Commentaries to Foundations of Analysis

1 Proofs and definitions

The main goal of the course is for the student to learn to read and understand mathematics. This is no easy thing and the course is generally considered to be quite "heavy". One reason for the difficulty in learning how to read mathematics is that mathematical notation has been developed for a very long time. The result is that there are a large number of conventions that are intended to to help readers but of course won't unless they understand them.

There is however a more important reason for the difficulties. The problem is that mathematics does not deal in anything specific that one may refer to (contrary to for instance the different natural sciences). For psychological/pedagogical reasons it is however important to create one's own images (that are not necessarily visual) of various mathematical notions. Exactly because there is no "reality" to refer to these images become very individual and it is almost impossible to use them to discuss mathematics with others or to write about mathematics.

The solution to this problem, in particular when written mathematics is concerned, is to confine oneself to a very precise and rather formal language whose interpretation everyone learns to agree upon. This means that in order to be able to read mathematics one must

- learn to understand this formal and precise common language,
- create one's own language or images and
- learn to translate between the common formal language and one's private.

Note that it is a necessity to create one's *own* language. This means that the teacher's attempt to explain a mathematical text should not be considered the "correct" explanation but at most a possible one. It is always necessary for the student to make up a personal explanation. In the beginning this will be difficult as the personal language has not been yet been developed. The only way to proceed in such a development and at the same time learn how to translate between the common and private languages is to work with mathematical definitions, statements and proofs. In the beginning one should expect to work a lot with them, to the point that one will as a consequence have learnt them by heart. It is important to realise however that that is not a goal in itself but that the development of one's own language that can be used to express mathematical notions is. Once this has been accomplished one will not have any need for learning them by heart.

As has already been said it is important to understand that the language and the images one will finally arrive at is one's own and that the suggestions of others do not necessarily suit oneself. There are however some principles that seem to help most people to learn how to analyse definitions, statements and proofs and we shall in these notes take up some of them.

The very first thing one should do when reading a mathematical definition or statement is to *understand what it says*. That means for instance that one checks that one understands all the terms involved, that one understands what is assumed and what is claimed under the given assumptions. In particular one should make clear the *exact* formulations; there is for example a difference between "for all x there exists a y" and "there is a y such that for every x".

If it is then a question of a definition one can continue by

- try to find some examples that fulfil the conditions of the definition,
- try to find some examples that do *not* fulfil the conditions of the definition and
- try to understand the purpose of the definition.

As far as examples and non-examples of a definition are concerned it is important to find such not just to understand a definition but also for the future. When a definition is applied as part of a new definition, statement or proof one needs to understand the new ones. Having good examples as well as non-examples of the original definition is then likely to be a help. What is meant by "good examples" varies fo course but as a general principle it is important on the one hand not to have too many examples, on the other hand that they cover all important aspects of the definition. If one only picks one example there is a risk that one will fixate too much on it and that one, inadvertently no doubt, lets special properties of the particular example sneak in, properties that do not follow from the definition. Too many examples on the other hand may make them unusable.

Non-examples on the other hand are there to mark the boundaries of the definition and it is thus important to choose them as close as possible to examples that do fulfil the definition.

Finally, understanding the purpose of a definition is a rather vague task. On the one hand, there probably is one as otherwise one would hope that noone would have bothered to introduce the definition. The purpose may on the other hand not be clear immediately. It could be that only later in the text will there be a use of the new notion.

Let us look at an example of a definition. We begin by giving a version whose purpose is to make it as difficult as possible to understand its purpose.

Definition 1.1 *A* field *is a quintuple* (K, p, m, a, b) *where*

- K *is a set,*
- p *is a function* $p\colon K \times K \to K$*,*
- m *is a function* $m\colon K \times K \to K$*,*
- a *and* b *are elements of* K*.*

These data must fulfil the following conditions for all $x, y, z \in K$*:*

1. $p(x, y) = p(y, x)$.
2. $p(x, p(y, z)) = p(p(x, y), z)$.
3. $m(x, y) = m(y, x)$.
4. $m(x, m(y, z)) = m(m(x, y), z)$.
5. $m(p(x + y), z) = p(m(x, z), m(y, z))$.
6. $p(a, x) = x$.
7. $m(b, x) = x$.
8. *There is an* x' *such that* $p(x, x') = a$.
9. *If* $x \neq a$ *then there is an* x'' *such that* $m(x, x'') = a$.

This definition is difficult to penetrate. It is not even easy to find examples (non-examples are easier!) that fulfil it and the purpose is not clear at all. We make a new attempt where we formally give the same definition but formulated in such a way so as to make it easier to understand.

Definition 1.2 *A* field *is a quintuple* $(K, +, \cdot, 0, 1)$ *where*

- K *is a set,*
- $+$ *is a function* $+\colon K \times K \to K$*,*

- $\cdot$ *is a function* $\cdot\colon K \times K \to K$,
- 0 *and* 1 *are elements of* K.

These data must fulfil the following conditions for all $x, y, z \in K$*:*

1. *(Commutativity for addition)* $x + y = y + x$.
2. *(Associativity for addition)* $x + (y + z) = (x + y) + z$.
3. *(Commutativity for multiplication)* $x \cdot y = y \cdot x$.
4. *(Associativity for multiplication)* $x \cdot (y \cdot z) = (x \cdot y) \cdot z$.
5. *(Distributivity)* $(x + y) \cdot z = x \cdot z + y \cdot z$.
6. *(Unit element for addition)* $0 + x = x$.
7. *(Unit element for multiplication)* $1 \cdot x = x$.
8. *(Additive inverse) There is an* x' *such that* $x + x' = 0$.
9. *(Multiplicative inverse) If* $x \neq 0$ *then there is an* x'' *such that* $x \cdot x'' = 1$.

Now it looks much more understandable. Note to begin with that we have written the function values in so called *infix form*, i.e., what really should be written $+(x, y)$ (just as we wrote it $p(x, y)$ in our first definition) we write as $x + y$. This is an example of a mathematical convention; if one uses (very) special function names such as $+$ and $\cdot$ one may (and almost always does) use the infix notation.

Similarly, the choice of the name '+' for the first function, '$\cdot$' for the second and 0 and 1 for the two elements is no coincidence. In fact, with these choices it suddenly becomes very simple to find examples of quintuples that fulfil these conditions. We have made these even clearer by writing small comments in front of each conditions.

Exercise 1: i) Find examples of fields.

ii) Find quintuples fulfilling all conditions except

- the last,
- the last two,
- the second to last,
- the third and the last and
- the third.

(The last case is difficult.)

There are more examples than the ones that first comes to mind and as one of them is somewhat peculiar it is good to have among one's collection of examples.

Exercise 2: Let K be the set $\{0, 1\}$ and define addition by

$+$	0	1
0	0	1
1	1	0

and multiplication by

$\cdot$	0	1
0	0	0
1	0	1

.

Show that the quintuple $(K, +, \cdot, 0, 1)$ is a field.

If one then considers the purpose of this definition, it would seem to be reasonably obvious. One could summarise it by saying that a field is a set for which one has access to the four basic arithmetic operations and all the rules among them that we are used to. The example given in the last exercise shows however that one has to be a little bit careful. We are for example not allowed to assume that $1 + 1$ is different from 0.

Remark: There is a little bit more to the purpose of a definition. A definition should also be useful in that it has interesting consequences. The notion of fields is indeed very useful. One example of that is that a very large part of linear algebra (which one first only meets over the real, or possibly complex, numbers) makes sense and is true for all fields. This has many applications.

We shall now have a look at statements and their proofs. For statements the same holds as for definitions: First one must understand what they are saying and what needs to be done to do that is roughly the same; understand all terms involved, figure out assumptions and conclusions and so on. The next step is to understand what they mean but this stage differs from the corresponding stage for definitions. In the beginning the difference is not so large however. It may be a good thing to do the following.

- Try to find some examples of situations where the assumptions are fulfilled and try to understand which consequences the statement gives in such examples.
- Try to find some examples of situations where the assumptions are *not* fulfilled (and for this step to be useful, the conditions should be close to being fulfilled). If one can come up with examples where the conditions are almost fulfilled but the conclusion is false it is even better.

What one then can try to do is to see if one can see what the statement is good for. This can be quite difficult and may not even be clear when one is reading the statement.

2 Ett analyserat bevis

We shall show how to use the supremum axiom to show that there is a positive real number r such that $r^2 = 2$. At the same time we shall see how one can organise a proof. Just like in "real life" we start with a lemma whose significance will be made clear only in a little while.

Lemma 2.1 *i) Every downwards bounded set (of real numbers) has a largest lower bound.*
ii) If r and s are positive real numbers and $r^2 < s^2$, then $r < s$.
iii) For all real numbers $r > 0$ we have that $(r^2 + 4/r^2)/4 \geq 1$.

Proof. For the first statement we assume that S is a set of real numbers that is downwards bounded by M, i.e., if $s \in S$ then we have that $M \leq s$. Consider now the set $S' := \{-s \mid s \in S\}$. As $-s \leq -M$ for all $s \in S$ we have that S' is upwards bounded by $-M$ and according to the supremum axiom there is a smallest upper bound S'. Then $-N$ is a largest lower bound for S.

For the second part we may assume that $0 \leq s \leq r$. From this it follows $s^2 \leq r^2$ which is a contradiction.

For the last statement we have

$$0 \leq (r - 2/r)^2 = r^2 - 4 + 4/r^2$$

which gives $1 \leq (r^2 + 4/r^2)/4$ by moving the four to the other side and then dividing by 4. □

One problem with proofs is that they are adapted to the imagined reader and hence a proof having one type of audience in mind may skip a lot of details that wouldn't be skipped otherwise. This is usually a good thing as a proof with less unnecessary details is easier to read. It does however require a reader in training to realise that usually some details are skipped.

Exercise 3: Give an account of the details that have been skipped in the proof and write a more careful proof that fills them in.

We are now ready to show our result.

Proposition 2.2 *There is a positive real number r such that $r^2 = 2$.*

Proof. Consider the set $S := \{s \in \mathbf{R} \mid s > 0, s^2 \geq 2\}$. This set is bounded downwards by for instance 0 and according to (2.1:i) it has a largest lower bound, r say, which is positive as 0 is a lower bound of S. We shall now show the following:

- It is not true that $r^2 < 2$.
- It is not true that $r^2 > 2$.

Together these statements show that $r^2 = 2$.

If we start with the first statement we may assume that $r^2 < 2$. Now put $\epsilon := (2 - r^2)/5$ which according to the assumption is a positive real number. According to (2.1:ii) we have $r < 2$ because $r^2 < 2 < 2^2$ and thus $\epsilon \leq 1$ which implies that $2r + \epsilon \leq 5$ and hence that $\epsilon(2r+\epsilon) \leq (2-r^2)/5 \cdot 5 \leq 2-r^2$. We have therefore that $(r+\epsilon)^2 = r^2+\epsilon(2r+\epsilon) \leq r^2+2-r^2 = 2$ but this gives that for $s \in S$ we have that $(r+\epsilon)^{\leq} 2 \leq s^2$ and by (2.1:ii) we get that $r + \epsilon \leq s$ so that $r+\epsilon$ is a lower bound but as $\epsilon > 0$ it is a greater lower bound than r which is a contradiction.

