The Anti-Cookbook of Proofs

A course in mathematical intuition

by Vadim Zavalishin

rev. 1.0.0-alpha

The book presents various math topics by "rediscovering" them, focusing on intuitional and motivational aspects of mathematical thinking, and is supposed to complement technical university-level math courses. The text includes materials on calculus of one and several variables, linear, tensor and exterior algebra, ordinary linear differential equations as well as Fourier transform. For students and graduates of technical engineering universities.

Disclaimer: this book is not suitable for usage as a conventional math textbook or a reference math text. It is a course in mathematical intuition, not a course in math. Please check the Introduction chapter for further details.

The right is hereby granted by the book's author to freely copy this revision of the book in software or hard-copy form, as long as the book is copied in its full entirety (including this copyright note) and its contents are not modified.

Built on December 13, 2023.

Contents (overview)

Introduction 5

Why this book was written and what is its intended goal and target audience. The approach and presentation style of this book. How to read this book.

I Mostly algebra 17

1 Some experiments 19

Using experiment and observation to obtain mathematical results. Quadratic equation solution formula and fundamental theorems of algebra and arithmetic.

2 Real numbers 33

Introduction of real numbers as the values of infinite decimal fractions. Relationship to the postulate of Cauchy sequence convergence and to other equivalent postulates. Relationship to continuity of arithmetic operations. Uncountability of real numbers: do we need that many numbers? Continuity of the real line as justification of uncountability.

3 Linearity everywhere 63

Defining linear spaces based on addition only (without scalar multiplication). Defining linearity based on addition only. Application of linearity to the analysis of the dot product and the rotation transformations. Application of linear algebra concepts for the analysis of nonlinear functions.

4 Beyond the basis 93

Basis-independent (covariant-form) equations. Linear basis-independent entities (tensors) as generalizations of the ideas of vectors, linear operators, bilinear forms etc. Tensor index notation and its relationship to the matrix notation. Basic operations on tensors. Review of some of the linear algebra topics from the tensor algebra perspective. Dual space. Tensor-like entities.

5 Area and volume 135

Derivation of signed area of a parallelogram by bilinearity. Exterior product in 2D and in higher dimension counts. Relationship to dot product. Relationship to determinants. Relationship to tensors and pseudotensors.

II Mostly calculus 195

6 Instantaneous and additive quantities 197

Motivation-based construction of differentiation and integration framework basics. Derivative. Riemann and Riemann–Stieltjes integrals. Indefinite integral (antiderivative). Improper integrals. Differentials. Multiple dimensions. Mean value theorem. Common particular cases of generalized Stokes' theorem.

7 Exponential function 297

Analytic solution of differential equations as the limiting case of a finite-step scheme, and the exponential function as the limiting case of integer powers emerging in that scheme. Matrix diagonalization as a linear basis change for a differential equation system. Rotational behavior of exponentials of imaginary numbers as a limiting case of complex multiplication. Jordan normal form as a limiting case fallback of diagonalization. Discovering the spectral theorems. Linear space rotations as exponentials.

8 Superpositions of sinusoids 379

Four types of Fourier transform: discrete Fourier series and the other three types as its limiting cases. Normalization placement. Transform domains and their dual symmetry. Transform properties. Kronecker and Dirac delta functions. Convergence. Dual domain effects of discontinuities and features. Energy and its momenta. Reciprocal duality of transform domains. Sampling and STFT as a means to finite computability. Application to random processes.

Revision history 535

Detailed table of contents 537

Index 543

Introduction

Why this book was written and what is its intended goal and target audience. The approach and presentation style of this book. How to read this book.

For whom is this book 5
Why this book 6
The approach of this book 7
Time and place for rigor 8
Proofs and cookbooks 9
Separation of concerns vs. holistic approach 9
Motivation and intuition 10
How to read this book 11
Acknowledgements 15

For whom is this book

The intended target audience of this book is the students and graduates of technical engineering universities, who are looking to deepen and broaden the intuitional part of their understanding of math. The book is supposed to *complement* a typical university math course for engineers. A loose set of prerequisites expects familiarity with

- basic calculus of the real line: real numbers, limits of sequences and functions, infinite series, differentiation and integration
- introductory level of calculus in more than 1 dimension: partial derivatives, double and triple integrals
- linear algebra basics: vectors, matrices, linear spaces
- complex number algebra and, ideally, an introductory level of complex calculus
- a good number of rigorous definitions and proofs in the contexts of calculus and linear algebra (for the sake of discussion focus, the book intentionally refrains from a fully rigorous presentation, leaving quite a number of blanks that a reader is supposed not really or necessarily to be able to fill, but at least to recognize and to be aware of).

In many cases it might even be a good idea to allow a period of a few years between taking a university math course and reading this book, to let the formal math knowledge obtained from the university course settle a bit. In particular, some students may get overwhelmingly fascinated by the power and elegance of the formal aspects of math, and it might take a bit of time before this original fascination somewhat subsides and (re-)opens the room (and a new perspective) for the appreciation of intuitional and motivational aspects.

Conversely, if a subset of the earlier mentioned prerequisites is not met, it is probably not a showstopper, as the missing bits of knowledge can be picked up "on-demand" from textbooks or reference sources during the reading of the book. However it might be a bit too early to read this book at the very beginning of a university math program. The book assumes a certain level of established familiarity with the aforementioned concepts, and aims to further structure and extend the reader's knowledge of the matters, hopefully answering a number of questions that an inquisitive student might have developed during their math course. Most importantly, the book tries to answer questions like *"why would one bother with this or that idea in the first place, and what was (or what could have been) the path of the development of that idea?"*

On the other hand, this book might be not suitable for readers with knowledge levels of a professional mathematician or close to that, as well as for students of the respective specializations. For them, the book might be lacking new information, as well as lacking proper formal presentation of the materials, and the goals of the book would be probably covered for such readers elsewhere and in other ways. This book is written by an engineer (with some mathematical background), for engineers and with an engineering mindset in mind, although it's still possible that some mathematicians might find the book an interesting read nevertheless. If nothing else, the book still presents the author's personal point of view on the subject, which hopefully at least some of the readers are going to find quite valuable.

Why this book

The goal of this book could be described as *to help building an intuitional and motivational structure around the mathematical knowledge of the reader* in a number of key areas of math, including the areas already studied by the reader, as well as possibly a number of new areas. Now what do we mean by that?

- By *intuitional aspects* of understanding we mean the viewpoints and approaches allowing one to manipulate mathematical entities in one's mind, rather than by formal application of rules on paper.
- By *motivational aspects* of understanding we mean being aware of why certain concepts or statements are there, what purpose they are serving, or which problem they are solving, and how do they solve it.

These two sets of aspects are extremely important if one wishes to be able to use math efficiently.

- Good intuitional understanding of math allows one to more quickly and easily find ways of combining the math concepts as well as identifying the

missing pieces, in order to arrive at the desired goal. It can be seen as an efficient navigation helper, improving one's awareness of where they are and where they are going, thereby helping one to avoid "getting lost amongst the roads of math".

- Understanding the motivational part of the math allows one to "notice less obvious turns" and even "navigate beyond the official roads" at times, modifying and tweaking the building blocks of mathematical knowledge beyond their standard form as necessary. Effectively, it gives one a larger freedom of application of math.

Both these sides are often having a limited, if any, coverage in standard math courses. Any math course has certain focus, consistency and size restrictions. Respectively it needs to choose priorities and approaches, which in turn determine what fits into the course and what doesn't. Often a typical math course will prioritize such things as

- rigorous and formal presentation,
- conciseness,
- abstraction.

These things are extremely important and have their place and value, but often may come at the cost of leaving intuitional and motivational aspects aside. While it's probably a good idea to begin a university-level math study with and focus on formal and abstract topics and approaches, thereby providing an overall solid foundation, it's probably equally not such a good idea to leave the intuition and motivation completely, or to a large part, aside. This book suggests to handle the latter two in a complementary set of materials provided by the book's text.

By taking a different focus and approach in this book, it simultaneously becomes possible and desirable to not only review parts of standard math course contents from a somewhat different angle, but also to go beyond such contents. In particular, the topics of tensor and exterior algebra will be often omitted from a technical university math course, but besides their own value, they also provide critical "missing links" for intuitive understanding of basic linear algebra course contents. Such areas of the book's text mostly do not require preexisting familiarity with the respective concepts, but it might be a good idea to catch up, to some extent, with proper mathematical treatment of these topics after reading this book.

The approach of this book

The book's contents can be loosely described as "excerpts from an imaginary history of math". While not trying to be historically correct at all, the book takes the reader on a journey, telling a story of not necessarily how math *was* developed, but how it *could have been* developed. Or it can be seen as "a diary of trying to recreate a number of key mathematical concepts from scratch" (which is what the book de-facto is).

Bringing intuitional and motivational aspects into the foreground, the book chooses a different focus and a different set of priorities. The book thereby proposes a point of view which might be seen as strongly deviating from standard math texts. It needs to be noted, however, that this presented point of view is not being promoted by the book as *the* point of view, but rather as *a* point of view, which complements whatever other points of view exist. In general, it's the author's belief, that it's often quite valuable to have a number of different points of view. Points of view can be seen as tools, and it's often beneficial to have more than one tool in one's toolbox.

The book also doesn't aim at giving definite answers or presenting fully finalized results. As we would be trying to finalize a particular point of view, the fewer and fewer of the intuitional and motivational aspects would usually remain there, and we would just be left with technicalities. In this book we are trying to concentrate on other things, without getting lost in those technical details. Thus many passages in the book will be intentionally left vague (to various degrees), in order to avoid the detours of filling up all the missing details, otherwise we could be risking losing the bigger picture, which is our main focus here.

Time and place for rigor

The biggest step away from a classical math presentation, that the book takes, is dispensing away with rigor, which it does in order to free up space and mental resources to concentrate on other things. For some people this might immediately constitute a rude violation of one of the key principles of a proper math study, however let's be not so quick with judgment, but try to analyze the value and the purpose of rigor in math, especially in a math course.

One could say that rigor is used in math solely *as a tool* to avoid mistakes. From that perspective, rigor is merely a verification tool rather than an inseparable integral part of math. As we want the ultimate reliability from math, this tool is absolutely indispensable as such, but still has its application time and place. We would like to suggest that this time and place occur:

- in the *final form* of a proof, where it is our explicit goal to check for mistakes in the proof, or
- in the *final form* of a definition or a theorem, where rigor is used as a means to clearly specify the statement's applicability boundaries, excluding possible misunderstandings.

Here, the words "final form" explicitly exempt intermediate forms of definitions, theorems and proofs, occurring during their development.

Outside of the contexts mentioned above, the involvement of rigor can in many cases be premature and constitute and unnecessary hurdle and an inhibitor to the progress and freedom of our thought. So it might be more efficient not to apply rigor 100% of the time, but only during occasional verification steps, or when wrapping up the results. At other times the rigor can be, if not fully ignored, then significantly lowered in priority. So hopefully we can agree to temporarily demote the rigor from its high position and make it optional in many situations. With that in mind, let's now try to assess the place and value of a particularly large and rigor-sensitive part of a typical math course, this part being the theorem proofs.

Proofs and cookbooks

It could be suggested that the main value of presenting theorem proofs in a math course is not in the proofs themselves, but in getting the student well familiar with how such proofs need to look, as well as maybe developing some proof-construction skills. From that perspective, as more and more proofs get studied, and the student gets more and more familiar with the proof language and conventions, the value of each new studied proof decreases, and at some point the time might be more efficiently spent elsewhere.

One also could question how efficient it is to let the student learn to construct proofs by solely looking at the *final results* of such construction. These final results are usually stripped down of the largest part of the thinking process that led to their development, their primary purpose being verification and verifiability. This issue becomes particularly prominent with "proofs by *Ansatz*", where one pretends to somehow have guessed the correct answer. Again, "guessing the right answer" is completely fine for final verification purposes, but it can be questioned, whether in reality the answer was "just guessed", or there actually was a different, possibly longer, path to arrive at this "guess" in the first place.

So, showing one a number of proofs and expecting one to learn to construct proofs just from that is a little bit like giving one a cookbook and saying "now go and make fundamentally new recipes of your own". The problem is that the final result is not equal to the process of its obtaining, and one might need to go through a lot of guessing and "reverse engineering" before one might even get a chance to be able to come up with a comparably decent result. Now this guessing and "reverse engineering" might be a valuable exercise in itself, but should it be the only way how one learns to do things? Maybe not.

Therefore, as proofs go, the book concentrates on the process of obtaining those proofs, rather than on presenting the final form of the proofs. Dropping the rigor requirement serves here as a way to stay focused on the intuition and "navigation". It is however expected that the readers recognize and are aware of the "holes", which are thereby being left in the "proofs" obtained that way, and (roughly) understand what kind of extra steps would be needed to fill those holes.

