b f(x)\mathrm{d}x = \int_a^c f(x)\mathrm{d}x + \int_c^b f(x)\mathrm{d}x.$$
the area is zero: $\int_a^a f(x)\mathrm{d}x = 0$.
- If we add two functions, we can split the area into two parts, using one of the functions. That is
$$\int_a^b f(x) + g(x)\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_a^b g(x)\mathrm{d}x$$

Besides the obvious geometric applications, we can use this area for example to define the *average* of the function f on the interval from a to b as

$$\frac{\int_a^b f(x)\mathrm{d}x}{b-a}.$$

Riemann Sums

While we can define the area under the graph, we have so far no general formula from geometry to calculate it, unless f has a very particular form[5].

[5] say a straight line, or a half-circle

But we can approximate or estimate the area by covering it with smaller pieces, whose area we can calculate. In the picture in Figure VIII.1, the grid depicted helps with such an approximation, giving an area between about 32 and 36 units[6].

While it seems clear that cutting in finer pieces will give a better result, it is not clear how to do this cutting so that it will work for arbitrary functions. The approach we shall employ, named *Riemann sums*[7] is simple enough to allow for an easy description, while allowing for the definition of exact results: We split the area in vertical stripes of equal width, such that the height of the stripe matches the function value on the left side. Figure VIII.2 depicts such approximations for $n = 1, 2, 3, 4, 10, 25$ stripes. The area of the stripes becomes a better and better approximation of the area under the graph.

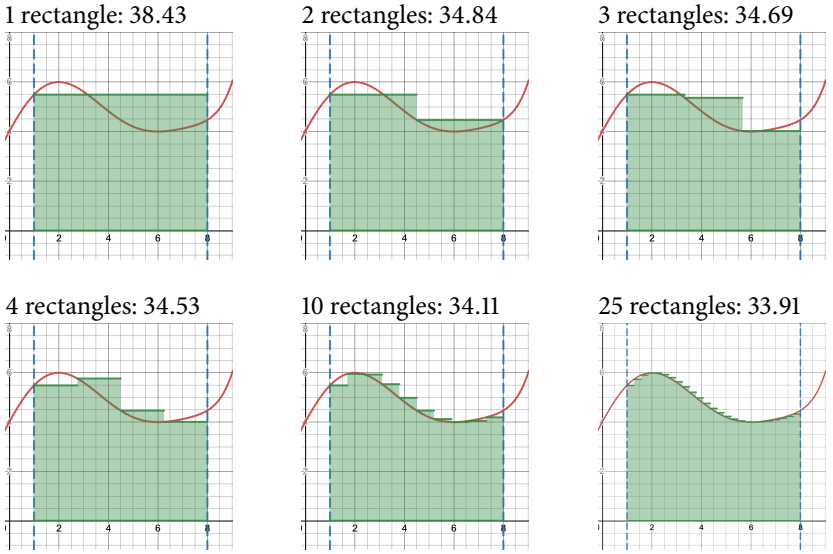


Figure VIII.2: Integral approximations by Riemann sums

Lemma VIII.7: If we split the interval into k stripes of equal width, each stripe has width $\Delta x = (b-a)/k$ and the i-th stripe starts at $x_i = a + (b-a)/k \cdot (i-1)$. The total area of the k stripes thus is

$$A_k = \sum_{i=1}^{k} f(x_i) \cdot \Delta x = \sum_{i=1}^{k} f\left(a + \frac{(b-a)(i-1)}{k}\right) \cdot \frac{b-a}{k}.$$

In our example, (beyond the pictures shown), 100 intervals give us an area of 33.81, 200 intervals give 33.79, 1000 intervals 33.779, and 10000 intervals 33.7759.

[6] while the correct area is about 33.7756 units

[7] named after the German mathematician Bernhard Riemann, 1826-1866

Note VIII.8: The decision to set the height of a stripe at the function value on the left side does not necessarily give the best approximation. Others options are right side, middle, or largest or smallest value of the function in the interval. Indeed such different strategies are all used if one wants to approximate the area numerically. For our purposes however the difference between these methods is irrelevant, as we aim to make the stripes infinitesimally narrow.

We note that, unsurprisingly, a choice of finer and finer stripes gives an increasingly better approximation of the area. Indeed, one can show if we consider the values of the stripe approximations A_k as a sequence, (A_k), indexed by k, that this sequence converges in our example.

Definition VIII.9: More generally, we call a function *integrable* on the interval from a to b, if the limit

$$\lim_{k\to\infty} A_k$$

exists (and is finite). This is for example the case if f is continuous. And the value of the definite integral then is defined as this limit:

$$\int_a^b f(x)\mathrm{d}x = \lim_{k\to\infty} A_k.$$

Note VIII.10: The summation formula in Lemma VIII.7 is the origin for the integral notation. The $\sum$ is transformed in a lengthy "S", the integral sign $\int$, and the interval width Δx becomes the end delimiter $\mathrm{d}x$.