If we instead assume that $r^2 > 2$ we have that $2/r < r$ which means that if we put $t := (r + 2/r)/2$ then we have that $t < r$. We further have that $t^2 = (r^2 + 4 + 4/r^2)/4 \geq 1 + (r^2 + 4/r^2)/4 \geq 1 + 1 = 2$, where for the last inequality we have used (2.1:iii), so that $t \in S$ but as $t < r$ we have that r is not a lower bound. □

When analysing a proof suspicion (paranoia even) is an appropriate frame of mind. It is necessary to question *every* statement that is made in the proof and get to understand why it is true. This is also where a proof suitable for an unexperienced reader should be more detailed than one written for a more experienced one; an experienced reader will be able to fill in small details immediately and will only be disturbed by extra details.

Exercise 4: i) Give an account of skipped details of the the proof and write a more complete one that fills them in.

ii) Mark where the different results of the lemma are used in the proof. Rewrite the proof so that what is proved in the lemma is instead proved within the proof of the proposition. Compare readability of the two approaches.

iii) Try to explain how one arrives at the choice $\epsilon := (2 - r^2)/5$.

iv) Trye to figure out why one puts $t := (r + 2/r)/2$.

v) Show that for each $a > 0$ there is an $r > 0$ such that $r^2 = a$.

3 Cardinality

A comment in section 2.3 makes some claims that are not is clear as the author's "clearly" would suggest. Fact is that it is quite instructive to go through the argument that is hidden behind these words. The section defines a relation $A \sim B$ between sets; $A \sim B$ if there is a bijection $f\colon A \to B$ and claims the following for it:

- $A \sim A$ for all sets A.
- If $A \sim B$ then $B \sim A$.
- If $A \sim B$ and $B \sim C$ then $A \sim C$.

It is often a good idea to think of mathematical staement and their proofs in terms of a game between two persons where a proof gives a strategy that guarantees a win for the first player. A round for the first statement ($A \sim A$) would look like this (at each step the player about to make a move may concede defeat but we will not mention that explicitly):

1. Player 1 makes the claim $A \sim A$ for all A.
2. Player 2 now chooses a specific set A.
3. Player 1 must now show that for the choice of A made by player 2 we have that $A \sim A$. According to the definition of the relation $A \sim A$ this means that player 1 must choose a function $f\colon A \to A$ and make the claim that it is a bijection.
4. Player 2 can now aim at either showing that f is not onto or that it is not 1-1. In the former case an $a \in A$ is chosen by player 2.
5. Player 1 must now choose a $b \in A$ such that $f(b) = a$. Once this has been done the round has been won by player 1.
6. The second possibility is that player 2 aims at questioning that $f$is 1-1. In that case the player must choose two elements $a, a' \in A$ such that $f(a) = f(a')$.
7. Player 1 now wins if $a = a'$ and player 2 wins if $a \neq a'$.

If player 2 should happen to start a round by choosing $A = \{1, 2, 3\}$ then player 2 has many different ways of choosing different (6 to be precise) bijections from A to A and will then win by choosing any of them. This however is done by "improvisation", i.e., it depends on the precise nature of the particular choice of A. To prove $\forall A\colon A \sim A$ we must come up with a strategy that is guaranteeing that player 1 wins no matter which A is chosen (and there are *many* very strange sets). That player 1 has specified a function f means that if player 2 chooses an $a \in A$, player 1 must be able to find a $b \in A$ such that $f(b) = a$. The only thing that player 1 can assume about A at that stage is that a is in it. Hence it looks difficult to come up with another element b (of course f could be chosen very cleverly but as we know nothing about A we can not hope to be able to be very clever about its choice). This argument is not something that would be a part of a winning strategy but instead part of an argument that, hopefully, will lead to one. We have therefore come up with a candidate strategy not because we are sure it is a good one but because it seems to be the only possible one. The strategy is that no matter what set A player 2 will come up with player 1 should choose the *identity functino*, id, that is defined by $\mathrm{id}(a) = a$. This is not a complete strategy however only its first step; we must also specify how player 1 should react at the following steps.

- If player 2 chooses to question that the function is onto and presents player 1 with some $a \in A$ then player 1 will win as a can be an answer and indeed $\mathrm{id}(a) = a$.
- If player 2 chooses to question that the function is onto 1-1 and send back $a, a' \in A$ with $\mathrm{id}(a) = \mathrm{id}(a')$ then player 1 wins as $a = \mathrm{id}(a) = \mathrm{id}(a') = a'$.

We thus see that we have found a winning strategy for player 1. When one is writing a mathematical proof it is (usually) not formulated in terms of games and strategies but it is not difficult to see how close a more standard formulation lies to a gametheoretic one. This can be seen if we write down a proof of $A \sim A$ in traditional terms.

Given a set A we consider the identity function $\mathrm{id}\colon A \to A$. *This is a bijection because given $a \in A$ we have* $\mathrm{id}(a) = a$ *which shows that*id *is onto and given $a, a' \in A$ with* $\mathrm{id}(a) = \mathrm{id}(a')$ *we have that $a = \mathrm{id}(a) = \mathrm{id}(a') = a'$ and thus* id *is* 1-1.

We even see that the word given corresponds exactly to what the opponent, player 2, gives us, player 1 as data that are part of a move.

As further illustration let us consider the third part $A \sim B \wedge B \sim C \Rightarrow A \sim C$. In this game the first of move of player 2 is to produce three sets A, B och C together with two bijections $f\colon A \to B$ and $g\colon B \to C$. Note that suddenly player 2 is in a more exposed situation; player 1 may now claim that f or g are bijections and player 2 must accept that. In any case player 1 is now forced to in one way or other cook up a function $h\colon A \to C$ and must be able to defend that it is a bijection. The player can, it seems, only expect to be able to use f and g to do that and it is difficult to see how to construct a function $A \to B$ in any other way than to compose them; $h = g \circ f$. If this now is the move of player 1, player 2 may now choose to question either that h is 1-1 or that it is onto. In the first case player 2 will return $a, a' \in A$ such that $h(a) = g(f(a))$ is equal to $h(a') = g(f(a'))$. However, player 2 has guaranteed that g is 1-1 and as $g(f(a)) = g(f(a'))$ player 2 is forced to agree when player 1 claims that $f(a) = f(a')$. As player 2 also has guaranteed that f is 1-1 again the player is forced to accept that $a = a'$ and player 1 wins.

If instead player 2 questions that h is onto and send a $c \in C$ to player 1 the latter is allowed to insist that player 2 must deliver a $b \in B$ such that $g(b) = c$ as player 2 is the guarantor for the fact that g is onto. Player 1 can then insist that player 2 choose a $a \in A$ with $f(a) = b$ as again player 2 has guaranteed that f is onto. Player 1 now wins by delivering a as $h(a) = g(f(a)) = g(b) = c$. That means that we have found a winning strategy.

Exercise 5: Rewrite this winning strategy into a proof in the ordinary sense.

Exercise 6: Find a winning strategy for $A \sim B \Rightarrow B \sim A$ and write it as an ordinary proof.

Exercise 7: That a function $f\colon A \to B$ is a bijection is the same thing as it having an inverse $g\colon B \to A$, i.e., the composites $g \circ f$ and $f \circ g$ are the identity functions on A resp. B. When a player claims that a function is a bijection the other player may instead demand that the first player produce an inverse. Find winning strategies for the three games where the players make such demands instead och rewrite them as ordinary proofs.

4 Mathematical notation

Mathematical notation is complicated as it on the one hand aims at extreme precision and on the other tries to be as readable as possible. This means for instance that it often uses certain conventions that however can not be automatically assumed to be in play. An example is that one often use m or n to denote integers and r or s to denote real numbers. This does *not* mean that one can write "Consider r" and then assume that one has made clear that r must be a real number. One must rather say "Consider $r \in \mathbf{R}$". It is on the other hand a good idea to use r (or s...) for a real number as it makes it much easier for a reader to understand what is going on.

Remark: A very striking example on how deeply rooted these conventions are is the integral

$$\int_1^2 e^x \, de.$$

It is perfectly legitimate to use any name for the variable of integration and once e has been chosen it supercedes in the integrand any special meaning e might otherwise have (such as being the base of the natural logarithm...). It is clearly however a bad idea to use e in this manner and the way to deal with this integral is probably to immediately replace the variable of integration so as to arrive at (for instance)

$$\int_1^2 y^x \, dy.$$

4.1 Sequences

As somewhat more serious example of notation is for sequences. The general principle is that if a is a sequence of elements of a set S then a_n is the n'th element of the sequence. Note that one speaks also of "The sequence (a_n)". This is really the same thing, one may even write "The sequence $a = (a_n)$" and then it is exactly the same thing; one gives the sequence a name and at the same one gives the name a_n to the n'th element of the sequence. The advantage of the (a_n)-notation is flexibility. One can to begin with use (n^2) to denote a sequence a that has $a_n = n^2$. One can then reindex the sequence and speak of the sequence (a_{2n}) which with the given definition has $(2n)^2 = 4n^2$ as its n'th element. The notation (a_n) does however have its disadvantages. The first thing that can confuse is that (a_p) is exactly the same sequence as (a_n); in the first case one has a sequence whose p'th element equals a_p and in the second a sequence whose n'th element is a_n. It is thus not possible to use different index variables to define different sequences. Things can become even more complicated if S has some special form. A confusing example is when S itself is a set of sequences of another set T. This means that if a is a sequence of elements of S, then each individual element a_n is also a sequence (of elements of T). It is likely that one sooner or later needs some kind of notation for those elements. Luckily we do not need to come up with some new notation for them. We have already said that if x is a sequence, then x_m denotes its m'th element. Thus the m'th element of a_n is denoted as a_{nm} but as it is somewhat difficult to quickly see what this should mean one usually throws in some parentheses, $(a_n)_m$. One could of course also write this as $a_{m,n}$ but way of writing is not part of the standard conventions so one should explicitly tell what it means. The advantage of $(a_n)_m$ is that there one does not need to say anything extra as its meaning follows from a single (standard) rule. This rule can then be combined with other rules so that $(n^2)_4$ means the fourth element of the sequence (n^2), i.e., $(n^2)_4 = 4^2 = 16$.