The same applies to definitions and statements found in this book: they intentionally lack some rigor, the reader being expected to notice the respective "holes" in their formulations.

Separation of concerns vs. holistic approach

Another thing with standard math courses which may be somewhat of a problem is separation of math into disconnected areas, the most obvious separation probably occurring between calculus and linear algebra. The problem is that these areas are not disconnected in reality, their separation rather serves the purpose of structuring a math course into a number of self-contained parts. Except that assuming that these parts are really self-contained is somewhat wishful thinking.

This separation might be a "necessary and lesser evil" from the perspective of the course structuring, but, once the separation has fulfilled its purpose, we don't need to stick to it any longer. In particular, in this complementary text, we can take advantage of not separating the topics too strongly. Of course,

we still need to maintain some structure, but this structure is way less strict. In particular there is quite strong mixing of calculus and linear algebra topics. This is also why the two parts of the book are called *Mostly algebra* and *Mostly calculus*, rather than *Algebra* and *Calculus*.

One example of benefiting from the lowered amount of separation is provided by the topics of matrix diagonalization and Jordan normal form, which we, among other things, review in this text. Those might look totally abstract from the point of view of a pure linear algebra course, but get an obvious and clear motivation in the context of differential equation solving, additionally becoming much more intuitive if put into the context of tensor algebra. Also the "proof" of the Jordan normal form can be done rather intuitively in terms of calculus and exterior algebra, while a pure "linear algebraic" proof might require quite a bit of "bookkeeping". However a technical university math course is likely to separate calculus from linear algebra, also treating differential equations at a later point, while tensor and exterior algebras might often not be a part of the curriculum whatsoever. This could impose rather severe restrictions on the presentation and the proof approaches.

Similarly, various Fourier theory topics have can have a very intuitive understanding from the perspective of linear algebra. However the separation between calculus and linear algebra (as well as possibly the occasional lack of rigor, if we try to stick to finite-dimensional linear algebra) again may prevent presentation from this perspective.

In this book, having lowered the amounts of separation and rigor, we are having much more freedom for an intuitive treatment of such topics.

Motivation and intuition

The mentioned earlier "excerpts from an imaginary history of math" approach of this text serves a number of purposes.

- The "excerpts" part means that we concentrate on a selection of topics we believe benefit most from the approach taken by this book.
- The "imaginary" part means that our main purpose has little to do with historical correctness. Historically, the math development could have taken a number of detours or even gone astray at times. The order in which the topics were studied historically is probably at times far from optimal. By writing an "imaginary history" we could try to fix all those issues, pretending that we "were just lucky" to immediately do things in a right direction and order.
- The "history" part means that we present or review the topics as if they would be discovered and studied from scratch, which naturally allows to concentrate on the motivation for these topics as well as on the intuition behind the thinking.

The "proofs" contained in this text, lacking the full rigor, should be probably better referred to as "proof drafts" or "proof sketches" (the same applies to statements and definitions). In order to be convincing as examples of a person's thinking about the matters, these proof sketches were developed from scratch, almost or even fully refraining from consulting any sources (which could have

given a number of hints, thus providing an additional kind of "an unfair advantage" and thereby making them "less plausible" as real thinking examples).[1]

These newly derived proof sketches may therefore deviate in the approach and the ideas from the proofs commonly found in texts. For our purposes here, however, this is rather a benefit. Commonly found proofs, even when looking highly concise and elegant, might often be the results of a long iterative process, which is then dropped in the finally presented form of the proof. Or they might be very difficult to come up with. Either case will make them rather unsuitable for the purposes of this book. We would rather have here a proof which doesn't require many iterations to develop and which can be (maybe with some realistic amounts of skill and luck) found within days at most.

Errors and mistakes

There is a seeming downside of deriving proof sketches from scratch, which comes from the fact that they haven't been proofread and verified by thousands of mathematicians, the latter usually being the case with the proofs found in common texts. In principle this applies to the entire contents of the book. While some extra checking of the book contents has been done, there can still easily be some mistakes. For that reason the readers are *strongly advised against using this book as a standard math textbook or as any kind of reference source.* At the same time, the mistakes of this kind naturally occur in the course of development of math and sometimes could take years to be noticed and fixed. From that perspective, such mistakes (if any are contained in the text) also naturally fit into the book's purpose of demonstrating how math could have been developed and do not affect the intended value of the book. The readers looking for exactly formulated and properly verified mathematical facts should consider other sources.

How to read this book

First and foremost, the reader has to keep in mind, that, as we explained before, the amount of rigor in this text has been noticeably dropped compared to the typical math standard, and the reader is expected to identify the "missing holes" themselves. Nothing needs to be immediately done about those holes though, the focus of the book being the motivational and intuitional aspects of mathematical thinking. The completion of the proof sketches, if really desired as an exercise, might be taken into a separate context closer to a proper math course.[2]

According to the purpose of practicing mathematical intuition, the reader is also suggested not to simply rush through the text, but to regularly make a stop and ask themselves, whether they can already figure out where this is going. It might be a good exercise to try taking a number of further steps without reading on through the book's text and then compare the obtained results to

[1] To be completely honest, this also helped the author to further structure and refine his own knowledge and understanding, but that's beyond the point we're trying to make here.

[2] According to the made earlier reservation of possibly contained mistakes, it might be that some of these proof sketches are found to be "leading nowhere" and cannot be completed. This doesn't devalue the practical use of trying to complete them, as the reader would also practice identifying such mistakes.

what follows in the book (where the reader's results do not necessarily need to match what's written in the book, as long as these results generally make sense). As an aid and a reminder for the reader to follow that process, the book contains a number of explicitly declared goals throughout the text, which look like following:

Goal 1.3. This is some example goal which is set in the book's text as a reminder for the reader to stop and think. *Hint:* some hint text will be often given. It is however suggested that the readers first give it a shot without reading the hint.

These goals are thought to be stop points, where the reader is suggested to try to take a number of further steps on their own. The goals vary in complexity, and figuring out the next steps could take any amount of time between a fraction of a second and up to a few days in some cases. If the reader is still stuck after a few days, it's probably advisable to simply read on. Depending on the reader's background, preferences and motivation, even taking a few days may be a bit too much.

Once each specific goal is reached, there will be a respective mark in the text looking like

Goal 1.3 completed. ■

Sometimes a goal might stretch over a few subsections or even sections, but never exceeds a chapter. After reading through the text until a specific goal is reached it might be worth taking another stop and trying to memorize the main ideas used in reaching the goal, and after that trying to recreate the same steps again based on the memorized ideas.[3]

More complicated goals will be further split into subgoals. E.g.

Goal 1.4a. This is an example subgoal. Readers who didn't manage to reach a particular specified goal completely on their own, even after reading a hint, may then further stop at each of the subgoals, taking these subgoals as further hints.

More impatient readers can read through the goals without stopping whatsoever, merely taking the goals as some kind of structural separators in the text.

Highlighted excerpt

Besides the goals, important definitions and statements in the text will be highlighted like this paragraph. We'd like to remind the readers once again that the definitions and statements in this text are intentionally having a loose form, lacking a strict formalism.

▷ The book is also extensively employing "digressions", which look like the present paragraph. Depending on the context, these can be used as

- a kind of inline footnote, or
- additional highlighting means, or
- just a means to separate certain passages from the rest of the text.

[3] It is highly unrecommended to memorize all the details of how the goal was reached. For one, this goes against the idea of practicing the intuition. Also quite a number of details can be not fully or properly worked out in this text, thus the text is not a good source for memorizing.

▷ The triangular mark at the beginning of each "digression section" is intended to help distinguishing between the situation where a page break occurs between two paragraphs of one and the same digression and the situation where a new digression is started on the next page.

Conventions commonly used throughout the book include:

- Writing an equation number on top of an equality sign, e.g.

$$a \overset{(1.14)}{=} b$$

  means that the transformation from one side of the sign to the other is done by that equation. In the above example we are saying that $a = b$ by Equation 1.14.

- A question mark on top of an equality sign:

$$a \overset{?}{=} b$$

  means that at this point in the text we are not sure yet, whether the equality holds or not. Later we might find that it does hold, or that it doesn't.

- The word "iff" occasionally encountered throughout this book's text is a common shorthand convention for "if and only if", which might be unfamiliar to some of the people who studied math in a language other than English.

Chapter order

The chapters of the book are relatively independent and therefore, to a good extent, can be read in a random order. However, occasionally, some knowledge contained in earlier chapters may be required and the reader may need to keep an eye on that. As the focus is taken away from formal presentations to intuitional insights, the names of the book chapters are usually more related to the starting motivational point than reflect the formal name of the respective area of math. E.g. the chapter on the exterior algebra is called *Area and volume*. The contents of the chapters are also structured based on the motivational path rather than on the formal areas of mathematics, e.g. we discuss matrix diagonalization in the chapter on exponentials.

Motivational starters

Somewhere close to the beginnings of most (but not all) of the chapters, there will be a *Motivational starter* for the chapter. This motivational starter formulates a specific (seemingly small) goal that we want to achieve. As we proceed through the chapter text, the readers are encouraged to observe, how this seemingly small goal unfolds into further and further details, all those details effectively originating from that one starting question.

Intentional vagueness and iterations

It is important to keep in mind, that the book text presents examples of how mathematical ideas can develop essentially from zero into whole frameworks. Therefore the phrasing is in many cases intentionally vague and often imprecise, as the initial thoughts often would be while the idea unfolds. Often things won't be presented in a "final form", but rather refined at later steps as we pay attention to further details. Sometimes "further details" will be completely left out of scope.

Formulas as part of the text

Not unlike the proof sketches, the formulas and expression transformations in this book shouldn't be looked at simply as calculation recipes or formal transformations where the sole value is just the final result. Rather, they often tend to complement the normal text with extra meaning encoded in the expressions. E.g. the derivation of (3.31) in Chapter 3 is intended to illustrate how bilinearity "transfers" the basis vector components from the original vectors into a combined matrix of dot products of basis vectors, and each step in the respective transformation illustrates a part of that process. From that perspective, in quite a number of cases the equations should be taken for their illustrative purposes as if they were just normal text, or maybe as if they were pictures, and the readers are expected to pay attention to such equations. Whether a particular equation has this extra informational role, needs to be judged from the context of the discussion around it.

Prerequisite awareness

Since the set of prerequisites for this book is rather loose, there are no strict assumptions about the reader's knowledge at the beginning of discussion of each topic. Some parts of the text are written with review-like purposes in mind, concentrating on trying to offer a different point of view on the already known materials, and therefore expecting some prior familiarity of the reader with the topic. Some other parts are written with a primary intention of introducing the reader to new areas, which may fall outside of the standard curriculum, thus no prior knowledge is expected (but it might be a good idea to catch up afterwards at least using some reference texts for more rigorous definitions and statements). Sometimes it could be a mixture of both. There are occasional hints in the text, in regards to whether prior familiarity with a particular subject is expected, but the readers are also advised to exercise their own judgment in this regard.

The reading speed

It is common to try to read a book at a more or less constant pace, maintaining the "momentum". Respectively, having to slow down may feel rather uncomfortable, creating a subconscious urge to continue moving forward at the previous speed. It is highly recommended to be aware of this effect. The complexity of the book's text varies randomly, simply due to the nature of the discussed materials. This is similar to walking a road, where some parts of the road are going uphill and are more difficult to travel than the horizontal parts. This complexity may often be hidden and misleading: a long and complicated

formula may contain fully trivial transformations, while a very simple one may require a good amount of thinking before the contained idea really sinks in. It is therefore totally natural and even advised to stop at a complicated place and not try to rush through by all means. At certain places it might be okay to take a few days pause to meditate on a certain idea. At times it might also be helpful to read on a bit, finding out a small number of further details and afterwards returning back to the skipped point. Should some particular text fragment turn out to be a real blocker, it's probably okay to simply move on, maybe returning to this fragment much later.

Acknowledgements

The author would like to thank Norbert Nemec for a productive discussion, which to a significant extent affected the shape and the focus of the book's contents. The author is also grateful to Helene Kolpakova for her feedback. Very special and warm thanks to Romanov I.V.

There are virtually no sources which have been used during writing of this book, with the exception of occasional checkups with materials in Wikipedia. A number of details in the discussion of Fourier series was inspired by *Foundations of Mathematical Analysis* by Richard Johnsonbaugh and W. E. Pfaffenberger.

Part I

Mostly algebra

Chapter 1

Some experiments

Using experiment and observation to obtain mathematical results. Quadratic equation solution formula and fundamental theorems of algebra and arithmetic.