Note VIII.11: The reader might worry that we now defined the definite integral twice – once as area, and once as limit. Indeed, the proper approach would be to define the improper integral as a limit as in definition VIII.9, show that it obeys the properties, in particular the area additivity of Note VIII.6, and show that in cases that a graph defines an area that is a geometric object defined in school, the geometric area, and the value of the improper integral agree. Doing so is not particularly illustrative, and thus we do not do so here. For areas that were never calculated in geometry class, the definite integral then *is* the definition of their area.

VIII.6 The Fundamental Theorem

The definition of the definite integral as a limit is not particularly helpful to calculating it. We thus instead use another approach that will lead us to a connection with antiderivatives.

For this, assume now that $f(x)$ is an integrable function. We define a new function that gives values of certain definite integrals:

$$F(z) = \int_0^z f(x)\mathrm{d}x.$$

(The choice of 0 is arbitrary, one could chose any other number instead.) Note that we need to call the integration variable (here x) differently from z to avoid confusion of variables.

That is, the function F assigns to every number z the value of the definite integral from 0 to z over the function f; that is the area under the graph of f from 0 to z.

We now consider how $F(z)$ changes when z changes:

THEOREM VIII.12 (Fundamental Theorem of Calculus):
a)The function $F(z)$ is differentiable with respect to the variable z, and we have that

$$\frac{\mathrm{d}}{\mathrm{d}z}F(z) = f(z),$$

that is $F(x)$ is an antiderivative of $f(x)$.
b) For any $a, b \in \mathbb{R}$, we have that

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a).$$

c) If G is *any* antiderivative of f, we have that

$$\int_a^b f(x)\mathrm{d}x = G(b) - G(a).$$

To show the connection and to save space, we write $F(x)\Big|_a^b = F(b) - F(a)$.

Proof: a) We use the definition of the derivative as limit of the difference quotient:

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = \lim_{h\to 0}\frac{F(x+h) - F(x)}{h}$$

The difference in the numerator is the difference of the two areas, as depicted in green in Figure VIII.3. As $h \to 0$, this area is approximated by a single vertical strip of width h and height $f(x)$, and thus area $f(x)\cdot h$. The quotient thus has the limit $f(x)$.
b) By the rules for additivity of area, we have that

$$\int_a^b f(x)\mathrm{d}x = \int_0^b f(x)\mathrm{d}x - \int_0^a f(x)\mathrm{d}x = F(b) - F(a).$$

c) If $G(x)$ is another antiderivative, we have that $G(x) = F(x)+c$ for some constant c. But then

$$G(b)-G(a) = F(b)+c-(F(a)+c) = F(b)-F(a)+c-c = F(b)-F(a) = \int_a^b f(x)\mathrm{d}x.$$

□

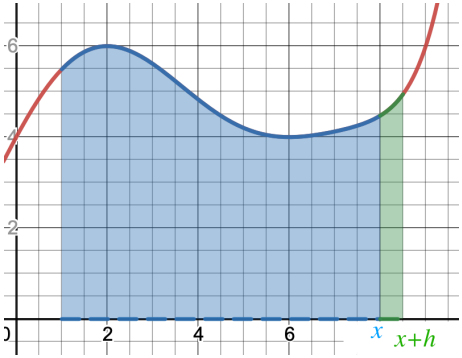


Figure VIII.3: Proof of the Fundamental Theorem

Example VIII.13:

$$\int_0^{\pi} \sin(x)\mathrm{d}x = (-\cos(x))\Big|_0^{\pi} = -\cos(\pi) - (-\cos(0)) = -(-1) - (-1) = 2$$

and (verify the antiderivative by differentiation!)

$$\int_{-1}^{1} \sqrt{1-x^2}\mathrm{d}x = \left(\frac{1}{2}\left(x\sqrt{1-x^2} + \arcsin(x)\right)\right)\Big|_{-1}^{1} = \frac{1}{2}(\pi/2 - (-\pi/2)) = \frac{\pi}{2}$$

the area of a half-circle (thus verifying the circle-area formula from school).

Some examples of applications of definite integrals are:

Areas and Volumes If we can describe an regions using functions, we can often use integrals to calculate their areas. Building on this we can calculate not only the volumes of prisms (area of the base times height), but also volumes of objects that can be built from areas in a regular way – pyramids and cones, or volumes obtained by rotating a region through higher dimensional space.

Averages We have already seen that one can define an "average" value of a function over an interval as the quotient of the definite integral by the interval length. In similar way one can calculate other statistical measures, determine centers of mass, or renormalize functions to make them comparable.

Infinite summations The definition of definite integrals is as limit of a sum. There are other measures that can be described as such a limit – for example the length of a curve, or a more complicated surface area, and the limit expression then can be interpreted as a definite integral.

Geometry on Functions What is probably the most important application of definite integrals for computer science might also initially seem the most cryptic one. You will see in Linear Algebra classes, that definite integrals can be used to define an "length" of functions on an interval $[a, b]$. Concretely, the length of the functions f can be defined as

$$\sqrt{\int_a^b f(x)^2 \mathrm{d}x}$$

Using the concept of length, it is possible to continue in defining angles, orthogonality, projections, close approximations, and ultimately use the tools of geometry to investigate functions.