4.2 Existential notation

Particularly in analysis statements of the form "For all ... there exists ... such that ...". It is extremely important to understand such statement and in particular to understand how seemingly minute variations in formulations may have a huge importance. An example of such variations is the difference between punctual and uniform convergence:

- *If* $f, f_n \colon X \to Y$ *are functions where* Y *is a matric space, then* f *is the* pointwise limit *of* (f_n) *if*
$$\forall \epsilon > 0 \forall x \in X \exists N \colon d_Y(f_n(x), f(x)) < \epsilon \text{ when } n > N.$$

- *If* $f, f_n \colon X \to Y$ *are functions where* Y *is a matric space, then* f *is the* uniform limit *of* (f_n) *if*
$$\forall \epsilon > 0 \exists N \forall x \in X \colon d_Y(f_n(x), f(x)) < \epsilon \text{ when } n > N.$$

These two definitions do indeed look very much alike even though a close look reveals that there is a difference; for pointwise convergence we have "for all x there is an N" while for uniform convergence we have "there is an N such that for all x". This is of course a difference but it may not be clear what it means. One way of getting a hold of that is to think in terms of a game between two persons as above.

Pointwise convergence:

1. I claim that (f_n) converges pointwise towards f.
2. My opponent gives me an $\epsilon > 0$ *and* an $x \in X$.
3. I win if I can produce an N such that for all $n > N$ we have $d(f_n(x), f(x)) < \epsilon$.

Uniform convergence:

1. I claim that (f_n) converges uniformly towards f.
2. My opponent gives me an $\epsilon > 0$.
3. I now produce an N.
4. My opponent wins if there is an $x \in X$ and an $n > N$ such that $d(f_n(x), f(x)) \geq \epsilon$.

This makes the difference clearer but there is an even better way to make it clear: In a "for all ... there exists ..." one first chooses something and then something else is supposed to exist, i.e., what is to exist may depend on the first choice. This dependence can be made clear in the notation. For pointwise convergence this results in

$$\forall \epsilon > 0 \forall x \in X \exists N_{\epsilon,x} \colon d_Y(f_n(x), f(x)) < \epsilon \text{ when } n > N_{\epsilon,x},$$

where it has been made clear that N is allowed to depend on both ϵ and x whereas for uniform convergence one has

$$\forall \epsilon > 0 \exists N_\epsilon \forall x \in X \colon d_Y(f_n(x), f(x)) < \epsilon \text{ when } n > N_\epsilon,$$

where it is now clear that N is only allowed to depend on ϵ. If the expressions are complicated one may allow oneself to skip the indices so that for instance one might write

$$\forall \epsilon > 0 \forall x \in X \exists N_{\epsilon,x} \colon d_Y(f_n(x), f(x)) < \epsilon \text{ when } n > N,$$

where $N_{\epsilon,x}$ are N are the same number. One may however *not* write

$$\forall \epsilon > 0 \forall x \in X \exists N \colon d_Y(f_n(x), f(x)) < \epsilon \text{ om } n > N_{\epsilon,x},$$

it must be completely clear that N may depend on ϵ and x and one must thus put them in when N is *introduced*.

5 More analysed proofs

We shall now have a look at some of the proofs in the book and make an attempt at analysing them.

5.1 The Weierstrass approximation theorem

The proof of the Weierstrass approximation theorem (Sats 7.26) is instructive as it quite naturally can be divided up into smaller steps.

- An initial reduction is made to the interval $[0, 1]$. (The book does not say why but it is a matter of considering $t \mapsto f((1-t)a + tb)$ on the interval $[0, 1]$.)
- A second initial reduction is to the case when $f(0) = f(1) = 0$. This allows us to define f for all real numbers by setting its value to 0 outside of $[0, 1]$ and f is then still uniformly continuous.
- The proof then consists in writing down an explicit formula for an approximating sequence:

$$P_n(x) := \int_{-1}^{1} f(x+t)Q_n(t)\,dt.$$

- Two things must no be shown; that P_n is a polynomial and that $P_n \to f$ uniformly. Interestingly enough these two facts are true for completely different reasons.

- To show that P_n is a polynomial we first note that as f is 0 outside $[0,1]$ we get the same integral if we modify the interval of integration a little bit.

$$\int_{-1}^{1} f(x+t)Q_n(t)\,dt = \int_{-x}^{1-x} f(x+t)Q_n(t)\,dt$$

and we may then make a change of variable $t \to t-x$ so that we get

$$P_n(x) = \int_0^1 f(t)Q_n(t-x)\,dt.$$

As Q_n is a polynomial we may write $Q_n(t-x)$ as

$$Q_n(t-x) = \sum_i Q_n^i(x)t^i$$

where $Q_n^i(x)$ are polynomials which gives

$$P_n(x) = \sum_i Q_n(x)\int_0^1 f(t)t^i\,dt$$

which in turn shows that P_n is a polynomial.

- To show that $P_n \to F$ uniformly we use three properties of Q_n:
  1. $Q_n(x) \geq 0$ for all x.
  2. $\int_{-1}^{1} Q_n(t)\,dt = 1$.
  3. For each $\delta > 0$ Q_n converges uniformly to 0 on the set $[-1,1]\backslash]-\delta,\delta[$.

- Let us show that for a sequence of (integrable) functions Q_n with these properties and a continuous function f on $[0,1]$ with $f(0) = f(1) = 0$ we have that $P_n \to f$ uniformly, where

$$P_n(x) = \int_{-1}^{1} f(t+x)Q_n(t)\,dt,$$

and where we let f be 0 one $[-1,1]$ outside $[0,1]$. This makes f a continuous function on $[-1,1]$. We can thus put $M := \sup_{x\in[-1,1]} |f(x)|$ and choose $\delta > 0$ so that $|f(s)-f(t)| < \epsilon$ if $|s-t| \leq 2\delta$ (for $s,t \in [-1,1]$). Finally we choose N such that if $n \geq N$ then we have $Q_n(x) < \epsilon$ if $|x| \geq \delta$. In that case we have, for$n \geq N$,

$$\begin{aligned}
|P_n(x) - f(x)| &\overset{1)}{=} \left|\int_0^1 (f(t+x)-f(x))Q_n(t)\,dt\right| \overset{2)}{\leq} \\
\left|\int_{|t|\geq\delta} (f(t+x)-f(x))Q_n(t)\,dt\right| &+ \left|\int_{-\delta}^{\delta} (f(t+x)-f(x))Q_n(t)\,dt\right| \overset{3)}{\leq} \\
\int_{|t|\geq\delta} |f(t+x)-f(x)|Q_n(t)\,dt &+ \int_{-\delta}^{\delta} |f(t+x)-f(x)|Q_n(t)\,dt \overset{4)}{<} \\
4M\epsilon + \epsilon &\int_{-\delta}^{\delta} Q_n(t)\,dt \leq 2M\epsilon + \epsilon = (2M+1)\epsilon
\end{aligned}$$

and the right hand side tends to 0 when $\epsilon \to 0$. Equality 1) follows from

$$f(x) = f(x)\cdot 1 = f(x)\int_{-1}^{1} Q_n(t)\,dt = \int_{-1}^{1} f(x)Q_n(t)\,dt$$

and inequality 2) follows from a division of the interval of integration into $\{t \in [-1,1] \mid |t| \geq \delta\}$ (which consists of two intervals) and $[-\delta, \delta]$ and use the triangle inequality. After that we get inequality 3) from the fact that one may estimate an integral by the integral of the absolute value of the integrand together with the fact that $Q_n \geq 0$. Inequality 4) is true as we have that $|f(t+x) - f(x)| \leq 2M$, $Q_n(t) < \delta$ för $|t| \geq \delta$ and the length of $\{t \in [-1,1] \mid |t| \geq \delta\}$ is < 2 together with the fact that $|f(t+x) - f(x)| < \epsilon$ when $|t| \leq \delta$. The last inequality follows from

$$\int_{-\delta}^{\delta} Q_n(t) \leq \int_{-1}^{1} Q_n(t) = 1$$

as $Q_n(t) \geq 0$.

- The remaining somewhat tricky part that remains is the construction of the polynomials Q_n fulfilling the three conditions. The polynomials $((1-x^2)^n)$ fulfil 1) and 3) but not 2) and we must *normalise* them by multiplying them by a constant c_n so that 2) is fulfilled. The problem is that if the c_n should happen to be too large we may destroy 3). More precisely $c_n(1-x^2)^n$ is bounded by $c_n(1-\delta^2)^n$ in $[0,1]\backslash]-\delta, \delta[$ so the condition is that $c_n a^n \to n$ when $n \to \infty$ for all $0 \leq a < 1$. The estimate in the book gives $c_n \leq \sqrt{n}$ which implies this (with large margins).

It might be interesting to think about what kind of construction $\int_{-1}^{1} f(x+t)Q(t)\,dt$ is. To give it a form that looks like the one we obtained after the change of variables one often make a change of variables $t \mapsto -t$ to make it equal to $\int_{-1}^{1} f(x-t)Q(-t)\,dt$ and one then uses $Q(-t)$ instead of $Q(t)$. Apart from that one also extends the interval of integration to $]-\infty, \infty[$. (This does not change the integral as f is 0 outside of $[0,1]$.) This gives an example of the so called *convolution*

$$(f * g)(x) := \int_{-\infty}^{\infty} f(x-t)g(t)\,dt.$$

(Some conditions are required to make this integral convergent. That f is 0 outside of a finite interval is more than enough.) If one makes the change of variables $s = x - t$ we get

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-t)g(t)\,dt = \int_{-\infty}^{\infty} f(s)g(x-s)\,ds = (g * f)(x).$$

(Again some assumptions must be made in order for the change of variables to be permissible, that f or g vanishes outside a finite interval is enough.) This relation is exactly the relation we used to show that P_n was a polynomial. If we then want to understand the approximation part, it is useful to consider a discrete version of the convolution. Let therefore (a_n) be a sequence of real numbers where n runs over *all* integers and let (b_n) be another such sequence. We assume that all but a finite number of b_n are different from 0 and defines the convolution sequence $((a*b)_n)$ by

$$(a * b)_n := \sum_i a_{n-i} b_i.$$

This looks like an infinite sum but as only a finite number of the b_i are non-zero it is in reality finite (it would also be enough that only a finite number of the a_i are non-zero).

Exercise 8: Show that if a and b are two sequences, where one of them has only a finite number of non-zero values, then we have $a * b = b * a$.

The discrete convolution has many applications. In some of these applications one imagines that (a_n) specifies some kind of signal (for instance a sound signal) and the convolution is expected so "smooth" the signal with the aid of the (b_n). This means that

- only values that are close to a given position should affect the new value in the same position, and thus the b_i should be different from zero only if $|i|$ is small,

- we do not want to turn the signal "upside down" and thus we should have $b_n \geq 0$ for all n and
- we only want to smooth out the signal not increase or decrease it which is true if (and only if) $\sum_i b_i = 1$.