1.1 Quadratic equation 19
1.2 Fundamental theorem of algebra 23
1.2.1 Real case 24
1.2.2 Complex case 25
1.3 Fundamental theorem of arithmetic 28
1.3.1 Modified Euclid's lemma 28
1.3.2 LCM and GCD 30
Summary of Chapter 1 31

At a first look one might think that experiments belong to physics and have no place in mathematics, since the latter is based on formal proofs rather than on experimental data. However this idea would not be really correct. While it is true that the *final results* in mathematics must be based on formal proofs, there is nothing which stops us from using experiments to derive various hints and ideas leading towards these final results.[1]

In this chapter we shall give a number of demonstrations of obtaining mathematical results by an approach based on experiment and observation. We are going to address the derivation of the solution formula for quadratic equation as well as derivations of the fundamental theorems of algebra and arithmetic.

1.1 Quadratic equation

The derivation of the commonly known solution formula for quadratic equations of the form $x^2+px+q=0$ is usually done by employing the idea of "completing the square". Specifically, we rewrite the equation in the form

$$x^2+px+q=x^2+2\frac{p}{2}x+\left(\frac{p}{2}\right)^2-\left(\frac{p}{2}\right)^2+q=\left(x+\frac{p}{2}\right)^2-\left(\left(\frac{p}{2}\right)^2-q\right)$$

[1] There is also such thing as a *proof by a counterexample*, which *is* an "experimental kind of proof", but this is not the point which we're currently trying to make.

This very elegant idea is however not completely obvious to come up with. Let's see how it could have been found in a relatively straightforward way.

Goal 1.1. Find a path leading to the idea of completing the square as a way of solving quadratic equation. *Hint:* study the behavior of the graphs of quadratic polynomials.

Suppose we want to solve the equations of the general form

$$ax^2 + bx + c = 0 \tag{1.1}$$

and suppose we didn't know how to do that. As the first step, we could try to make things a bit simpler by dividing the entire equation by a, which gives us an equation of the form

$$x^2 + px + q = 0 \tag{1.2}$$

known as *reduced quadratic equation*, where $p = b/a$, $q = c/a$. However now it feels we're stuck. Just trying a few random transformations such as

$$x(x + p) = -q$$

doesn't seem to be getting us anywhere.

In order to obtain some ideas how to move further we may choose to experiment with the equation of the form (1.2). In particular, we may decide to play with graphical plots of various functions of the form $y = x^2 + px + q$ (with various specific values of p and q) and look for intersections of such plots with the x axis, corresponding to the solutions of the equation (1.2).

In such case we are going to notice that the graph of $y = x^2 + px + q$ is a parabola of "one and the same size", where the values of p and q are connected to the horizontal and the vertical positioning of that graph (Fig. 1.1). Specifically, q is connected only to the vertical positioning, while p is connected to both. We can further notice that $p = 0$ corresponds to the graph being symmetric with respect to the y axis (Fig. 1.2). In this case (1.2) takes the form $x^2 + q = 0$ and is very easy to solve: $x = \pm\sqrt{q}$.

This may lead us to the following idea: let's shift the graph horizontally so that it becomes symmetric with respect to the y axis and respectively p becomes equal to zero. Horizontal shifts are achieved by the variable replacement $x' = x - x_0$. Substituting $x = x' + x_0$ for x into (1.2) we obtain

$$\begin{aligned} x^2 + px + q &= (x' + x_0)^2 + p(x' + x_0) + q = x'^2 + 2x'x_0 + x_0^2 + px' + px_0 + q \\ &= x'^2 + (p + 2x_0)x' + (q + px_0 + x_0^2) = x'^2 + p'x' + q' = 0 \end{aligned}$$

We therefore notice that $p' = p + 2x_0 = 0$ is attained by $x_0 = -p/2$:

$$\begin{aligned} x^2 + px + q &= x'^2 + \left(p - 2\frac{p}{2}\right)x' + \left(q - p\frac{p}{2} + \left(-\frac{p}{2}\right)^2\right) \\ &= x'^2 + \left(q - \left(\frac{p}{2}\right)^2\right) = x'^2 - d = 0 \end{aligned} \tag{1.3}$$

where

$$d = \left(\frac{p}{2}\right)^2 - q$$

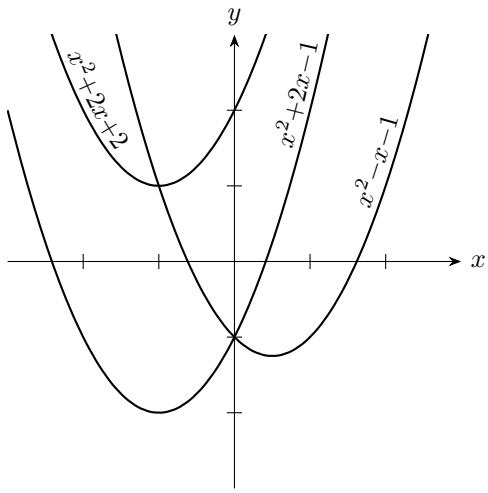


Figure 1.1: Various parabolic curves.

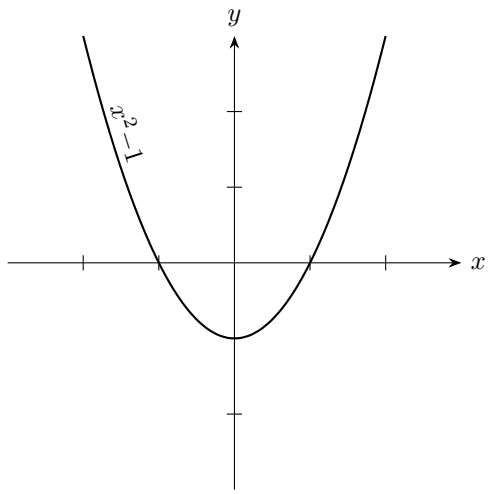


Figure 1.2: Symmetric parabolic curve.

and it follows that $x' = \pm\sqrt{d}$ and respectively

$$x = x' + x_0 = x_0 + x' = -\frac{p}{2} \pm \sqrt{d} = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q} \tag{1.4}$$

We can also reexpress the solution (1.4) in terms of a, b and c, where we recall that $p = b/a$, $q = c/a$:

(quadratic equation solution formula)

$$\begin{aligned} x &= -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q} = -\frac{b}{2a} \pm \sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}} \\ &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-b \pm \sqrt{D}}{2a} \end{aligned} \tag{1.5}$$

where $D = b^2 - 4ac$.

Now we can "rediscover" the approach of completing the square. Aiming at streamlining our steps of the derivation of the solution formula, we want to get rid of the auxiliary variable x' and therefore substitute $x' = x - x_0$ directly into the result of (1.3), which yields

$$\begin{aligned} 0 &= x'^2 - d = (x - x_0)^2 - d = \left(x + \frac{p}{2}\right)^2 - d \\ &= x^2 + px + \left(\frac{p}{2}\right)^2 + \left(q - \left(\frac{p}{2}\right)^2\right) = x^2 + px + q \end{aligned} \tag{1.6}$$

Going through (1.6) from right to left:

$$\begin{aligned} x^2 + px + q &= x^2 + px + \left(\frac{p}{2}\right)^2 + \left(q - \left(\frac{p}{2}\right)^2\right) \\ &= \left(x + \frac{p}{2}\right)^2 + \left(q - \left(\frac{p}{2}\right)^2\right) \end{aligned}$$

we obtain the procedure of completing the square.

Goal 1.1 completed. ■

At this point it might be worth it making a short detour into numerical methods to illustrate that mathematical formulas are not always directly applicable in machine computations. Specifically, not only the formula (1.5) doesn't work for $a = 0$ (and obviously neither does (1.4)), it also has computation precision issues at small a, where we have $\sqrt{D} \approx |b|$ and respectively $-b + \sqrt{D} \approx 0$ (if $b > 0$) or $-b - \sqrt{D} \approx 0$ (if $b < 0$).

We can however notice that it's only one of the two expressions $-b \pm \sqrt{D}$ that is having the precision issues, while the other one doesn't. Yes, it might deliver a large value of x (and even the one out of range of machine representation), but it doesn't have the precision issues per se. We can then recall that, by Vieta's formulas, $x_1 x_2 = c/a$ and we can use the solution which doesn't have precision issues to obtain the other solution:

(quadratic equation solution formula, alternative version)

$$x_{1,2} = \frac{c/a}{x_{2,1}} = \frac{(c/a)\cdot 2a}{(-b \mp \sqrt{D})} = \frac{2c}{-b \mp \sqrt{D}} = \frac{2c}{-b \mp \sqrt{b^2 - 4ac}} \tag{1.7}$$

▷ Notice how the out-of-range intermediate values concern is addressed in (1.7). E.g. imagine that $a \approx 0$, $b > 0$ and $-b + \sqrt{D} \approx 0$, thereby $x_1 = \frac{-b-\sqrt{D}}{2a}$ is very large while $x_2 = \frac{-b+\sqrt{D}}{2a}$ has precision issues. Suppose x_1 is so large that it doesn't fit into machine representation. Therefore it doesn't matter which formula we use to compute x_1, the result cannot be represented anyway. However, the value x_2, which we obtained from x_1 as $x_1 = c/ax_2$, still can be computed directly from (1.7), without going out of range, since we no longer explicitly use x_2 as an intermediate value in these computations.

Now we can use (1.5) to compute one of the roots of the quadratic equation and use (1.7) to compute the other root. Remember, that to minimize precision losses we want the expression $-b \pm \sqrt{D}$ to essentially add two values of the same sign, so that $|-b \pm \sqrt{D}| = |b| + |\sqrt{D}|$. Thus we obtain:

(a more numerically robust solution formula for quadratic equations)

For $b > 0$ we should choose the expression $-b - \sqrt{D}$ and respectively compute

$$x_1 = \frac{-b - \sqrt{D}}{2a} \qquad x_2 = \frac{2c}{-b - \sqrt{D}}$$

while for $b < 0$ we should use $-b + \sqrt{D}$ to express the solutions:

$$x_1 = \frac{2c}{-b + \sqrt{D}} \qquad x_2 = \frac{-b + \sqrt{D}}{2a}$$

For $b = 0$ either approach can work, provided $a \neq 0$. The case $a \approx 0$ should be accounted for separately, as at some point there will be only one solution instead of two within the machine-representable number range.

1.2 Fundamental theorem of algebra

The *fundamental theorem of algebra* states that an order N polynomial equation has N roots (counting their multiplicities). In this section we shall derive a visual approach to proving the theorem. By far, we won't get to all the details necessary for a rigorous proof, however the presented argument is hopefully going to be convincing enough anyway and can be used as a basis for a proper proof.

An N-th order polynomial equation has the form

$$P(x) = \sum_{n=0}^{N} a_n x^n = 0$$

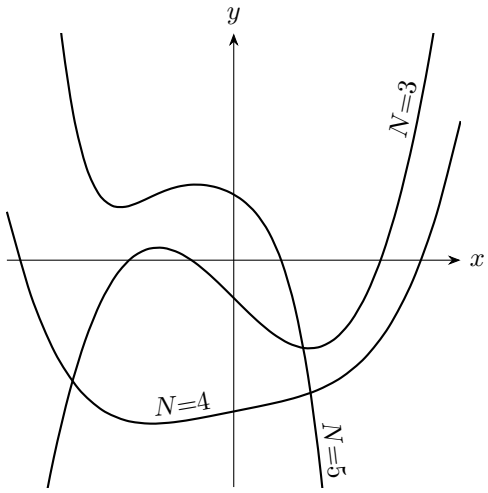


Figure 1.3: Polynomial curves of different orders N.

where $a_N \neq 0$ (otherwise the order of the equation would be lower than N). Without loss of generality we can assume that $a_N = 1$, since otherwise we can simply divide the entire equation by $a_N \neq 0$, obtaining an equation with $a_N = 1$:

$$P(x) = x^N + \sum_{n=0}^{N-1} a_n x^n = 0 \tag{1.8}$$

1.2.1 Real case

Let's start with the purely real case, that is the coefficients of the polynomial are real and we are interested only in the real roots.

Goal 1.2. Come up with a sketch of a proof for the fundamental theorem of algebra, where we limit ourselves just to real polynomials and real roots. *Hint:* study the behavior of the graphs of real polynomials.

We begin our research by considering a few example polynomial graphs (Fig. 1.3). Looking at these graphs we can easily come to the idea that all odd-order polynomials have at least one root. Indeed, for an odd-order $P(x)$ of the form (1.8), at $x \to +\infty$ we have $P(x) \to +\infty$, while at $x \to -\infty$ we have $P(x) \to -\infty$. Apparently this occurs due to x^N dominating the other terms of the polynomial at large $|x|$. This means that for a sufficiently large positive x_+ we are guaranteed to have $P(x_+) > 0$ and for a sufficiently large (in absolute magnitude) negative x_- we are guaranteed to have $P(x_-) < 0$. By continuity of $P(x)$, the graph has to cross the horizontal axis at least once somewhere within (x_-, x_+) in order to change sign. That is there must be some $x_1 \in (x_-, x_+)$ such that $P(x_1) = 0$ and we should be able to factor $(x - x_1)$ out of $P(x)$.