One doesn't need to make a convolution in one dimension only but one may for instance consider sequences $(a_{m,n})$ and $(b_{m,n})$ in two discrete variables and the convolution given by

$$(a * b)_{m,n} := \sum_{i,j} a_{m-i,n-j} b_{i,j}.$$

In this case $a_{i,j}$ might specify the amount of gray in a picture $b_{i,j}$ could be user in an attempt to improve resolution. A given resolution corresponds to a given size of pixels, i.e., the picture is divided into squares of a certain size where the grayness level is constant in each square (see Fig. 1). Each square can then be divided up in (for instance) four subsquares. The grayness level

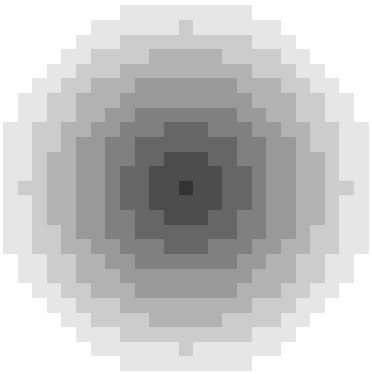

Figure 1: Low resolution picture.

of each smaller square is then modified by weighting it with the levels of neighbouring squares. The weight factors can for instance be specified by the following matrix.

0.05	0.1	0.05
0.1	0.4	0.1
0.05	0.1	0.05

where the central factor 0.4 is the weight given to the gray level in the given square, 0.05 is the weight given to the gray level in the squares up to the left or right and down to the left and right and so on. The transformed picture will then have smoother transitions which normally means a picture with a higher resolution (see Fig. 2). In terms of convolution, the transition from the first picture to the second is given by a convolution $a \mapsto a * b$, where the gray level in the square (m, n) is given by $a_{m,n}$ and b is given by the above matrix, i.e., $m_{0,0} = 0.4$, $m_{1,1} = 0.05$, $m_{1,-1} = 0.05$, $m_{-1,1} = 0.05$, $m_{-1,-1} = 0.05$ and so on (with $a_{m,n} = 0$ if $|m| > 1$ or $|n| > 1$).

It is clear in the discrete case (and at least by analogy in the continuous) that the more one concentrates the values of b to $(0, 0)$ (or 0 in the one-dimensional case) the closer will the transformed sequence be to the original. In the discrete case one may even choose b with $b_{0,0} = 1$ and the rest $b_{m,n} = 0$ in which case $a * b$ will be equal to a. In the continuous case there is no function g such that $f * g = f$ for all f but the result above gives precise condition for us to have $f * Q_n \to f$ and it is clear that these conditions say that Q_n becomes more and more concentrated around 0. If we draw different Q_n for our specific example $Q_n(x) = c_n(1 - x^2)^n$ we see (Fig. 3) this very concretely.

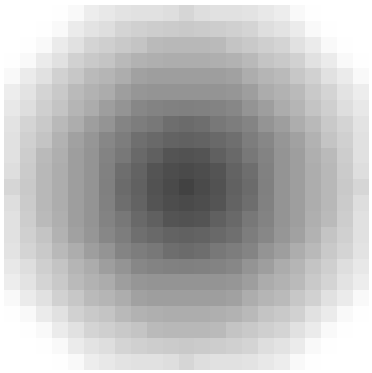

Figure 2: Convolution of low resolution picture.

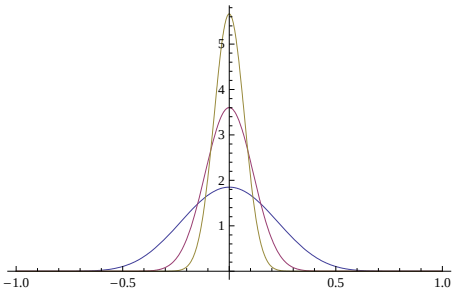

Figure 3: $Q_n(x)$ for $n = 10, 40, 100$.

Exercise 9: A *trigonometric polynomial* is a function of the form $\sum_{n=0}^{M} a_n \cos(nt) + b_n \sin(nt)$.

i) Show that products and sums of trigonemetric polynomial is again a trigonemetric polynomial.

ii) Show that each continuous function f on $[-\pi, \pi]$ with $f(-\pi) = f(\pi)$ may be uniformly approximated by trigonemetric polynomials.

Hint: Consider convolution with $Q_n(t) = d_n \cos^{2n}(t)$ for suitable d_n.

5.2 The chain rule

The proof of the chain rule for the differential of functions (Sats 9.15) is quite "polished" making it difficult to see how one might arrive at it. A step by step analysis can look as follows:

- That a function $h\colon U \to \mathbf{R}^n$, U open in $\mathbf{R}^m$, is differentiable in a point $z_0 \in U$ means that there is a function $\epsilon\colon V \to \mathbf{R}^n$ defined in a neighbourhood V of $0 \in \mathbf{R}^m$ such that $h(z_0+h) = h(z_0)+Ch+|h|\epsilon(h)$ for some linear map $C\colon \mathbf{R}^m \to \mathbf{R}^n$ and such that $\epsilon(h) \to 0$ when $h \to 0$.
- If we use the notations of the book there are therefore ϵ and η such that $f(x_0 + h) = f(x_0) + Ah + |h|\epsilon(h)$ and $g(y_0 + k) = f(y_0) + Bk + |k|\eta(k)$ with $\epsilon(h) \to 0$ when $h \to 0$ and $\eta(k) \to 0$ when $k \to 0$.
- If we put $F = g \circ f$ we have $F(x_0 + h) = g(f(x_0 + h))$. We can expand $f(x_0 + h)$ as above which gives us

$$g(f(x_0 + h)) = g(y_0 + Ah + |h|\epsilon(h)).$$

- When h is small so is $Ah + |h|\epsilon(h)$ and hence it seems reasonable to put $k = Ah + |h|\epsilon(h)$ and apply the expansion of g to it:

$$g(y_0 + Ah + |h|\epsilon(h)) = g(y_0 + k) = g(y_0) + Bk + |k|\eta(k).$$

- If we look more closely at Bk it is equal to $B(Ah + |h|\epsilon(h))$ but B is linear so this becomes equal to

$$BAh + |h|B(\epsilon(h))$$

and if we write everything out we get

$$F(x_0 + h) = F(x_0) + BAh + |h|B(\epsilon(h)) + |k|\eta(k).$$

This in turn looks like F is differentiable with differential BA provided that $|h|B(\epsilon(h)) + |k|\eta(k)$ tends towards 0 faster than h.

- We therefore compare it with $|h|$

$$|h|B(\epsilon(h)) + |k|\eta(k) = |h| \left(B(\epsilon(h)) + \frac{|k|}{|h|}\eta(k) \right)$$

and thus what is left to prove is that

$$B(\epsilon(h)) + \frac{|k|}{|h|}\eta(k) \to 0$$

when $h \to 0$.

- We have that $\epsilon(h) \to 0$ and as B is continuous we get that $B(\epsilon(h)) \to 0$ which takes care of the first term.

- We have that

$$|k| = |Ah + |h|\epsilon(h)| \leq |Ah| + |h||\epsilon(h)| \leq \|A\||h| + |h||\epsilon(h)|$$

which gives that

$$\frac{|k|}{|h|}\eta(k) \leq (\|A\| + |\epsilon(h)|)\,\eta(k).$$

The first term, $\|A\| + |\epsilon(h)|$, is smaller than $\|A\| + 1$ if h is small and $\eta(k) \to 0$ when $h \to 0$ as $k \to 0$.

6 Specific comments

6.1 Ordered sets

In Definition 1.5 of ordered sets there is (at least in some printings) a misprint: The correct formulation of (ii) is "If $x, y, z \in S$, if $x < y$ and $y < z$, then $x < z$." (The second z has incorrectly become x.)

6.2 Rational density

The second part of Thm 1.20 is easier to understand if divided up into two pieces (both of which are interesting in themselves, the second part being 1.20 (b)):

- If $x, y \in \mathbf{R}$ with $y - x > 1$, then there is an integer m with $x < m < y$.

- If $x < y$, then there is a $p \in \mathbf{Q}$ with $x < p < y$.

For the first we would like m to be the least integer such that $x < m$. The existence of such an m is stated in the book's proof without comment and is a special case (of a version) of the principle of induction. However, it should be noted that this principle is a consequence of the supremum axiom. (The proof of this fact is a little bit involved as we must first specify exactly the definition of the integers as a subset of $\mathbf{R}$ and is given in an exercise below.) In any case as m is the least integer $> x$ we must have that $m - 1 \leq x$, i.e., $m \leq x + 1 < y$ which proves the first part. For the second we proceed as in the book and use (a) to get an integer $n > 0$ such that $n(y - x) > 1$, i.e., $ny - nx > 1$. By the first part there is an integer m such that $nx < m < ny$ which gives $x < m/n < y$.

Exercise 10: i) We call a subset S of $\mathbf{R}$ *stable* if $0 \in S$ and $m \in S \Rightarrow m+1 \in S$. Show that the intersection of all stable subsets is stable and is hence is contained in all stable subsets. Denote this intersection $\mathbf{N}$.

ii) Show that if $0 \neq m \in \mathbf{N}$ then $m-1 \in \mathbf{N}$. (Hint: Show that the set $\{0\} \cup \{m \in \mathbf{R} \mid m - 1 \in \mathbf{N}\}$ is stable.)

iii) Show that if $n, m \in \mathbf{N}$ then $n < m \Rightarrow n \leq m - 1$. (Hint: Use ii) to show that for fixed n, the set of m that fulfils the implication is stable.)

iv) Show that if S is a non-empty subset of $\mathbf{N}$ then it contains a smallest element. (Hint: If x is the infimum of S then there is an $m \in S$ such that $m < x + 1$. Use iii) to show that m is a minimal element.)

v) Show that a non-empty subset of $\mathbf{N}$ bounded from above contains a largest element.

vi) Define the integers $\mathbf{Z}$ to be the union of $\mathbf{N}$ and $-\mathbf{N} := \{-m \mid m \in \mathbf{N}\}$. Show that every subset of $\mathbf{Z}$ bounded from below has a minimal element.

vii) Show that $\mathbf{Z}$ is closed under addition and multiplication. (Hint: Reduce to appropriate statements for $\mathbf{N}$ and use the same techniques as before.)

6.3 Equivalence relations and classes

In Definition 2.3 the notion of an equivalence relation is introduced. Later on (as well as before!) the associated notion of equivalence class is used without comment. Given an equivalence relation $\sim$ on a set S, the *equivalence class* containing an element $s \in S$ is the subset $\overline{s} := \{t \in S \mid s \sim t\}$. The conditions for being an equivalence relation ensure first that s does indeed belong to $\overline{s}$, then that S is the disjoint union of the different equivalence classes and finally that t and t' belong to the same equivalence class precisely when $t \sim t'$.

Exercise 11: Verify these statements.

The main point about equivalence relation is that one often wants to regard two equivalent element as essentially the same. This can be formalised by considering the set $\overline{S}$ of equivalence classes so that $s \mapsto \overline{s}$ may be regarded as a function $S \to \overline{S}$. Equivalence between elements is thus replaced with actual equality of equivalence classes.

6.4 Countable union of countable sets

It is somewhat difficult to get a proper overview of the proof of Theorem 2.12. A better way to organise is to start with a lemma.

Lemma 6.1 *If $f\colon S \to T$ is a function and S is countable, then the image $f(S)$ is finite or countable.*

Proof. Let $g\colon \mathbf{Z}_+ \to S$ be a bijection (where $\mathbf{Z}_+$ is the set of positive integers) and let h be the composite $f \circ g$. By definition h is a surjective map to $f(S)$. Put

$$T := \{n \in \mathbf{Z}_+ \mid m < n \Rightarrow h(m) \neq h(n)\}.$$

Then we have that the restriction of h to T is still surjective and it is also injective, i.e., h gives a bijection between T and $f(S)$. According to Theorem 2.8 of the book T is either finite or countable and thus so is $f(S)$. □

The proof of Theorem 2.12 is now that we (if we use the notations of the proof) first find a surjective map from $\mathbf{N}$ to S by the enumeration $x_{11}; x_{21}, x_{12}; x_{31}, x_{22} \dots$ and the lemma gives that S is finite or countable but as $E_1 \subseteq S$ it is countable.

Note also that the enumeration that is presented in the proof does not provide a proper mathematical proof. Rather the reader is expected to turn the graphical description into a mathematical proof. What must be proved is that there is a bijection between the set of positive integers $\mathbf{Z}_+$ and the set of pairs of positive integers, $\mathbf{Z}_+^2$. We can use the graphical description as a starting point but are required to turn it into explicit formulas. The key computation is that one needs to compute which position one has reached one a diagonal has been finished. This means counting the number of element in the set $\{(i,j) \in \mathbf{Z}_+^2 \mid i+j \leq n\}$. For this one computes the number of elements of each diagonal and arrives at the conclusion that the number of elements of the set equals $1+2+\cdots+n-1$ which we know to be equal to $n(n-1)/2$. With this a starting point we can give a proper proof.

Proposition 6.2 *The map $f\colon \mathbf{Z}_+^2 \to \mathbf{Z}_+$ given av $f(i,j) = (i+j-1)(i+j-2)/2 + j$ is a bijection.*

Proof. We start by proving a statement that we shall then see is equivalent to what we want to prove.

For each $m \in \mathbf{Z}_+$ there are unique $n, j \in \mathbf{Z}_+$ with $0 < j < n$ and $m = (n-1)(n-2)/2 + j$.

To show this we let n be the largest integer for which $(n-1)(n-2)/2 < m$. Such an n exists as $(n-1)(n-2)/2 < m$ implies that $n \leq m$ so we can use the supremum axiom (as in 10). We then have that $n+1$ does not fulfil the condtion so that $n(n-1)/2 = (n+1-1)(n+1-2)/2 \geq m$. This gives that $m - n(n-1)/2 \leq n(n-1)/2 - (n-1)(n-2)/2 = n-1$ so that if we put $j := m - n(n-1)/2$ we have that $0 < j < n$ and $m = (n-1)(n-2)/2 + j$. If we instead assume that we have written m as $m = (n-1)(n-2)/2 + j$ with $0 < j < n$ we may work backwards and see that n is the largest integer n so that $(n-1)(n-2)/2 < m$ which means that n is uniquely determined by m and then so is j.

We now define a function $g\colon \mathbf{Z}_+ \to \mathbf{Z}_+^2$ by given $m \in \mathbf{Z}_+$ write m as $(n-1)(n-2)/2 + j$ with $0 < j < n$ and then put $g(m) = (n-j, j)$. By what we have already proven g is well-defined and man sees easily that f and g are inverses to each other. □

Note that this makes the proof of the theorem a bit more complicated than one would like (though the graphical argument is rather attractive). There are however other ways of constructing a bijection between $\mathbf{Z}_+ \times \mathbf{Z}_+$ and $\mathbf{Z}_+$ or, which amounts to the same, between $\mathbf{N} \times \mathbf{N}$ and $\mathbf{N}$. The following exercise gives one of them.

Exercise 12: i) Show that each natural number n can be written uniquely in the form $n = \sum_{k=0}^{\infty} a_k 10^k$, where $0 \leq a_k < 10$ and all but a finite number of the a_k are equal to 0.

ii) Show that the map

$$(\sum_{k=0}^{\infty} a_k 10^k, \sum_{k=0}^{\infty} b_k 10^k) \mapsto \sum_{k=0}^{\infty} c_k 10^k,$$

where $c_k = a_{k/2}$ if k is even and $c_k = b_{(k-1)/2}$ if k is odd, gives a bijection $\mathbf{N} \times \mathbf{N} \to \mathbf{N}$.

6.5 Countability of the rational number

The proof of Cor. 2.13 is quite brief and there is reason to look more closely at it. We start by showing the countability of the positive rational numbers $\mathbf{Q}_+$. We have a map $\mathbf{Z}_+^2 \to \mathbf{Q}_+$ taking a pair (m,n) to the number m/n. This map is onto and as we have a bijection $\mathbf{Z}_+ \to \mathbf{Z}_+^2$ we get that the composite $\mathbf{Z}_+ \to \mathbf{Z}_+^2 \to \mathbf{Q}_+$ also is onto (as the composite of two surjective maps is

surjective). According to Lemma 6.1 we get that $\mathbf{Q}_+$ is countable or finite but as $\mathbf{Q}_+$ contains $\mathbf{Z}_+$ which is not finite $\mathbf{Q}_+$ itself can not be finite.

Exercise 13: We have an explicit surjective map $\mathbf{Z}_+ \to \mathbf{Q}_+$ and Lemma 6.1 then gives an explicit enumeration of $\mathbf{Q}_+$. Determine the image of 20 under it.

6.6 Convergence of Cauchy sequences

Theorem 3.11(b) can be proven by a somewhat different argument than that of the book which maybe is clearer: One uses Theorem 3.6(a) to conclude that there is a $p \in X$ and a subsequence (p_{n_k}) such that $p_{n_k} \to p$. It is then enough to show that the whole sequence converges towards p. Let therefore $\epsilon > 0$ be arbitrary and pick N so that we have $d(p_{n_k}, p) < \epsilon/2$ (we do this by first choosing a K so that it is true for $k \geq K$ and then choose a N so that $n_k \geq N \Rightarrow k \geq K$) if $n_k \geq N$ as well as $d(p_m, p_n) < \epsilon/2$ if $m, n \geq N$ (which can be done as the subsequence converges and the whole sequence is a Cauchy sequence). Choose a $n_k \geq N$ and let$m \geq N$. The triangle inequality then gives us $d(p_m, p) \leq d(p_m, p_{n_k}) + d(p_{n_k}, p) < \epsilon/2 + \epsilon/2 = \epsilon$.

6.7 Connected sets

The proof of Theorem 4.22 becomes more difficult to see through than necessary because the definition of connected subset that is used. The proof can be simplified (at the price of doing some more initial work). What one should start doing is to define the notion of connected metric space (the book defines when a subset of a metric space is connected).

Definition 6.3 *A metric space X is* connected *if it* cannot *be written as a disjoint union of two open non-empty sets.*

Remark: i) This is a negative definition (in that it rather defines when a space is not connected). The positive version (and the one that is used in practice) is if one has written a connected metric space as a disjoint union of two open subsets, then one of the subsets is empty.

ii) Note that if the metric space X is the disjoint union of the open subsets U and V, then U is the complement of V which means that U (as well as V) is also closed. Conversely, if U is both open and closed, then X is the union of U and its complement V which is also open. A metric space is therefore connected precisely when it does not contain any proper non-empty *clopen* (= *cl*osed + *open*) subsets.

iii) It is not necessary to use the definition used in the book of connected subset, one can instead just use the one just given (though we shall show that the definitions are equivalent).

The result analogous to Theorem 4.22 using instead this definition has a much cleaner proof.

Theorem 6.4 *If $f\colon X \to Y$ is a continuous function between metric spaces, then it is image $f(X)$ is connected (considered as a metric space through the metric defined on Y) if X is.*

Proof. The function f gives rise to a function (which we also call f) $f\colon X \to f(X)$. It is easy to show that f as function from X to Y is continuous precisely when it is continuous as function from X to $f(X)$ (where again $f(X)$ is a metric through the metric induced from Y). We may therefore assume that $Y = f(X)$. Assume now that Y is the disjoint union of the non-empty open subsets U and V. Then X is the disjoint union of $f^{-1}(U)$ and $f^{-1}(V)$. These sets are non-empty as $f(X) = Y$ and open as f is continuous and U and V are open. As X is assumed to be connected this is a contradiction. □

Exercise 14: Show that if X and Y are metric spaces and $Z \subseteq Y$, then a function $f\colon X \to Z$ is continuous (where Z is given the induced metric) exactly when the function $X \to Y$ given by f is continuous.

To get a simple proof of Theorem 4.22 it remains to show how our definition of connected metric spaces is connected to the definition of connected subset in the book.

Proposition 6.5 *Let X be a metric space and $E \subseteq X$ a subset. Then E is connected as metric space (with the metric induced from X) precisely when it is connected as subset of X (i.e., according to Definition 2.45).*

Proof. Assume that E is the union of two separated sets A and B and let $U' := X \setminus \overline{A}$ and $V' := X \setminus \overline{B}$ and put $U := U' \cap E$ and $V := V' \cap E$. It is clear that U' and V' are open in X and therefore U and V are open in E. We have that U contains B as $B \cap \overline{A} = \emptyset$ and U is evidently disjoint from A so that $U = B$ as E is the disjoint union of A and B. In the same way we get $V = A$ so that E is the disjoint union of the two open non-empty subsets U and V.

Assume now that E is the disjoint union of the two open non-empty subsets U and V. Thus for each $x \in U$ there is a $r_x > 0$ so that $B_E(x, r_x) := \{ y \in E \mid d(x, y) < r_x \}$ lies in U and thus we have that $B_E(x, r_x) \cap V = \emptyset$. This means that $B_X(x, r_x) \cap V = \emptyset$ eftersom $B_X(x, r_x) \cap E = B_E(x, r_x)$. If we put $U' := \cup_{x \in U} B_X(x, r_x)$ we have that U is open in X and hence $\overline{V} \subseteq X \setminus X'$. We further obviously have that $U \subseteq U'$ so that $\overline{V} \cap U = \emptyset$ and in the same way we get that $V \cap \overline{U} = \emptyset$ and thus U and V are separated. □

To further demonstrate that our definition of connected spaces can be used without involving the definition in the book we prove the characterisation of connected subsets of $\mathbf{R}$.

Proposition 6.6 *A subset E of the extended real line is connected precisely when $z \in E$ if $x, y \in E$ and $x < z < y$.*

Proof. Assume to begin with that there are x, y, z with $x < z < y$ and $x, y \in E$ but $z \notin E$. If we put $U := \{ t \in E \mid t < z \}$ and $V := \{ t \in E \mid z < t \}$. Then we have that U and V are open in E, E is the disjoint union of U and V as $z \notin E$ and U and V are non-empty because $x \in U$ and $y \in V$. Hence E is not connected.