▷ To factor $(x - x_1)$ out of $P(x)$ we simply perform polynomial division of $P(x)$ by $(x - x_1)$. The remainder of the division, if any, must be a polynomial of an order smaller than the polynomial order of $(x - x_1)$. That is the polynomial order of the remainder must be zero, which means that the remainder is simply a constant r_0:

$$P(x) = (x - x_1)P_1(x) + r_0$$

Substituting x_1 for x into the above equality yields

$$P(x_1) = 0 \cdot P_1(x) + r_0 = r_0$$

But $P(x_1) = 0$ by choice of x_1 and therefore $r_0 = P(x_1) = 0$.

We are thereby left with an even-order polynomial $P_1(x)$. Or the polynomial $P(x)$ could have had an even order to begin with. Therefore let's now consider the situation for an even-order $P(x)$. In this case we have $P(x) \to +\infty$ for $x \to \pm\infty$ (recalling that $a_N = 1$) and we cannot employ the continuity to argue the existence of an intersection of the polynomial's graph with the horizontal axis. Nor is there a guarantee that there *is* an intersection, as e.g. can be seen with the 2nd-order polynomials in Fig. 1.1. However there *may* be an intersection. In that case we perform the same polynomial division, as we did in case of an odd-order polynomial, obtaining an odd order polynomial.

Repeating this step, we either get stuck at some even-order polynomial, or we manage to factor the original polynomial $P(x)$ into N different first-order factors

$$P(x) = \prod_{n=1}^{N} (x - x_n)$$

thereby obtaining N (or less) different roots.

Goal 1.2 completed. ■

1.2.2 Complex case

Now that we demonstrated the fundamental theorem of algebra for the case of real polynomials and purely real roots we might be ready to deal with the general case of (potentially) complex roots of (potentially) complex polynomials. Maybe the approach we used for the real case can be somehow generalized and/or modified to handle the complex case as well?

Goal 1.3. Starting with the previously used "purely real" approach as a basis, design a new one which handles the complex case. *Hint:* we cannot plot graphs of complex-valued functions of a complex argument, since that would require 4 dimensions. But we could at least imagine how the function's value "moves" in the complex plane in some specific situations that we construct.

In case of complex numbers, the domination of the term x^N at large $|x|$ turns into $P(x) \to \infty$ at $x \to \infty$, where ∞ is the complex infinity that is lying "in all directions" on the complex plane. Therefore there is no "sign change" like we had with real numbers, but we still wish to argue that we are going to hit $P(x) = 0$ for some x.

In case of real numbers the zero value of $P(x)$ occurring at $x = x_1$ was located *between* two (sufficiently large in absolute magnitude) numbers $x_- < 0$ and $x_+ > 0$ such that $P(x_-) < 0$ and $P(x_+) > 0$. This 1-dimensional "between" needs to turn into a "within" on the 2-dimensional complex plane. Specifically, "between two large numbers" needs to turn into a "within some large closed contour". The simplest possible closed contour is arguably a circle. Let's therefore turn the 1-dimensional interval $[x_-, x_+]$ into a large circle.

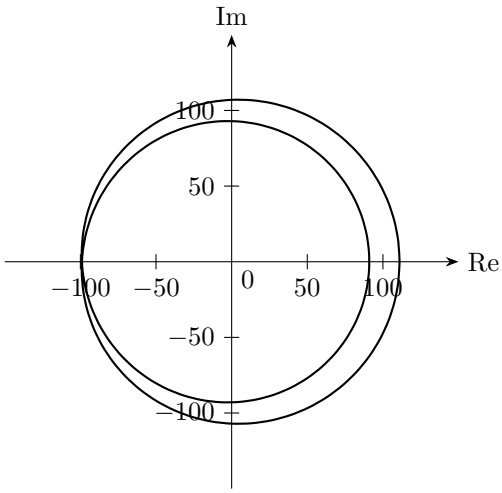


Figure 1.4: Trajectory covered by $x^2 + x + 1$ for $|x| = 10$.

If we fix the value of $|x|$ at some large value r, we are going to have $x = re^{i\varphi}$ and

$$P(x) \approx x^N = (re^{i\varphi})^N = r^N e^{iN\varphi}$$

that is, as φ varies from 0 to 2π, the value of x draws a full circle of radius r centered at the origin on the complex plane and the value of $P(x) \approx x^N$ draws N full circles of radius r^N centered *approximately* at the origin.[2] Since the equality $P(x) \approx x^N$ is approximate, these N circles might slightly differ from each other, but the N-th circle ends exactly at the start of the first circle (since x is identical at $\varphi = 0$ and $\varphi = 2\pi$ and so is x^N). Fig. 1.4 illustrates.

Now the origin point lies inside these circles of radius r^N. But is this origin reachable by some x, that is is there such x that $P(x) = 0$? Let's try to "scan" all the points x lying inside the circle of radius r. If we continuously decrease the radius r, at some point the origin will lie outside of the trajectory described by $P(x)$. At least if $a_0 \neq 0$. Indeed, at small $|x|$ the term a_0 dominates, therefore $P(x) \approx a_0 \neq 0$ and the origin must lie outside of the trajectory described by $P(x)$ (Fig. 1.5)

▹ If $a_0 = 0$, we immediately have $P(0) = 0$, that is $x_1 = 0$.

So, at large r the origin lies inside the trajectory, while at small r the origin lies outside the trajectory. By continuity, at some r the origin must lie *on* the trajectory (Fig. 1.6), that is at some x_1 we are going to have $P(x_1) = 0$, and we can perform the same polynomial division step as we did in the real case. The difference is that in the complex case our argument didn't assume odd or even order of $P(x)$, therefore there is always a solution of $P(x) = 0$ and we always can perform the polynomial division step. Thus we always obtain N solutions.

Goal 1.3 completed. ■

[2] Formally, N is the *winding number*.

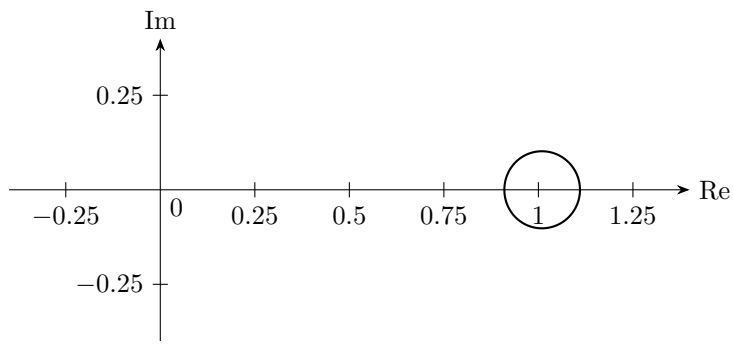


Figure 1.5: Trajectory covered by $x^2 + x + 1$ for $|x| = 0.1$.

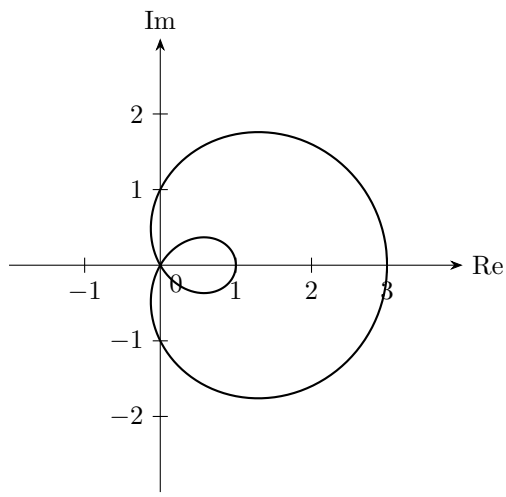


Figure 1.6: Trajectory covered by $x^2 + x + 1$ for $|x| = 1$. The trajectory is going twice through the origin because both roots of the polynomial, being mutually conjugate, are having the same absolute magnitude $|x_1| = |x_2| = 1$.

1.3 Fundamental theorem of arithmetic

The fundamental theorem of arithmetic states the uniqueness of the prime number factoring of any natural number. The standard proof of this theorem is based on Euclid's lemma, which states that a prime dividing a product mn must divide at least one of m or n. Now, the statement of Euclid's lemma is quite intuitive and is not too difficult to come up with. However its common proofs, despite being short, are not too intuitive and not too easy to find. We shall take a look at employing an "experimental" approach to find a way to prove the entire theorem.

Goal 1.4. Find a path leading to some kind of a proof or a sketch of a proof of the fundamental theorem of arithmetic. Note that we don't care if the proof is longer than the common proofs of the theorem, our focus is intuition, not conciseness.

1.3.1 Modified Euclid's lemma

As we want to find out whether an alternative factorization of a natural number into primes is possible, it makes sense to look into the question of divisibility of products of numbers by other numbers, maybe just starting with the products of two numbers.

Goal 1.4a. Study the question of when and why a product of two natural numbers $a \cdot b$ would be divisible by another natural number c and see what kind of further ideas we can derive from that. Note that we don't have to arrive at Euclid's lemma in its common form, we might come up with a somewhat different idea. *Hint:* start with $c = 2$ and increment c one by one.

Let's start with divisibility by 2, which is probably the simplest case, as we can talk in terms of odd and even numbers. We can notice that a product $a \cdot b$ of two natural numbers a and b is odd iff both factors are odd. If at least one of the two factors is even, the product is also even. The reason for this is demonstrated by the following three equalities listing 3 different cases of a and b being odd or even:

$$\begin{aligned}
a \cdot b &= 2n \cdot 2m = 2 \cdot (2nm) \\
a \cdot b &= (2n+1) \cdot 2m = 2 \cdot ((2n+1)m) \\
a \cdot b &= (2n+1) \cdot (2m+1) = 2 \cdot (2nm + n + m) + 1
\end{aligned}$$

Notationally we write $2|ab \iff (2|a) \vee (2|b)$ where $2|a$ means "2 divides a", that is a is divisible by 2, and $2|ab$ means "2 divides the product ab".

The same property for divisibility by 3 can be analyzed using the following equalities:

$$\begin{aligned}
a \cdot b &= 3n \cdot 3m = 3 \cdot (3nm) \\
a \cdot b &= (3n + r) \cdot 3m = 3 \cdot ((3n + r)m) \\
a \cdot b &= (3n + r_1) \cdot (3m + r_2) = 3 \cdot (3nm + mr_1 + nr_2) + r_1 r_2
\end{aligned}$$

where $r_1 \neq 0$ and $r_2 \neq 0$ are the division remainders. From the first two lines we can already see that $(3|a) \vee (3|b) \implies 3|ab$, however we're not immediately sure about the third line. Let's see. In the third line we obtain a division remainder determined by the term $r_1 r_2$ (with $1 \leq r_i < 3$, where $i = 1, 2$), unless $r_1 r_2$ is

also a multiple of 3. However the only options for $r_1 r_2$ are to be equal to $1 \cdot 1$, $1 \cdot 2$ and $2 \cdot 2$, that is 1, 2 and 4 respectively, and none of those is a multiple of 3. Therefore the 3rd line describes the case of $\neg(3|ab)$ (that is $a \cdot b$ being not divisible by 3), and we obtain $3|ab \iff (3|a) \vee (3|b)$, same as with division by 2.

In comparison, the same analysis for 4 essentially gives

$$a \cdot b = (4n + r_1) \cdot (4m + r_2) = 4 \cdot (4nm + mr_1 + nr_2) + r_1 r_2$$

(where we dropped the first two equalities, since they are not too interesting), however this time $1 \leq r_i < 4$, and in case $r_1 = r_2 = 2$ we would have $r_1 r_2 = 4$. Therefore it's possible to have none of the factors divisible by 4 but the whole result divisible by 4. The simplest example is given by $2 \cdot 2 = 4$. Therefore we only have the obvious $(4|a) \vee (4|b) \implies 4|ab$, but the converse implication doesn't hold: $4|ab \not\Longrightarrow (4|a) \vee (4|b)$.

Looking at the above analysis of divisibility by 4, by now we already can realize that the divisibility property $p|ab \iff (p|a) \vee (p|b)$ doesn't seem to hold for non-prime numbers p, however maybe it still might hold for primes. Let's check 5:

$$a \cdot b = (5n + r_1) \cdot (5m + r_2) = 5 \cdot (5nm + mr_1 + nr_2) + r_1 r_2$$

It's getting a bit too cumbersome to analyze all possible combinations of r_1 and r_2, although for the number 5 they are still manageable. Let's try to find a more general approach to analyze the divisibility of $r_1 r_2$ by 5. The divisibility of $r_1 r_2$ by 5 means that $\exists k \in \mathbb{N}\colon r_1 r_2 = 5k$. That is the number $l = r_1 r_2 = 5k$ is a *common multiple* of 5 and r_1. This means we are looking for a common multiple of 5 and r_1, and we are specifically looking for the one strictly smaller than the one attained at $r_2 = 5$, since we have a restriction $r_i < 5$. Our assumption is that we won't find one, which would mean that divisibility by 5 follows the same pattern as the one we observed earlier with 2 and 3: $5|ab \overset{?}{\iff} (5|a) \vee (5|b)$. Actually, we could explicitly confirm this assumption by diligently trying out all possible combinations of r_1 and r_2. Manual exhaustive tryouts is not what we're looking for, but at least they can confirm that we haven't discovered any contradictions to our assumption yet.

Generally, for a given number p we have

$$a \cdot b = (pn + r_1) \cdot (pm + r_2) = p \cdot (pnm + mr_1 + nr_2) + r_1 r_2$$

and we want to know whether $r_1 r_2$ would be divisible by p, under the restriction $1 \leq r_i < p$. As we have seen for $p = 5$, this is equivalent to the question of the existence of a common multiple of p and r_1, smaller than the one equal to their product. In other words, we want to know whether the *least common multiple* of p and r_1 is equal to pr_1, which is notationally written as $\operatorname{lcm}(p, r_1) = p \cdot r_1$. If that's the case, we would have thereby confirmed that our divisibility property holds for p:

$$\Big(\operatorname{lcm}(p, r_1) = p \cdot r_1 \quad \forall 1 \leq r_1 < p\Big) \implies \Big(p|ab \iff (p|a) \vee (p|b)\Big) \tag{1.9}$$

Should we find that the left-hand side of (1.9) holds for all prime p, so would the right-hand side.

▷ Suppose we have demonstrated that the left-hand side of (1.9) holds for all prime p. This would turn (1.9) into Euclid's lemma in its common form. Thereby (1.9) can be seen as a modified version of Euclid's lemma, containing an explicit extra condition that is the left-hand side of (1.9), instead of simply stating the right-hand side.

Goal 1.4a completed. ■

By applying the right-hand side of (1.9) recursively we would be able to generalize it to a product of n factors: $p|a_1a_2\dots a_n \iff \exists a_i\colon p|a_i$. Assuming $a_1a_2\dots a_n$ is a factorization of some number into primes (where a_i are the primes), there would thereby be no other prime p which can divide the entire number, meaning that no other factorization into primes is possible. We shall therefore concern ourselves now with the left-hand side of (1.9) and dive deeper into the idea of the least common multiple.

Goal 1.4b. Study the least common multiples to demonstrate that the left-hand side of (1.9) holds for all prime p.

1.3.2 LCM and GCD

Given two numbers m and n, their obvious common multiple is $l = m \cdot n$, that is l is a multiple of m and is also a multiple of n. Apparently, there are other common multiples l' of m and n.

It's totally obvious that there are such other multiples which are larger than l. However sometimes some of these other common multiples l' might be also smaller than l. Let λ be the smallest of all l'. That is λ is the *least common multiple*, denoted as $\operatorname{lcm}(m, n)$:

$$\lambda = \operatorname{lcm}(m, n) = \alpha \cdot m = \beta \cdot n \tag{1.10}$$

where $\lambda \leq m \cdot n$ (otherwise $m \cdot n$ would have been a smaller common multiple than λ).

Now if we look at how common multiples come about to be, we can notice a strict periodicity related to that. E.g. let's look at the first few nonnegative integer numbers and the remainders of their division by 6 and 4 shown in Table 1.1. We can notice that each of the two remainder rows is strictly periodic. Furthermore, each subsequent number in a given row is fully determined by its preceding number. That is, we could fully discard the top row of the table, and simply build each of the remainder rows in an incremental fashion.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$k \bmod 6$	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2
$k \bmod 4$	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2

Table 1.1: Periodicity of remainders.

This means that after we reach the common multiple at $k = 12$ (which is also the least one), we are fully restarting the cycle which was started at $k = 0$, since both entries in the table for $k \bmod 6$ and $k \bmod 4$ at $k = 12$ are identical to

those at $k = 0$ (specifically, both of them are zero). That is, starting at $k = 12$ we are fully repeating the cycle which was started at $k = 0$. And then the next cycle will start at $k = 12 + 12 = 24$, the next at $k = 24 + 12 = 36$, etc., each cycle being fully identical to other cycles. This also means there are no common multiples of 6 and 4, other than the ones occurring at the multiples of the least common multiple at 12, since no other common multiples were occurring within the first cycle, which is simply being repeated afterwards.

Obviously, the picture would be the same for any other two natural numbers m and n taken instead of 6 and 4, the cycle will fully restart after their least common multiple. This means that *all common multiples of two numbers m and n are multiples of their least common multiple.* In particular, this applies to $m \cdot n$, meaning that $m \cdot n$ is a multiple of $\mathrm{lcm}(m, n)$:

$$m \cdot n = d \cdot \mathrm{lcm}(m, n) \qquad \text{for some } d \in \mathbb{N}$$

and respectively

$$\mathrm{lcm}(m, n) = \frac{m \cdot n}{d} \tag{1.11}$$

Now by (1.10)

$$m \cdot n = d \cdot \mathrm{lcm}(m, n) = d \cdot \alpha \cdot m$$

therefore $d \cdot \alpha = n$ and $d|n$. Similarly,

$$m \cdot n = d \cdot \mathrm{lcm}(m, n) = d \cdot \beta \cdot n$$

therefore and $d \cdot \beta = m$ and $d|m$. Apparently, d is a *common divisor* of m and n. We conjecture that d is the greatest of the common divisors of m and n.

Indeed, suppose m and n have a yet greater common divisor $d' > d$. This would mean that $\lambda' = mn/d'$ is another common multiple of m and n. But since $d' > d$, it would follow that $\lambda' < \lambda$ and thereby violate the fact that λ is the least common multiple. It follows that d is the greatest possible common divisor of m and n, denoted $\gcd(m, n)$ and we can rewrite (1.11) as

$$\mathrm{lcm}(m, n) = \frac{m \cdot n}{\gcd(m, n)} \tag{1.12}$$

E.g. for two prime numbers p and q there are no common divisors besides 1, therefore $\gcd(p, q) = 1$ and therefore $\mathrm{lcm}(p, q) = p \cdot q$.

Now, if we return to (1.9), then in its left-hand side we have the equality $\mathrm{lcm}(p, r_1) = p \cdot r_1$, where p is prime and $r_1 < p$. Apparently p and r_1 cannot have common divisors besides 1, therefore the left-hand side of (1.9) is true for all prime p, and therefore so is its right-hand side as well, thereby confirming Euclid's lemma in its common form, thereby confirming the fundamental theorem of arithmetic.

Goal 1.4b completed. ■

Goal 1.4 completed. ■

Summary of Chapter 1

- We have demonstrated the usage of "experimental" approach of finding solutions to mathematical problems. We used it to find the idea of completing the square for the purpose of solving quadratic equations, where

we experimented with parabolic graphs to come to the idea that we just need to shift the parabola to a position symmetric around the y axis.

- We observed the behavior of graphs of polynomials to build a sketch of a constructive proof of the fundamental theorem of algebra in the real case, which we then extended to the complex case.
- We used the experimental approach to find a way to prove the fundamental theorem of arithmetic, where we started by looking into divisibility of products of two numbers by other numbers and quickly arrived at a modified (reduced) version of Euclid's lemma. Then we promoted it to a full version by doing some experiments on the behavior of the least common multiple of two numbers.

Chapter 2

Real numbers

Introduction of real numbers as the values of infinite decimal fractions. Relationship to the postulate of Cauchy sequence convergence and to other equivalent postulates. Relationship to continuity of arithmetic operations. Uncountability of real numbers: do we need that many numbers? Continuity of the real line as justification of uncountability.

2.1 Periodic decimal fractions 34
2.2 Non-periodic decimal fractions 36
2.3 Decimal fraction convergence 37
2.4 Limits of sequences 38
2.5 Postulated existence of real numbers 42
2.5.1 Cauchy sequences 42
2.5.2 The least bound axiom 45
2.5.3 The Dedekind cut 46
2.6 Continuity of operations 47
2.7 Comparisons 49
2.8 Uncountability 52
2.8.1 Countable sets 52
2.8.2 Cantor's diagonal argument 55
2.8.3 Uncountability implications 56
2.9 Continuity of the real line 58
Summary of Chapter 2 60

Axiomatic introduction of real numbers (by postulating their properties), which is commonly done in many sources, looks like the one which makes sense, but leaves certain questions open. In particular:

- Why are two infinitely close real numbers considered equal? Intuitively, the first thought would be that one of them is "infinitesimally larger than the other".

- Formal introduction of real numbers needs an additional postulate, e.g. "the least bound axiom", although some other equivalent options are also in use. From a certain perspective, this need feels like a "bug in the reality". Why do we need to postulate anything like that?

- It can be demonstrated that an absolute majority of the real numbers cannot ever be "referred to". That is, there is no way that the numbers from this "absolute majority" will be ever mentioned in some text, or obtained as an answer to a problem, or described in any other way (except in "bulk descriptions", like. e.g. referring to the entire set of real numbers at once). This is formally expressed by saying that the majority of the real numbers is not *computable*[1] (where computability is being used as a generic formal way of "referring to" a number). This feels pretty weird. Does it have to be that way?

In fact, there is a variety of different ways to define real numbers. One of these ways might not look like the most elegant from a formal mathematical point of view, but arguably could be the most intuitive and insightful one. It also, possibly, if not completely answers the questions asked above, then at least makes them not so much bothering. This is introduction of real numbers via decimal fractions.

> ▷ In principle, this introduction approach is not restricted to decimal fractions. Any other base will do. E.g. we could use binary, or ternary, or whatever-base fractions.

Also, during this approach we will come to the necessity to get into other fundamental ideas of calculus, in particular the idea of limits of sequences and related topics.

Motivational starter for Chapter 2

Extend the set of rational numbers to accommodate the results of the inverses of the exponentiation operation: taking the root and the logarithm.

> ▷ The introduction of real numbers will only partially achieve the just declared goal of the chapter. It is fully achieved by complex numbers. Complex numbers have limited coverage in this text, but one possible motivation for their introduction is discussed in Subsection 7.7.3.

2.1 Periodic decimal fractions

We shall start by reviewing the idea of periodic decimal fractions being used to represent rational numbers. It is well known that some rational numbers have no difficulty being represented as decimal fractions:

$$\frac{1}{10} = 0.1 \qquad \frac{2}{5} = 0.4 \qquad \frac{15}{8} = 1.875$$

while some others cannot be represented by a finite decimal fraction and we should use an infinitely long one, e.g.

$$\frac{1}{3} = 0.3333333\ldots \tag{2.1}$$

[1] As it is obviously generally impossible to represent a real number by a finite decimal fraction, a real number being computable doesn't mean that we can compute a full decimal fraction representation of that number, but rather that we can describe a way (specify an algorithm) to compute that number to any given precision. Such algorithm is expected to have a finite description.

This is actually the key moment, which effectively enters a rabbit hole of an unknown depth:

- In order to represent numbers like $\frac{1}{3}$ by decimal fractions we need to start dealing with infinities (in the form of an infinite length of the respective decimal fraction).
- Furthermore, by understanding the infinite decimal fraction $0.33333\ldots$ as a representation of $\frac{1}{3}$ we have to agree that *two infinitely close numbers are equal*. Indeed, $0.33333\ldots$ is not really equal to $\frac{1}{3}$, but is rather infinitely close to $\frac{1}{3}$. Treating $0.33333\ldots$ as being equal to $\frac{1}{3}$ also immediately implies that
$$0.99999\ldots = 3 \cdot 0.33333\ldots = 3 \cdot \frac{1}{3} = 1 \tag{2.2}$$

Now, with rational numbers the situation is actually not completely bad yet. The "rescue" comes from the fact that, as it turns out, all infinite decimal fractions representing rational numbers are periodic decimal fractions (where finite decimal fractions can be understood as having an infinite number of repeating zeros at the end, thereby also becoming periodic). Indeed, suppose we convert a rational number $\frac{m}{n}$ to a decimal fraction using long division. At some point we are going to run out of the numerator digits and will be just repeatedly dividing the remainders from the previous step. However there are only n possible different remainders from a division by n, therefore the situation is going to repeat latest after n steps. Actually, latest after $n-1$ steps, since a zero remainder just terminates the process.