For the converse we may assume, by way of contradiction, that E is the disjoint union of the open non-empty sets U and V and that if $x, y \in E$ we have that all z with $x < z < y$ belongs to E. We may further choose $x \in U$ and $y \in V$. After possibly having permuted U and V we may assume that $x < y$ and our assumption gives that $[x, y] \subseteq E$. We may then replace E by $[x, y]$, U by $U \cap [x, y]$ and V by $V \cap [x, y]$ which means that we may assume that $E = [x, y]$. Let now z be the supremum of U. We have that z either belongs to U or V. If it belongs to U then $z \neq y$ as $y \in V$ and as U is open in $[x, y]$ we have that there is a $\delta > 0$ such that $]z - \delta, z + \delta[\subseteq U$ but the fact that $z + \delta/2 \in U$ means that z is not an upper bound of U, a contradiction. If instead $z \in V$ by the same argument we get that $]z - \delta, z + \delta[\in V$ for some $\delta > 0$ which implies that $z - \delta/2$ is also an upper bound for U contradicting that z is the least upper bound of U. □

6.8 Countability of points of discontinuity

There is an alternative proof of Theorem 4.30 (that the set of points of discontinuity of a monotone function is at most countable) which may be easier to understand. We assume that $f \colon [a, b] \to \mathbf{R}$ is monotone and let E be the set of its points of discontinuity, By possibly replacing f by $-f$ we may assume that f is increasing. Put now, for a positive integer n, $E_n := \{ x \in E \mid f(x+) - f(x-) > 1/n \}$. If we can show that E_n is finite and that $E = \cup_n E_n$, then it follows from Corollary 2.12 that E is at most countable. We have $x \in E$ precisely when $f(x+) - f(x-) > 0$. For a $x \in E$ we can then choose n such that $f(x+) - f(x-) > 1/n$ which means that $x \in E_n$ which implies that E is the union of the E_n. On the other hand, let $\{x_1 < \cdots < x_k\}$ be a finite subset of E_n. Then we have that

$$\begin{aligned} f(b) - f(a) = f(b) - f(x_k+) + f(x_k+) - f(x_k-) + f(x_k-) - f(x_{k-1}+) + \\ f(x_{k-1}) + \cdots - f(x_1-) + f(x_1-) - f(a) = \\ (f(b) - f(x_k+)) + (f(x_k+) - f(x_k-)) + \\ (f(x_k-) - f(x_{k-1}+)) + (f(x_{k-1} + \cdots - f(x_1-)) + (f(x_1-) - f(a)) \,. \end{aligned}$$

All tems in the last sum is ≥ 0 as f is increasing and $f(x_i+) - f(x_i-) > 1/n$ for $i = 1, \ldots, k$. This means that the whole sum is $> k/n$ which gives that $k < n(f(b) - f(a))$. This means that every finite subset of E_n contains less than $< n(f(b) - f(a))$ element which in turn implies that E_n itself contains at most that number of elements and is therefore finite.

6.9 Def 4.33

In Definition 4.33 one must replace "such that $V \cap E$ is not empty" with "such that $V \cap E$ contains points different from x".

Exercise 15: Give an example that Definition 4.33 as it stands in the book is not equivalent with Definition 4.1 (in relevant cases).

6.10 Theorem 6.10

At the beginning of the proof the claim is made that one may choose $[u_j, v_j]$ around the points of E such that $\sum_j \alpha(v_j) - \alpha(u_j) < \epsilon$. The reason for this is that if p_j is the point in $E \cap [u_j, v_j]$ we may, by possibly shrinking the interval $[u_j, v_j]$, make $\alpha(v_j) - \alpha(p_j)$ and $\alpha(p_j) - \alpha(u_j)$ arbitrarily small (as α is continuous in p_j) and therefore $\alpha(v_j) - \alpha(u_j) = \alpha(v_j) - \alpha(p_j) + \alpha(p_j) - \alpha(u_j)$ can be made as small as needed. Now, the number of elements in E is fixed so $\sum_j \alpha(v_j) - \alpha(u_j)$ can be made arbitrarily small.

Note that the proof has many similarities with the proof of Weierstrass approximation theorem. In both case one divides the relevant interval in two pieces and treat them in completely different ways. In one case one uses just that the arbitrary function f is bounded and that another factor ($\alpha(v_j) - \alpha(u_j)$ resp. Q_n) is small, in the other case one uses that f is continuous and therefore has small variation on small intervals.

6.11 The norm of a linear map

Probably the best way to think of the norm of a linear map A is that it is the smallest number λ such that $|Ax| \leq \lambda|x|$ for all vectors x. Hence, if one want to give an upper bound of $\|A\|$ it is then a matter of estimating $|Ax|$ as a fixed number times $|x|$. We have for instance that $|(A+B)x| = |Ax + Bx| \leq |Ax| + |Bx| \leq \|A\||x| + \|B\||x| = (\|A\| + \|B\|)|x|$ which implies that $\|A + B\| \leq \|A\| + \|B\|$ (which is one of the results of Theorem 9.7).

There is however another way to look at the norm: If $x \neq 0$ then $|Ax| \leq \lambda|x|$ is equivalent with $A(x/|x|) \leq \lambda$ and the length of $x/|x|$ equals 1 so that

$$\|A\| = \sup_{|x|=1} |Ax|.$$

Now, we can easily see that A as a functino $\mathbf{R}^m \to \mathbf{R}^n$ is continuous; if $x = \sum_i \lambda_i e_i$ (where e_i are the elements of the standard base) we have $Ax = \sum_i \lambda_i Ae_i$ so that the continuity follows directly from the continuity of addition and multiplication. From this follows that the function $x \mapsto |Ax|$ is continuous (as $y \mapsto |y|$ is continuous). Further, the unit sphere $\{x \in \mathbf{R}^m \mid |x| = 1\}$ is closed and bounded and hence compact which implies that the continuous function $x \mapsto |Ax|$ is bounded and consequently $\sup_{|x|=1} |Ax| < \infty$.

We can be even more precise and first note that

$$\|A\|^2 = \sup_{|x|=1} |Ax|^2$$

and that $x \mapsto |Ax|^2$ is a quadratic form. More precisely we have that

$$|Ax|^2 = (Ax)^t(Ax) = x^t A^t A x,$$

where we have identified the linear map with its matrix (with respect to the standard base for $\mathbf{R}^n$). This means that the quadratic form is given by the symmetric matrix A^tA. We now know from the theory of quadratic forms that $\sup_{|x|=1} |Ax|^2$ is equal to the largest eigenvalue of the symmetric matrix A^tA so that $\|A\|$ is equal to the square root of this largest eigenvalue.

Example: Assume that $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Then we have that $A(x, y) = y$ and the norm $\|A\|$ is the smallest number λ such that $|y| \leq \lambda\sqrt{x^2 + y^2}$. If we put $(x, y) = (0, 1)$ we get $1 \leq \lambda$. On the other hand we have $|y| \leq \sqrt{x^2 + y^2}$ and thus $\lambda = 1$ work. This gives $\|A\| = 1$.

We may instead compute $A^tA = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ which clearly has the eigenvalues 0 and 1 and we get that $\|A\| = \sqrt{1} = 1$.

7 The inverse function theorem

It is not altogether easy to follow the proof in the book of the inverse function theorem. I shall give a proof that contains the same basic ideas but which is hopefully easier to follow. It takes its starting point in the method of Newton-Raphson.

Before we go into the details let us start by noting some common features between the two methods.

In both cases we want to solve an equation $f(x) = y$, where x is the value to be found. On the other hand y is a variable value close to some b where we as starting data have an a such that $f(a) = b$ and as a further condition we are looking for a solution x which is close to a. In both cases we try to find an *iteration scheme* $x_{n+1} = G_y(x_n)$, where we use the notation G_y to emphasise that G_y will depend on y. The idea is that the sequence (x_n) should converge to a solution x. As always with iteration schemes there are two things to do; show that the sequence converges and given that it converges that it converges to a solution of the original problem. Given that the sequence converges some x we may pass to the limit in the relation $x_{n+1} = G_y(x_n)$ we get, assuming that G_y is continuous, $x = G_y(x)$ and a first condition on G_y is that this relation should imply that $f(x) = y$ and if this is so we have taken care of the second part. For the convergence the general idea is that G_y should have the property that x_{n+1} and x_n should become closer and closer to each other quickly enough so that (x_n) will actually be a Cauchy sequence. (Note that for this it will not quite be enough that $x_{n+1} - x_n \to 0$ when $n \to \infty$ as we need $x_i - x_j \to 0$ when $i, j \to \infty$.) A crucial part to make this work for an iteration scheme is that one must have a good initial value x_0, in fact it should be close to an actual solution. This is in general a difficult problem. In our case y is supposed to be close to b and we are hoping that x will be close a. This suggests that we should pick a as the initial value and at least it gives us a candidate for initial value. That the whole method really works depends on the fact that we have the freedom to choose how close to b we require y to be. We choose this distance to be small enough to get a number of estimates that together will make all the parts work.

7.1 Newton-Raphson in one variable

The method of Newton-Raphson tries to solve an equation $g(x) = 0$, where we begin by assuming that g is a function from an open subset of $\mathbf{R}$ to $\mathbf{R}$, will start with an approximate value x_0 and then step by step try to find better and better approximations. In an attempt to find a better approximation of a zero of g one replaces g by its tangent in the point and hope that a zero of the tangent gives a better approximation. Then one repeats this procedure (see Fig. 4). If x_n is the result of the n'th step its tangent has the equation $y = g(x_n) + g'(x_n)(x - x_n)$ and if we solve for a zero we get

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}.$$

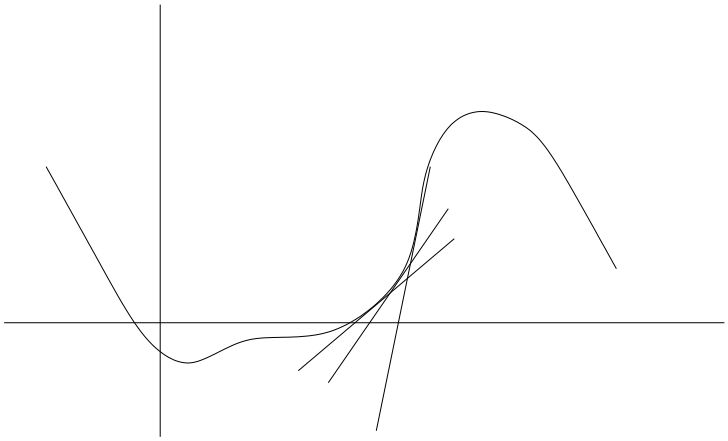

Figure 4: The Newton-Raphson method

If we assume that this sequence converges towards an x we may take the limit in this definition and obtain $x = x - g(x)/g'(x)$ which gives $g(x) = 0$ and we have indeed found a zero. This conclusion was however made under a number of assumption and the next step is to find conditions under which they are fulfilled.