Thereby we could introduce some kind of finite notation for these periodic infinite decimal fractions. The notation convention details may vary across different countries of the globe. One common convention is to write the periodic part in parentheses: $\frac{1}{3} = 0.33(3)\ldots$, $\frac{131}{84} = 1.55(952380)\ldots$. The bottom line being: all rational numbers can still be notated by a finite string of symbols. The latter fact also allows us to "restore" the rational number from its periodic infinite decimal fraction. E.g.

$$\begin{aligned} 0.7123123(123)\ldots &= 0.7 + 0.0123123(123)\ldots = 0.7 + 0.1 \cdot 0.123123(123)\ldots \\ &= 0.7 + 0.1 \cdot 123 \cdot 0.001001(001)\ldots = 0.7 + 0.1 \cdot 123 \cdot \frac{0.999999(999)\ldots}{999} \\ &= 0.7 + 0.1 \cdot 123 \cdot \frac{1}{999} = \frac{7}{10} + \frac{123}{9990} = \frac{7 \cdot 999 + 123}{9990} \end{aligned} \tag{2.3}$$

▷ Thereby not only any rational number is representable by a periodic infinite decimal fraction, but also conversely, any periodic infinite decimal fraction represents a rational number.

This in turn, could allow us to understand periodic infinite decimal fractions in a purely notational sense: by definition, they would just represent some rational numbers, which are supposed to be "restored" from the decimal fraction by a formal application of the idea demonstrated in (2.3). Respectively we wouldn't need to explicitly postulate the equality of "infinitely close numbers", as it won't be used anywhere else besides the formal conversion algorithm used in (2.3). However, this "way around the infinity" would no longer be possible once we get to nonperiodic decimal fractions in Section 2.2.

2.2 Non-periodic decimal fractions

It is well known that $\sqrt{2}$ cannot be represented as a ratio of two integers.

▷ Indeed, having $\sqrt{2} = \frac{m}{n}$ would imply $2 = \frac{m^2}{n^2}$. Now imagine both the numerator m^2 and the denominator n^2 being factored into primes. All these factors should cancel except for a single remaining factor 2 in m^2. However both m^2 and n^2 would have an even number of factors 2 each, therefore they cannot cancel to a single factor 2.

A similar argument can be made more generally for $\sqrt[m]{n}$.

From a certain perspective, it's not fully surprising that the values of such expressions as $\sqrt{2}$ are not expressible in the form $\frac{m}{n}$. Division is the multiplication's inverse, while root taking is the exponentiation's inverse. Why should a different inverse operation produce the same kind of values?

Alright, we cannot use the notation $\frac{m}{n}$ to represent the results of taking roots. But why don't we use the notation $\sqrt[m]{n}$ to directly represent the result of the expression $\sqrt[m]{n}$, similarly to how the notation $\frac{m}{n}$ not only expresses the division of m by n, but is also used to represents the result of that division? We could actually try, and this would even work to an extent, e.g.

$$\sqrt[3]{2} \cdot \sqrt[2]{3} = \sqrt[6]{8} \cdot \sqrt[6]{27} = \sqrt[6]{8 \cdot 27} = \sqrt[6]{216}$$

So we could consider $\sqrt[m]{n}$ (where $m, n \in \mathbb{Z}$ or maybe $m, n \in \mathbb{Q}$) as some kind of canonical notation form, similar to $\frac{m}{n}$. However it fails once addition or subtraction is involved. E.g. we cannot convert the expression $1 + \sqrt{2}$ to the form $\sqrt[m]{n}$. So, while with rational numbers and the four arithmetic operations on them (plus raising into integer powers) we can always represent the results in the form $\frac{m}{n}$ (where $m, n \in \mathbb{Z}$), the same doesn't hold for the notation $\sqrt[m]{n}$, which makes the latter not so useful.

So, in the absence of better ideas we may decide to resort to infinite decimal fractions as a means to represent the results of root taking and especially the results of further combining such values. Since we have found that periodic decimal fractions always represent rational numbers, it follows that the infinite decimal fractions representing irrational numbers will be nonperiodic.

This also precludes the purely notational interpretation of infinite decimal fractions, which we suggested at the end of Section 2.1. With rational numbers represented by infinite decimal fractions we always had the other representation in the form $\frac{m}{n}$, so for any infinite decimal fraction we could have said: it actually represents some number of the form $\frac{m}{n}$, obtained from the decimal fraction via the algorithm shown in (2.3). However we are not having a similar algorithm for irrational numbers and it doesn't look like this algorithm could exist either. To being with, we wouldn't be able to even provide the input to this algorithm, since this input would need to be an infinite string of decimal figures (which we can't represent in a finite form, since it's nonperiodic).

Thus we are going to accept infinite decimal fractions as a fundamental generic way to specify numbers, which includes both rational and irrational numbers, commonly making up the set of the real numbers $\mathbb{R}$. And this decision per se, as already discussed in Section 2.1, implies that we should treat infinitely close numbers as equal ones. And this fact alone is already motivating the introduction of the idea of the limits of sequences.

▷ As we cannot write an infinitely long string of digits, a nonperiodic decimal fraction is supposed to be specified in terms of some finite description. E.g. the number π is specified by describing it as the ratio of the circumference of a circle to the circle's diameter. The number $\sqrt{2}$ is specified by describing it as a positive number x such that $x^2 = 2$. Some numbers can be defined as specific limits, or solutions to specific problems, etc.

2.3 Decimal fraction convergence

The idea that infinitely close numbers should be treated as equal ones originates (at least within the course of this text) from the usage of infinite decimal fractions. When infinite decimal fractions were used to denote rational numbers, we probably could have stayed quite handwavy about this, if only due to the approach of (2.3), which allowed purely notational interpretation of infinite decimal fractions. However in case of irrational numbers, infinite decimal fraction representation becomes fundamental and we might appreciate a more formal treatment of the idea of "an infinite decimal fraction being infinitely close to a particular number". We shall first develop that formal treatment for the infinite decimal fractions representing rational numbers (where we are having a reasonably firm "finite" ground) and then proceed to irrational ones.

Convergence to a known limit

Let's start with the representation of the rational number $\frac{1}{3}$ by the infinite decimal fraction $0.33333\ldots$. To look at this formally, we can understand the infinite fraction $0.33333\ldots$ as a sequence of finite decimal fractions 0.3, 0.33, 0.333, 0.3333, etc. That is, we are having a sequence $\{a_n\}$, whose terms are made of truncated versions of the infinite decimal fraction $0.33333\ldots$:

$$a_1 = 0.3,\ a_2 = 0.33,\ a_3 = 0.333,\ a_4 = 0.3333 \text{ etc.} \tag{2.4}$$

The larger is the index n of the sequence term a_n, the better it approximates the "true result" $\frac{1}{3}$, because the error cannot exceed one decimal figure in the last position of the fraction, that is 10^{-n}. Thus

$$\varepsilon_n \leq 10^{-n}$$

where $\varepsilon_n = |a_n - \frac{1}{3}|$ is the approximation error. Intuitively, as $n \to \infty$, the terms a_n *converge* to the "true value" $\frac{1}{3}$. We say that the limit of the sequence $\{a_n\}$ as $n \to \infty$ is $\frac{1}{3}$:

$$\lim_{n\to\infty} a_n = \frac{1}{3}$$

So, how do we formally define the convergence of a sequence? Intuitively, a_n converge to $\frac{1}{3}$ iff the errors ε_n converge to zero:

$$\lim_{n\to\infty} a_n = \frac{1}{3} \iff \lim_{n\to\infty} \varepsilon_n = \lim_{n\to\infty} \left| a_n - \frac{1}{3} \right| = 0 \tag{2.5}$$

It doesn't look like (2.5) changes anything, since we just redefined one limit in terms of the other. Although, it actually might simplify the reasoning a

bit, since it might be easier to reason about convergence to zero than about convergence to a generic limit. We could therefore generally postulate

$$\lim_{n\to\infty} a_n = a \overset{\Delta}{\iff} \lim_{n\to\infty} \varepsilon_n = \lim_{n\to\infty} |a_n - a| = 0 \tag{2.6}$$

Convergence to an unknown limit

We haven't fully defined the idea of sequence convergence yet, but we already kind of intuitively agree that $\lim_{n\to\infty} 10^{-n} = 0$, and so is the limit of any sequence, the magnitude of whose terms is bounded by 10^{-n}, like the sequence ε_n in (2.5). This shall do for now.

We shall now apply the same ideas to the convergence of a nonperiodic decimal fraction, e.g. the one for

$$\sqrt{2} = 1.41421356237309504880168872 42\ldots \tag{2.7a}$$

which corresponds to the sequence of its truncated versions

$$a_1 = 1.4,\ a_2 = 1.41,\ a_3 = 1.414,\ a_4 = 1.4142,\ \ldots \tag{2.7b}$$

▷ Technically, the fractions (2.7b) can be obtained by brute-force trial. First, we find the largest possible decimal fraction a_1 with one decimal place after the decimal period, such that $(a_1)^2 \leq 2$. Then we find the largest possible decimal fraction a_2 with two decimal places after the decimal period, such that $(a_2)^2 \leq 2$. Etc.

Of course, this is a far from optimal way of obtaining the fractions (2.7b), but, from the theoretical standpoint, it is only important that we don't need to involve the framework of real numbers to obtain them.

We "don't know" the value that the sequence (2.7b) is converging to (except that we know that this value is the value of $\sqrt{2}$), but at least we know that the approximation error of a_n is subject to the same assessment

$$\varepsilon_n \leq 10^{-n}$$

that we had with the sequence (2.4) for $\frac{1}{3}$. We shall therefore conclude that the sequence (2.7b) also converges, we just "do not know" the value that it converges to (we "don't know" in the sense that we don't have a way of writing out this value other than as $\sqrt{2}$).

Thus, in a way, we are considering every infinite decimal fraction as converging, since the approximation errors decay as 10^{-n}. *Effectively, this implicitly postulates the convergence of Cauchy sequences.* We shall discuss further details in Section 2.5.

2.4 Limits of sequences

Now that we have found out that dealing with infinite decimal fractions effectively means using the limits of sequences, we shall look into the concepts of limits in more detail. This concept is usually covered with a large amount of detail in any standard text on calculus, however we would like to focus on the intuitive understanding of the ideas hidden behind "the ε-N language".

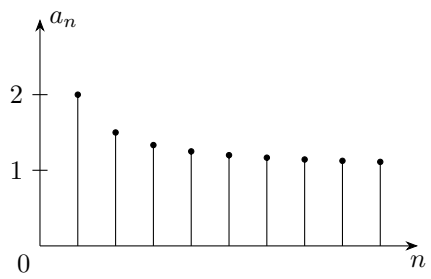


Figure 2.1: The sequence $a_n = 1 + \frac{1}{n}$ decays in absolute magnitude, but doesn't converge to zero.

By (2.6) we have reduced the general discussion of convergence to the question of convergence to zero, but we stopped somewhere around the intuitive conclusion that anything bounded by 10^{-n} in absolute magnitude is converging to zero. Apparently, this is too specific to be a general definition of convergence. E.g. $a_n = \frac{1}{n}$ is clearly not bounded by 10^{-n}, but intuitively we should agree that $\frac{1}{n}$ is converging to zero as well.

Goal 2.1. From scratch, step by step, develop a generic definition of convergence of a sequence to zero. Don't explicitly attempt to exactly repeat the "δ-ε language" of the commonly used definition, a more intuitive phrasing might be beneficial for our purposes here. *Hint:* construct some example sequences and test various definition ideas against those.

The first intuitive idea might be to simply say that a sequence $\{a_n\}$ converges if its terms decay in absolute magnitude:

$$|a_{n+1}| < |a_n| \quad \forall n \tag{2.8}$$

but a few counterexamples should quickly convince us that this is not going to do, where we would even have counterexamples of two kinds: those demonstrating that (2.8) is not sufficient and also those demonstrating that (2.8) is unnecessarily strong. That is, (2.8) is over- and under-conservative at the same time.

An example demonstrating that (2.8) is underconservative could be e.g. the sequence

$$a_n = 1 + \frac{1}{n} \tag{2.9}$$

(Fig. 2.1). The terms (2.9) clearly satisfy (2.8), but hardly anyone would say that (2.9) is converging to zero.

An example demonstrating that (2.8) is overconservative could be built from two overlapping converging sequences of different magnitudes, e.g.

$$a_n = \begin{cases} \frac{1}{n} & \text{if } n \text{ is odd} \\ \frac{2}{n} & \text{if } n \text{ is even} \end{cases} \tag{2.10}$$

(Fig. 2.2). Probably everyone would agree that (2.10) converges to zero, but it doesn't satisfy (2.8).