Example: Let $g(x) = x^2 - 2$. Then we get $x_{n+1} = x_n - (x_n^2 - 2)/2x_n$. This can be rewritten as $x_{n+1} = (x_n + 2/x_n)/2$ which in turn can be expressed as saying that that we let x_{n+1} be the average of x_n and the number whose product with with x_n becomes 2. If we start with $x_0 = 1$ we get $x_1 = 1.5$, $x_2 \approx 1.4167$, $x_3 \approx 1.414216$, $x_4 \approx 1.414213562375$, $x_5 \approx 1.4142135623709504880169$ which as can be seen converges very quickly towards $\sqrt{2} \approx 1.4142135623730950488016887242$.

To begin with we must of course assume that the derivative of $g(x)$ exists but in order for the final limit argument to work we must also assume that the derivative is continuous. This is also what is needed for some estimates that we shall need to work. More precisely we have the following lemma which says that the relative error in the approximation of a function by its tangent is uniformly bounded.

Lemma 7.1 *Let $g\colon]a, b[\to \mathbf{R}$ be a function with continuous derivative. Then we have that for every closed interval I n $]a, b[$ and every $\epsilon > 0$ there is a $\delta > 0$ such that*

$$|g(x) - g(y) - g'(y)(x - y)| < \epsilon|x - y|$$

for all $x, y \in I$ with $|x - y| < \delta$.

Proof. We may, because of uniform continuity choose a $\delta > 0$ such that $|g'(x) - g'(y)| < \epsilon$ if $x, y \in I$ and $|x - y| < \delta$. We can use the mean value theorem to find a ξ in $]x, y[$ such that $g(x) - g(y) = g'(\xi)(x - y)$. This gives

$$|g(x) - g(y) - g'(y)(x - y)| = |g'(\xi) - g'(y)||x - y| < \epsilon|x - y|,$$

the last equality coming from $|\xi - y| < \delta$. □

If we now consider our recursion, then we see that we have chosen x_n exactly so that $g(x_{n-1}) + g'(x_{n-1})(x_n - x_{n-1}) = 0$, i.e., so that

$$g(x_n) = g(x_n) - g(x_{n-1}) - g'(x_{n-1})(x_n - x_{n-1}).$$

According to the lemmam we may make this small in relation to $|x_n - x_{n-1}|$ and then according to the recursion formula $|x_{n+1} - x_n|$ will be small in relation to $|x_n - x_{n-1}|$ which means that

x_n and x_{n+1} will be closer to each other and that ought to mean that $\{x_i\}$ forms a Cauchy sequence. It is now quite easy to write down the conditions that makes this true. In fact, we want two conditions should be fulfilled. The first is that $|x_{n+1} - x_n|$ should be estimated from above by a fixed factor less than 1 times $|x_n - x_{n-1}|$. This will mean that not only will the differences between x_n and x_{n+1} tend towards 0 but also the differences between x_i and x_j will which will make (x_n) a Cauchy sequence. The second is that we need an estimate of $|x_1 - x_0|$ that will guarantee that the x_n will keep close enough to x_0 so that the estimate that is needed in the first will be true. We assume that $g'(x_0) \neq 0$ and because of the continuity we can find a closed interval I such that x_0 lies in the interior of I and such that $|g'(x)| \geq d > 0$ for $x \in I$. We then use the lemma to find a $\delta > 0$ such that if $|x - y| < \delta$ then we have that $|g(x) - g(y) - g'(y)(x - y)| < 1/2d|x - y|$. We further want that if $|x - x_0| < \delta$ then x will lie in I. We now define x_n by the inductive formula

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$$

and want to show that $|x_{n+1} - x_n| < 1/2|x_n - x_{n-1}|$ which by induction gives that $|x_{n+1} - x_n| < 2^{-n}|x_1 - x_0|$ and a repeated use of the triangle inequality gives

$$|x_{n+1} - x_0| < 1 + \frac{1}{2} + \cdots + \frac{1}{2^n} = 2\left(1 - \frac{1}{2^{n+1}}\right)|x_1 - x_0| < 2|x_1 - x_0|.$$

If we to begin with assume that $|x_1 - x_0| < \delta/2$, then the induction gives $|x_n - x_0| < \delta$ and also that $|x_n - x_{n-1}| < \delta$ so that $x_n \in I$ and

$$|g(x_n)| = |g(x_n) - g(x_{n-1}) - g'(x_{n-1})(x_n - x_{n-1})| < \frac{1}{2d}|x_n - x_{n-1}|$$

which gives that

$$|x_{n+1} - x_n| = \left|\frac{g(x_n)}{g'(x_n)}\right| < \frac{d}{2d}|x_n - x_{n-1}| = \frac{1}{2}|x_n - x_{n-1}|$$

which provides the next step of the induction. The whole induction now works provided that we know that $|x_1 - x_0| < \delta/2$. We have that $|x_1 - x_0| = |g(x_0)/g'(x_0)|$ so this is true if we can assume that x_0 is a good enough approximation so tht $|g(x_0)| < d\delta/2$, which is exactly what we do.

If now $i > j$ we may once again use that we know that $|x_{n+1} - x_n| < 2^{-n}|x_1 - x_0|$ to get that

$$|x_i - x_j| < \left(\frac{1}{2^j} + \frac{1}{2^{j+1}} + \cdots + \frac{1}{2^{i-1}}\right)|x_1 - x_0| < 2\frac{1}{2^j}|x_1 - x_0|$$

which shows that $\{x_i\}$ is a Cauchy sequence. This leads to the following result:

Proposition 7.2 *Let $g\colon I \to \mathbf{R}$ be a function from an open interval I with continuous derivative in the the interval. Let $x_0 \in I$ and assume given $\delta > 0$ and $d > 0$ such that*

1. *$]x_0 - \delta, x_0 + \delta[\subseteq I$,*
2. *$|g'(y)| \geq d$ for all $y \in I$,*
3. *$|g(x) - g(y) - g'(y)(x - y)| < d/2|x - y|$ if $x, y \in I$ and $|x - y| < \delta$,*
4. *and $|g(x_0)| \leq d\delta/2$.*

Then there is an $x \in I$ with $|x - x_0| \leq \delta$ and $g(x) = 0$.

Proof. The proof has already been given, let us just point out some things. The first is that the condition $|g(x_0)| \leq d\delta/2$ was needed to make $|x_1 - x_0| < \delta/2$ which in turn was needed to make sure that $|x_n - x_0| < \delta$ which in its turn was used to allow us to use the other inequalities to estimate $|x_{n+1} - x_n|$. By then passing to the limit in $|x_n - x_0| < \delta$ we get $|x - x_0| \leq \delta$. This extra piece of information is useful in various contexts. □

One could of course also wonder how one finds the starting value x_0. The answer that will interest us for the inverse function theorem is that $g(x)$ is of the form $f(x) - y'$; we are trying to solve the equation $f(x) = y'$ where y' is close to $y_0 := g(x_0)$. Allowing only values y' such that $|y' - y_0| \leq d\delta/2$ we may, as $g(x_0) = y_0 - y'$, obtain a solution x' with $f(x') = y'$. (Note that $g(x) - g(y)$ and $g'(y)$ do not depend on y' so that d and δ depend only on f and not on y'.)

Remark: For our purposes (i.e., the inverse function theorem) it is enough to show that there exists a neighbourhood of y_0 in which a solution always exists. Note however that one may then find a neighbourhood of y' in which one can solve the equation. In many cases this process can be repeated to find solution for y which do not necessarily lie close to y_0. (It does not work always however, one may for instance be unlucky and encounter a zero of $f'(x)$ along the way.)

We shal finish this part with some comments that are not directly relevant for the inverse function theorem. That $g(x)$ is of the form $f(x) - y'$ with y' close to $f(x_0)$ is not the only way to get an approximation to a zero of g (that can then be used as an initial value). There is in fact another method to find a sequence that converges to a zero of g, interval division. Assume therefore that we have found x_0 and t_0 such that $g(x_0) < 0$ and $g(t_0) > 0$. Consider $g((x_0+t_0)/2)$. If this value equals 0 we are finished, if it is > 0 we put $x_1 := x_0$ and $t_1 := (x_0 + t_0)/2$ and if it is < 0 we put $x_1 := (x_0 + t_0)/2$ and $t_1 := t_0$. Then we have that $g(x_1) < 0$ and $g(t_1) > 0$ and $|x_1 - t_1| = 1/2|x_0 - t_0|$. Repeating this procedure gives us (x_n, t_n) with $g(x_n) < 0$and $g(t_n) > 0$ so that $|x_n - y_n| = 2^{-n}|x_0 - t_0|$ which means that x_n and t_n converges to a common value which necessarily is a zero of g.

One may then ask what the point is of the method of Newton-Raphson when this method already exists, in particular as we see that $|x_n - x|$ is on the order of 2^{-n} which means that for both methods we must make on the order of n iterations to get n digits of the zero. The answer is that in practice the estimate we have made of the error in the Newton-Raphson method ismuch too large. If we do not only assume that g has a continuous derivative but that it also has a second derivative which is bounded then we get a much better estimate of the error; roughly the number of correct digits is double through each iteration. (This can be seen to be the case in the example above of computing $\sqrt{2}$.)

Exercise 16: i) Assume that $g\colon I \to \mathbf{R}$ is a function on the open interval I with first and second derivatives in each point and that we have $|f''(x)| \leq M$ for all $x \in I$. Show that

$$|g(x) - g(y) - g'(y)(x - y)| \leq \frac{M}{2}|x - y|^2$$

and

$$|g'(x) - g'(y)| \leq M|x - y|$$

for all $x, y \in I$.

ii) Show that if x_0 is used as a starting point for a Newton-Raphson iteration and if $x_n \in I$ for all n then we have that there is a constant C such that $|x_{n+1} - x_n| \leq (C|x_1 - x_0|)^{2^n}$. Give conditions ensuring that $C|x_1 - x_0| < 1$.

Remark: A convergence like this where $|x_n - x| = O(\epsilon^{2^n})$ for some $\epsilon < 1$ is called *quadratic convergence* and roughly means that then number of correct digits is doubled at each iteration.