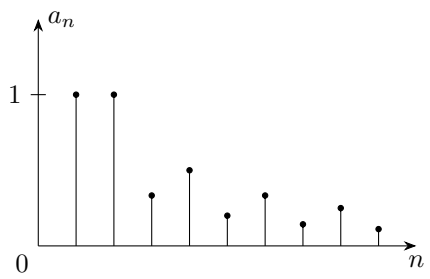


Figure 2.2: The sequence (2.10) converges to zero but doesn't monotonically decay in absolute magnitude, thereby violating (2.8).

These two problems, illustrated by (2.9) and (2.10) require two different amendments to (2.8), and we shall start with the second problem first.

Goal 2.1a. Address the problem illustrated by (2.10) and Fig. 2.2.

The trick addressing the problem of (2.10) is that instead of requiring the terms of the sequence itself to get progressively smaller, we shall require that the "tails" of the sequence get progressively smaller, where we define the n-th tail of the sequence $\{a_n\}$ as the subsequence consisting of all terms a_k for $k \geq n$. Thus the sequence itself is it's own 1st tail (assuming the first term of the sequence is a_1), the 2nd tail of the sequence consists of all the terms starting with a_2, the 3rd tail consists of all terms starting with a_3 etc. We shall denote the m-th tail of the sequence $\{a_n\}$ as $\{a_{m+}\}$, e.g.

$$\{a_{2+}\} = \{a_2,\ a_3,\ \dots\}$$

We shall define the "size" of the n-th tail as the supremum of the magnitudes of all its terms:

$$\|a_{n+}\| = \sup_{k \geq n} |a_k| \tag{2.11}$$

With this new concept in mind we are going to amend (2.10) to

$$\|a_{(n+1)+}\| < \|a_{n+}\| \quad \forall n \tag{2.12}$$

only to find out that, contrary to our expectation, it didn't fully solve the problem of (2.10). Fig. 2.3 illustrates.

As one can see in Fig. 2.3,

$$\|a_{1+}\| = \|a_{2+}\| > \|a_{3+}\| = \|a_{4+}\| > \dots$$

It looks like we need to allow non-strict inequalities in (2.12), but rewriting (2.12) as

$$\|a_{(n+1)+}\| \leq \|a_{n+}\| \quad \forall n$$

hardly makes sense, since it particularly allows $\|a_{(n+1)+}\| = \|a_{n+}\|\ \forall n$. This remaining issue, however, looks pretty similar to the one illustrated by (2.9), and we shall handle both issues together.

The introduction of sequence tails, however, still looks like a good idea. It did address the problem illustrated by (2.10), it's just that there is also the second problem in (2.10), which wasn't addressed yet.

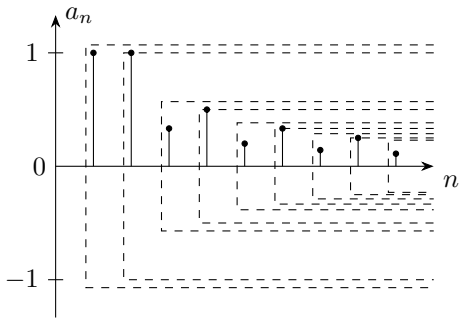


Figure 2.3: Tails of the sequence (2.10). For legibility purposes, the tail radii have been somewhat increased for odd n.

Goal 2.1a completed. ■

Goal 2.1b. Address the problem illustrated by (2.9) and Fig. 2.1, as well as the remaining part of the problem in (2.10).

To address the problem of (2.9) we shall notice that it's not sufficient for the terms or the tails to simply get smaller, they need to be getting *indefinitely smaller*. Now we need some formal language to express what does this "indefinitely smaller" mean. We suggest the following idea: no matter how small size we imagine, there will be some tail $\{a_{N+}\}$ fitting into this size, if we choose N to be large enough.

Let's put this into even more formal terms. Let $\varepsilon > 0$ denote our imagined small size. For this imagined size there will be some tail $\{a_{N+}\}$ such that $\|a_{N+}\| \leq \varepsilon$. That is

$$\lim_{n\to\infty} a_n = 0 \overset{\Delta}{\iff} \Big(\forall\varepsilon > 0 \implies \exists N\colon \|a_{N+}\| \leq \varepsilon\Big) \tag{2.13a}$$

where $\|a_{N+}\|$ is by (2.11).

It's not difficult to realize that (2.13a) is essentially the usual "ε-N" definition of the zero limit of a sequence:

$$\lim_{n\to\infty} a_n = 0 \overset{\Delta}{\iff} \Big(\forall\varepsilon > 0 \implies \exists N\colon \forall n > N \implies |a_n| < \varepsilon\Big) \tag{2.13b}$$

where we simply expanded the definition of the sequence tail size $\|a_{N+}\|$, plus nonstrict inequalities were replaced by strict ones, but the latter does not really matter, we could have also written

$$\forall\varepsilon > 0 \implies \exists N\colon \forall n \geq N \implies |a_n| \leq \varepsilon$$

instead.

Goal 2.1b completed. ■

Goal 2.1 completed. ■

Goal 2.2. Adapt the definition (2.13a) to allow for a (given a priori) nonzero limit value.

In order to adapt the definition (2.13a) to nonzero limits, we just need to adapt the definition of the tail size (2.11):

$$\|a_{n+}\| = \sup_{k \geq n} |a_k - a| \tag{2.14}$$

that is, we are measuring the tail size "relatively to the tentative limit a".

▷ We could have included the tentative limit a into the notation for the tail size, e.g. notating the tails size as

$$\|a_{n+} - a\| = \sup_{k \geq n} |a_k - a|$$

however the notation (2.14) is simpler and will prove to be more intuitive in the discussion of Cauchy sequences in Section 2.5.

With the amendment (2.14), the definition (2.13a) becomes applicable for arbitrary limit values:

(convergence of a sequence to a known limit a)

$$\lim_{n\to\infty} a_n = a \overset{\Delta}{\iff} \Big(\forall \varepsilon > 0 \implies \exists N \colon \|a_{N+}\| \leq \varepsilon\Big) \tag{2.15a}$$

where $\|a_{N+}\|$ is given by (2.14).

Respectively, by the expansion of the sequence tail size definition (2.14) in (2.15a), we obtain the classical definition of the sequence limit in the "ε-N language":

(convergence of a sequence to a known limit a in "ε-N" terms)

$$\lim_{n\to\infty} a_n = 0 \overset{\Delta}{\iff} \Big(\forall \varepsilon > 0 \implies \exists N \colon \forall n > N \implies |a_n - a| < \varepsilon\Big) \tag{2.15b}$$

Goal 2.2 completed. ■

2.5 Postulated existence of real numbers

2.5.1 Cauchy sequences

The switching from zero limits in (2.13a) to arbitrary limits in (2.15a) is possible only if the value of the limit is known (or at least guessed) in advance, and with a randomly given infinite decimal fraction this is not necessarily the case. For some infinite decimal fractions we still might know the value in some way, e.g. for (2.7a) we know that this value is a square root of 2 and we might somehow use this knowledge to assess the differences $|a_k - a|$ in (2.14). However in the general case we might be missing such extra knowledge, or it might be difficult or plainly impossible to use that knowledge.

On the other hand, we have postulated that any infinite decimal fraction denotes a value. In other words, any infinite decimal fraction converges. And this is still in agreement with (2.6) and (2.13a), since by (2.13a) the sequence

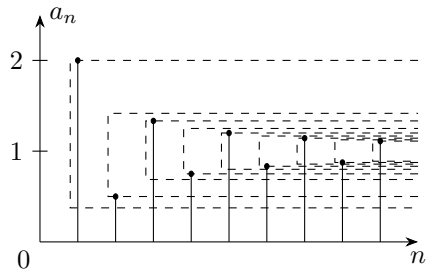


Figure 2.4: Diameters of sequence tails, as given by (2.16). For legibility purposes, the diameters have been somewhat increased from one of the sides.

of approximation errors ε_n is still converging to zero for any decimal fraction (since $0 \leq \varepsilon_n \leq 10^{-n}$). Or isn't it?

This is the question of how do we compute ε_n. For the sequence (2.4) we were simply computing ε_n as $\varepsilon_n = |a_n - \frac{1}{3}|$. But that involves knowing the limit in advance! On the other hand, for (2.7b) we didn't bother to provide any explicit expression for ε_n, we simply claimed that $\varepsilon_n \leq 10^{-n}$ by intuitive reasoning that the approximation error cannot exceed one decimal figure. However, can we provide any kind of formal definition for the approximation error in this case?

Goal 2.3. Find a way to define the "approximation error" which works in case the true value is not known. *Hint:* we do not need an exact error value, a reasonable assessment of the error's upper bound shall do.

The idea is to provide a way to assess the size of the sequence tails without using the limit value as the reference. One possibility would be to use the "diameters" of the sequence tails instead of their "radii measured with respect to a" (which is what (2.14) effectively does), where the diameter, as usual, is defined as the maximum possible distance between the terms of the tail:

$$\|a_{n+}\| = \sup_{k,l \geq n} |a_k - a_l| \tag{2.16}$$

Fig. 2.4 illustrates. The other possibility would be to still use the radius but measure it with respect to the first term of the tail:

$$\|a_{n+}\| = \sup_{k \geq n} |a_k - a_n| \tag{2.17}$$

Fig. 2.5 illustrates.

By using (2.16) or (2.17) as the tail size, we can apply the convergence definition (2.15a) even if we don't know the limit value a in advance. Except that in this case there is no way we can obtain the value of a or verify it from this definition. But this is fine, the infinite decimal fractions are serving as a means to denote the limit value by themselves anyway, we just need a formal statement about their convergence, and this statement is provided by the modified (2.15a):

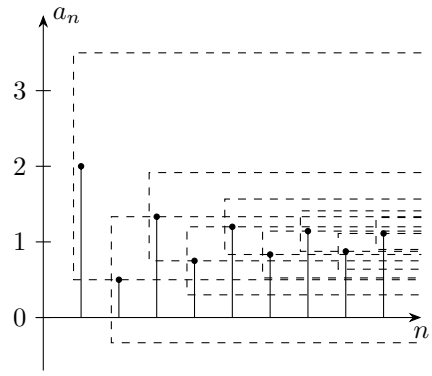


Figure 2.5: "Cauchy" radii of sequence tails, as given by (2.17).

(convergence of a Cauchy sequence)

$$\{a_n\} \text{ converges} \iff \Big(\forall \varepsilon > 0 \implies \exists N\colon \|a_{N+}\| \le \varepsilon\Big) \tag{2.18}$$

where $\|a_{N+}\|$ is commonly by (2.17), but in principle (2.16) will do equally well.

In the common mathematical language, the right-hand side of the statement (2.18) is effectively stating that $\{a_n\}$ is a Cauchy sequence.

Thus, if we use (2.14) as the definition of $\|a_{N+}\|$, we obtain the common definition of a converging sequence, while by using (2.17) or (2.16) as the definition of $\|a_{N+}\|$, we obtain the definition of a Cauchy sequence. It might thereby look as if both options are equivalent, but it turns out they are not: despite (2.14) intuitively having the same meaning as (2.17) and (2.16), all of them simply defining the tail size $\|a_{N+}\|$, they are not equivalent mathematically. We can easily infer (2.17) and (2.16) from (2.14), but not vice versa. This leads to what we could describe in the formal mathematical language as "we cannot demonstrate that any Cauchy sequence converges".

In particular, we cannot demonstrate that any infinite decimal fraction (seen as a sequence of its truncated fractions) converges. But, on the other hand, we kind of postulated the convergence of infinite decimal fractions by claiming that any infinite decimal fraction denotes a number. Our inability to demonstrate this convergence is closely related to that postulate. Indeed, on the set of rational numbers, there are at least some non-converging infinite decimal fractions, e.g. the infinite decimal fraction representing $\sqrt{2}$. It's only after we artificially extended the set of rational numbers to real numbers that this sequence became convergent. Thus, the statement about convergence of any infinite decimal fraction is, by itself, false and becomes true only with some additional postulated conditions.

The same holds for the statement about convergence of any Cauchy sequence: in the absence of additional postulated conditions it's not true. It turns out that it suffices to postulate the convergence of all infinite decimal fractions, the convergence of all remaining Cauchy sequences on $\mathbb{R}$ will follow (the readers are welcome to work out the details, if they like). Conversely, given the convergence

of all Cauchy sequences on $\mathbb{R}$, the convergence of infinite decimal fractions follows automatically, simply being a particular case of the former. Thus, we could postulate either the convergence of all infinite decimal fractions or the convergence of all Cauchy sequences on $\mathbb{R}$, both postulates are equivalent, despite the latter being seemingly a much stronger condition.