7.2 The Newton-Raphson method in several variables

To be able to generalise the Newton-Raphson method to several variables we are forced to forget the interpretation in terms tangents. This is actually quite natural as we saw in the proof (of the one-variable version) that the method works that what we used was that the definition of the derivative gave that $g(x) - g(y) - g'(y)(x - y)$ should be small (though we needed some uniformity of the smallness which goes beyond the mere definition). We can formulate what we did as saying that instead of trying to solve the equation $g(x) = 0$ we try (in the first step) to solve the equation we get by approximating $g(x)$ with the first two terms of its Taylor expansion, i.e., $g(x_0) + g'(x_0)(x - x_0) = 0$ instead of $g(x) = 0$. To do that there is nothing that stops us from assuming that g is a function from an open subset of $\mathbf{R}^n$ to $\mathbf{R}^n$. The only difference is that instead of a linear equation we get a system of linear equations. We can still solve it though if we assume that $g'(x_0)$ is an invertible linear map (to just assume that it is non-zero is not enough); we get $x = x_0 - g'(x_0)^{-1}(g(x_0))$. We now go through the same steps that we did in one variable to see that it works.

Lemma 7.3 *Let $g\colon U \to \mathbf{R}^n$ be a function from an open subset of $\mathbf{R}^n$ to $\mathbf{R}^n$ with continuous derivative. Then we have that for every compact subset K in U and every $\epsilon > 0$ there is a $\delta > 0$ such that*

$$|g(x) - g(y) - g'(y)(x - y)| < \epsilon|x - y|$$

for all $x, y \in K$ with $|x - y| < \delta$.

Proof. Because of uniform continuity we can choose a $\delta > 0$ such that $\|g'(x) - g'(y)\| < \epsilon$ if $x, y \in K$ and $|x - y| < \delta$. Fix $y \in K$ and consider the function $h(x) := g(x) - g'(y)(x)$. Then we have that $h'(x) = g'(x) - g'(y)$ and thus $\|h'(z)\| < \epsilon$ if $|z - y| < \delta$. According to Theorem 9.19 we have, for $|x - y| < \delta$

$$|g(x) - g(y) - g'(y)(x - y)| = |h(x) - h(y)| \leq \epsilon|x - y|.$$

□

Exercise 17: Use the proof of the lemma to give a new proof of Theorem 9.21.

We can use this lemma to give a criterion for the success of the method of Newton-Raphson in several variable which is almost identical with the one-variable case.

Proposition 7.4 *Let $g\colon U \to \mathbf{R}^n$ be a function from an open subset $U \subseteq \mathbf{R}^n$ with continuous derivative. Let $x_0 \in U$ and assume given $\delta > 0$ and $d > 0$ such that*

1. *The ball $D(x_0, \delta)$ lies in U,*
2. *$g'(y)$ is invertible and $\|g'(y)^{-1}\| \leq 1/d$ for all $y \in U$,*
3. *$|g(x) - g(y) - g'(y)(x - y)| \leq d/2|x - y|$ if $x, y \in U$ and $|x - y| < \delta$,*
4. *and $|g(x_0)| < d\delta/2$.*

Then there exists a $x \in U$ with $|x - x_0| < \delta$ and $g(x) = 0$.

Proof. We do as inte the 1-variable case and put

$$x_{n+1} = x_n - g'(x_n)^{-1}(g(x_n)),$$

where it of course must be part of our proof to show that $x_n \in U$ for all U (if this is the case, our conditions imply that $g'(x_n)$ is invertible for all n). We note first that $|x_1 - x_0| \leq \|g'(x_0)^{-1}\||g(x_0)| < 1/d \cdot d\delta/2 = \delta/2$. We want to show by induction that $|x_{n+1} - x_n| < 1/2|x_n - x_{n-1}|$. If this is true for a given n we have $|x_{n+1} - x_n| < 2^{-n}|x_1 - x_0|$ and therefore that

$$|x_n - x_0| = |x_n - x_{n-1} + x_{n-1} - \ldots x_1 - x_0| \leq |x_n - x_{n-1}| + \ldots |x_1 - x_0| < (2 - 2^{-n+1})|x_1 - x_0| < \delta$$

so that x_n lies in U. We have now chosen x_n such that $g(x_{n-1}) + g'(x_{n-1})(x_n - x_{n-1})$ which gives

$$|g(x_n)| = |g(x_n) - g(x_{n-1}) - g'(x_{n-1})(x_n - x_{n-1})| < d/2|x_n - x_{n-1}|$$

and this in turn gives

$$|x_{n+1} - x_n| = |g'(x_n)(g(x_n))| \leq \|g'(x_n)\||g(x_n)| < \frac{1}{d}\frac{d}{2}|x_n - x_{n-1}| = \frac{1}{2}|x_n - x_{n-1}|.$$

The same argument that gave that $|x_n - x_0| < (2 - 2^{-n+1})|x_1 - x_0|$ now gives for $i > j$

$$|x_i - x_j| < 2\left(1 - \frac{1}{2^{i-j}}\right)\frac{1}{2^j}|x_1 - x_0|,$$

which shows that $\{x_i\}$ is a Cauchy sequence. Let x be its limit. As we have that $|x_n - x_0| < (2 - 2^{-n+1})|x_1 - x_0|$ we get $|x - x_0| \leq 2|x_1 - x_0| < \delta$ by letting n tend towards infinity and in particular we have $x \in U$. If we let $n \to \infty$ in the equation $x_{n+1} = x_n - g'(x_n)^{-1}(g(x_n))$ and use that g and g' are continuous we get $x = x - g'(x)(g(x))$ which gives $g'(x)(g(x)) = 0$, which in turn, as $g'(x)$ is invertible, gives $g(x) = 0$. This proves the proposition. □

Remark: Note that this is essentially the same proof as that of Theorem 9.23 of the book.

7.3 The inverse function theorem

The inverse function theorem is now a rather direct consequence of the Newton-Raphson method.

Theorem 7.5 *Assume that $f\colon W \to \mathbf{R}^n$ is a function from an open subset W of $\mathbf{R}^n$ to $\mathbf{R}^n$ with continuous derivative. Assume that $f'(a)$ is invertible for some $a \in W$ and put $b := f(a)$. Then we have that*

1. *there are open subsets U and V i $\mathbf{R}^n$ sådana att $a \in U$, $b \in V$ such that f is a bijection from U to V and*
2. *if $g\colon V \to U$ is the inverse of f (that exists by the first part) we have that g has a continuous derivative in all of V.*

Proof. As $f'(x)$ is continuous and the invertible matrices form an open subset of $L(\mathbf{R}^n)$ the x for which $f'(x)$ is invertible form an open subset. Thus we may, possibly after having replaced U with a smaller set, assume that $g'(x)$ is invertible. We also have that $x \mapsto \|f'(x)^{-1}\|$ is a continuous function so we can, again after possibly having shrunk U, assume that $\|f'(x)^{-1}\| \leq 1/d$ for some fix d. We may further replace U with $D(a, \alpha)$ for some $\alpha > 0$ such that the closed ball $\overline{D}(a, \alpha)$ lies in U. This means that we may apply Lemma 7.3 to the compact set $\overline{D}(a, \alpha)$ and thus find $\delta > 0$ such that

$$|f(x) - f(y) - f'(y)(x - y)| < \frac{d}{2}|x - y|$$

if $|x-y| < \delta$. We now want to apply the Newton-Raphson method to the function $g(x) := f(x)-y$ for different y. Note that $g'(x) = f'(x)$ and $g(x) - g(x') = f(x) - f(x')$ so that the only thing that needs proving in order to use Proposition 7.4 on $g(x)$ with starting value x_0 is that $g(x_0) = f(x_0) - y = y_0 - y$ fulfils $|y_0 - y| < d\delta'/2$ where $\delta' \leq \delta$ has been chosen so that $D(x_0, \delta') \subseteq U$ and $y_0 := f(x_0)$. This means that if we choose y sufficiently close to y_0 this is always the case. The proof now proceeds as in Theorem 9.24. We may furthermore conclude that f is injective in U, because if $f(x) = f(y)$, then we have that

$$|f'(y)(x-y)| = |f(x)-f(y)-f'(y)(x-y)| < \frac{d}{2}|x-y| = \frac{d}{2}|f'(y)^{-1}f'(y)(x-y)| \leq \frac{1}{2}|f'(y)(x-y)|,$$

which gives $f'(y)(x - y) = 0$ which in turn gives that $x - y = 0$ when $f'(y)$ is invertible. Finally, to show that the inverse h of f has continuous differential it is enough to show that

$h'(y) = f'(h(y))^{-1}$, aas h' then is the composite of continuous functions. Now we have for $y, y' \in V$ that

$$h(y') - h(y) - f'(h(y))^{-1}(y' - y) = f'(h(y))^{-1}\left(f'(h(y))(h(y') - h(y) - (f(h(y')) - f(h(y)))\right)$$

which implies

$$|h(y') - h(y) - f'(h(y))^{-1}(y' - y)| \leq \|f'(h(y))^{-1}\||f(x') - f(x) - f(x)(x' - x)|,$$

where $x = h(y)$ and $x' = h(y')$. It follows from Proposition 7.4 and $g(x') - g(x) = y' - y$ that there is a constant C such that $|x' - x| \leq C|y' - y|$ which gives

$$\frac{|h(y') - h(y) - f'(h(y))^{-1}(y' - y)|}{|y' - y|} \leq \frac{\|f'(h(y))^{-1}\|}{C}\frac{|f(x') - f(x) - f(x)(x' - x)|}{|x' - x|}$$

and as $|x' - x| \leq C|y' - y|$ this gives that $x' \to x$ when $y' \to y$ and we get that the right hand side tends towards 0 when $y' \to y$ (as f is differentiable), this shows that $h'(y) = f'(h(y))^{-1}$. □

If we compare with the proof in the book we see that the difference is that in the proof of the book we use the iteration

$$x_{n+1} = x_n + f'(x_0)^{-1}(y - f(x_n))$$

while the iteration in the Newton-Raphson method is

$$x_{n+1} = x_n + f'(x_n)^{-1}(y - f(x_n)).$$

If we then look at the proof that these two methods converges one sees that the proof are fairly similar and in particular that the estimates of the speed of convergence are about the same. More precisely we get *linear convergence*, i.e., $x_n - x = \mathcal{O}(\epsilon^n)$ (which roughly means that each iteration gives one new correct digit). This may make one wonder what the point of the method of Newton-Raphson is, in particular is it seems more complicated; in each step one is forced to compute the inverse of $f'(x_n)$ while in the method of the book it is enough to compute $f'(x_0)^{-1}$ once and for all. The reason why the Newton-Raphson method is interesting is the same as in the one-variable case: If one puts extra conditions on f one will get quadratic convergence which is shown in the following exercises.

Exercise 18: Assume that for $f\colon [a, b] \to \mathbf{R}^n$ we have that $f^{(n-1)}$ is continuous on all of $[a, b]$ and that $f^{(n)}(t)$ exists for all $t \in]a, b[$. Show that there exists $x \in]a, b[$ such that

$$\left| f(b) - \sum_{k=0}^{n-1} \frac{(b-a)^k}{k!} f^k(a) \right| \leq \frac{b-a}{n!}|f^{(n)}(x)|.$$

Hint: Imitate the proof of Theorem 5.19 combined with Taylor's theorem.

References