From the motivational perspective, we probably should stick to postulating the convergence of infinite decimal fractions, since we do not really need to postulate more than that and can infer the rest. From the formal presentation perspective, however, postulating the convergence of infinite decimal fractions is a bit awkward. Indeed, why did we choose 10 and not any other base for those infinite fractions? Or, why did we choose positional fractions to begin with, those are just some specific notational representations of numbers, so it's the latter that we should treat as fundamental entities. Thus, from the formal perspective, it is more elegant to postulate the Cauchy sequence convergence.

Goal 2.3 completed. ■

2.5.2 The least bound axiom

By now we have identified the reason why Cauchy sequence convergence needs to be postulated and cannot be proven: effectively, we are just postulating that "any infinite decimal fraction denotes some number". We wouldn't be able to prove it on the set of rational numbers, because this statement is simply false there (some infinite decimal fractions do not converge to rational limits). And it looks like we wouldn't be able to prove it on the set of real numbers either, since without postulating Cauchy sequence convergence real numbers do not even exist.

Instead of treating infinite decimal fractions as implicitly defining sequences consisting of truncated versions of those fractions, we can treat them as (unordered) sets consisting of truncated versions of those fractions. Respectively, instead of taking the limit of the sequence we can take the supremum (the least upper bound). This leads to a different approach to the definition of real numbers, where instead of defining them as sequence limits we define them as suprema. E.g. instead of

$$\sqrt{2} = 1.4142\ldots \triangleq \lim \{1.4,\ 1.41,\ 1.414,\ 1.4142,\ \ldots\}$$

we let

$$\sqrt{2} = 1.4142\ldots \triangleq \sup \{1.4,\ 1.41,\ 1.414,\ 1.4142,\ \ldots\}$$

It's not difficult to realize, that since the sequence of truncated versions of a decimal fraction is monotonically increasing, its limit coincides with its supremum. Thus, both approaches are equivalent.

In this case instead of postulating the convergence of Cauchy sequences, we postulate the existence of the supremum of any bounded (from above) set and this postulate is referred to as *the least bound axiom*. Similarly to Cauchy sequence convergence, this axiom cannot be proven on the set of rational numbers, because it doesn't hold there. E.g. consider a set S, consisting of all rational numbers q such that $q < \sqrt{2}$ (which can be formally constructed, without involving the real number $\sqrt{2}$, by defining it as a set including negative rational numbers, zero, and positive rational numbers such that $q^2 < 2$). Apparently,

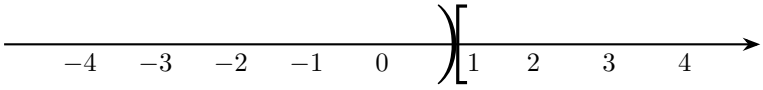


Figure 2.6: A Dedekind cut of the rational line at $x_0 = 1$. The number x_0 needs to be included either into the left or into the right interval. Here we choose to include it into the right interval.

$\sup S = \sqrt{2}$, however $\sqrt{2}$ doesn't belong to the set of rational numbers, and neither does the axiom's statement hold, if we don't have real numbers yet.

▷ Postulating the least bound axiom is not only equivalent to postulating Cauchy convergence in the context of the limits of infinite decimal fractions. Both postulates are generally equivalent. Convergence of Cauchy sequences implies the existence of suprema of the sets bounded from above, which can be demonstrated by building a monotonically increasing infinite sequence from the elements of such set. Conversely, the least bound axiom implies the convergence of Cauchy sequences, which can be demonstrated by fist noticing that existence of the supremum for sets bounded from above implies the existence of infimum (the greatest lower bound) for sets bounded from below, and then building a monotonically increasing or monotonically decreasing subsequence from a given Cauchy sequence.

2.5.3 The Dedekind cut

We have discussed that postulating the least bound axiom is equivalent to postulating the convergence of Cauchy sequences. Yet another equivalent postulate is based on the idea of *Dedekind cuts*. A Dedekind cut is just splitting the entire set of real numbers $\mathbb{R}$ into two nonoverlapping sets A and B, such that all elements of A are strictly less than all elements of B. Intuitively, we are just cutting the real line into the "left and right halves", where the cut can be made at an arbitrary place.

The idea is that if we apply a Dedekind cut to the rational line (the set of rational numbers $\mathbb{Q}$), we get one of the following two possible situations:

1. the cut is done "on a number", that is there is a specific number on one of the ends of the cut. E.g. we cut $\mathbb{Q}$ into $A = (-\infty, 1)$ and $B = [1, +\infty)$, that is A is all numbers to the left of 1, while B is all numbers to the right of 1, plus the number 1 itself (because we must have $A \cup B = \mathbb{Q}$). Fig. 2.6 illustrates.

2. the cut is done "between numbers", that is there is no specific number on either end of the cut. E.g. we cut $\mathbb{Q}$ into $A = (-\infty, \sqrt{2})$ and $B = (\sqrt{2}, +\infty)$. Notice that $A \cup B = \mathbb{Q}$, that is we didn't "drop" any number during the cut. Fig. 2.7 illustrates.

▷ We can formally specify the second situation without involving real numbers e.g. by defining B as the set of positive rational numbers q such that $q^2 > 2$ and defining A as the set of rational numbers complementary to B.

We can imagine a Dedekind cut of $\mathbb{Q}$ as if the cutter either cuts "between the numbers", in which case we already obtain $\mathbb{Q} = A \cup B$, or it cuts "right on some

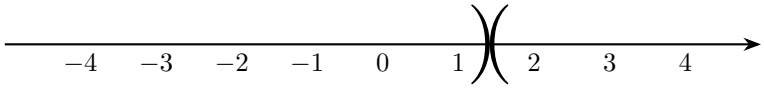


Figure 2.7: A Dedekind cut of the rational line at $x_0 = \sqrt{2}$. The number x_0 is not a part of the rational line and cannot be included into the left or into the right interval.

number x_0", so we first obtain $\mathbb{Q} = A \cup \{x_0\} \cup B$. But since the requirement is to have $A \cup B = \mathbb{Q}$, we "glue" the number x_0 into one of the sets.

Now it is postulated that with real numbers cutting "between the numbers" is impossible: we always get a number on one of the ends of the cut. E.g. $A = (-\infty, \sqrt{2})$ and $B = [\sqrt{2}, +\infty)$. It can be shown that this postulate is equivalent to the least bound axiom and thereby also to the convergence of Cauchy sequences. Respectively, the real numbers can be defined as numbers "on which" such cuts are occurring.

2.6 Continuity of operations

So far we have taken the infinite decimal fraction (2.7a) for the square root of 2 for granted. However, suppose we want to verify if this fraction is indeed the square root of 2. Then, we'd have to square this fraction and see if it yields 2 in return. However, how do we compute a square of an infinite decimal fraction? Since we implicitly understand each infinite decimal fraction as a sequence of its truncated versions, apparently we shall just take this sequence, square each of its terms and see if the resulting sequence converges to 2:

$$\begin{aligned} 1.4^2 &= 1.96 \\ 1.41^2 &= 1.9881 \\ 1.414^2 &= 1.999396 \\ 1.4142^2 &= 1.99996164 \\ &\dots \end{aligned} \tag{2.19}$$

However, in doing so, we have made an implicit assumption.

Let $f(x)$ denote the squaring operation: $f(x) = x^2$. What we have done in (2.19) is taking the sequence (2.7b) and applying $f(x)$ to each of its terms, thereby obtaining the sequence of $b_n = f(a_n)$.

$$b_1 = 1.96,\ b_2 = 1.9881,\ b_3 = 1.999396,\ b_4 = 1.99996164,\ \dots$$

Then we have claimed that by verifying $\lim b_n = 2$ we also verify $\lim a_n = \sqrt{2}$. That is we implicitly assumed that

$$\lim_{n\to\infty} a_n = \sqrt{2} \iff \lim_{n\to\infty} b_n = 2$$

and if we think of a more detailed justification for why we assumed that, it would be

$$\lim_{n\to\infty} a_n = \sqrt{2} \iff \lim_{n\to\infty} b_n = \lim_{n\to\infty} f(a_n) = f\left(\lim_{n\to\infty} a_n\right) = f(\sqrt{2}) = 2$$

Notice the equality $\lim f(a_n) = f(\lim a_n)$ in the middle. This is the well-known continuity condition for the function f. That is, we implicitly assumed the continuity of $f(x)$! Specifically, in our current situation, we assumed the continuity of $f(x) = x^2$.

Now, let's think how would we generally define the arithmetic operations on real numbers. If real numbers themselves are defined as limits of sequences, we would need to employ the limits in the definition of the arithmetic operations on them. E.g. suppose we want to add two real numbers a and b. Knowing those numbers implies that there are sequences formed by their infinite decimal fractions converging to these numbers

$$a = \lim_{n\to\infty} a_n \qquad b_n = \lim_{n\to\infty} b_n$$

▷ In principle a_n and b_n do not have to correspond to decimal fraction representations of a and b, they can be arbitrary sequences converging to a and b. But since we define real numbers as decimal fraction limits, it would make sense to focus on sequences implied by decimal fraction representations.

It would be natural to define their sum $a + b$ as

(definition of the sum of two real numbers)

$$a + b = \lim_{n\to\infty} (a_n + b_n) \tag{2.20}$$

thereby obtaining the infinite decimal fraction for $a + b$ by summing truncated infinite decimal fractions of a and b of progressively larger and larger lengths.

E.g.

$$\begin{aligned}
\sqrt{2} &= \lim \{\ 1.4,\ 1.41,\ 1.414,\ 1.4142,\ \dots\ \} \\
\sqrt{3} &= \lim \{\ 1.7,\ 1.73,\ 1.732,\ 1.7320,\ \dots\ \} \\
\sqrt{2} + \sqrt{3} &= \lim \{\ 3.1,\ 3.14,\ 3.146,\ 3.1462,\ \dots\ \}
\end{aligned}$$

▷ Actually, if we continue a bit further, we will notice that we don't exactly obtain the infinite decimal fraction for $\sqrt{2}+\sqrt{3}$ digit by digit, sometimes the previous digit can change:

$$\begin{aligned}
\sqrt{2} &= \lim \{\ \dots,\ 1.414213,\ 1.4142135,\ \dots\ \} \\
\sqrt{3} &= \lim \{\ \dots,\ 1.732050,\ 1.7320508,\ \dots\ \} \\
\sqrt{2} + \sqrt{3} &= \lim \{\ \dots,\ 3.146263,\ 3.1462643,\ \dots\ \}
\end{aligned}$$

Therefore *we need to adjust our approach* of treating real numbers as limits of sequences of truncated versions of one and the same infinite decimal fraction, to limits of sequences of some finite decimal fractions, or, even more generally, *to limits of sequences of rational numbers.*

However the definition (2.20) would be only justified if $f(x, y) = x + y$ is a continuous function, because what (2.20) effectively reads is

$$f\left(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n\right) = \lim_{n\to\infty} f(a_n, b_n)$$

But how can we establish the continuity of $f(x, y) = x + y$ before we have even defined the addition operation? *It will have to be postulated.* As a justification for this postulate, however, we could prove the continuity of $f(x, y) = x + y$ on the set of rational numbers. And then we could postulate that this continuity property extends to the real numbers.

▷ In typical math texts, the idea of function limit and the closely related idea of function continuity are introduced on the set of real numbers. However in this book we have been already applying those ideas on the set of rational numbers. This is not wrong at all, and actually can be found in more advanced math texts. The dense order property of rational numbers (rational numbers being indefinitely close to their "neighboring" rational numbers) provides a sufficient precondition for being able to speak of the function limit.

▷ Often, if not typically, in calculus courses (2.20) would be a theorem rather than a definition, and thereby would have a proof. This proof would however be based on some axiomatically introduced definitions and properties, which effectively state the continuity of addition. Thereby the continuity of addition is still going to be postulated, just implicitly, in an obscured fashion.

The same considerations that we just applied to addition by (2.20) would be obviously applicable to all other arithmetic operations. Thereby real numbers are not only closely related to the idea of the limit, but also to the idea of continuity.

▷ The exponentiation operation would be particularly interesting in this regard, with respect to the value 0^0. It is easily noticed that $f(x, y) = x^y$ is not continuous at $(0, 0)$, even if restricted to $x \geq 0$, $y \geq 0$. More precisely speaking, by choosing the value that $f(x, y)$ is defined to take at $x = y = 0$, it can be made continuous in one of the two arguments, but is cannot be made continuous in both simultaneously: there are limits $\lim_{x\to+0} f(x, 0) = 1$ and $\lim_{y\to+0} f(0, y) = 0$, but one can already see that there is no limit $\lim_{x,y\to+0} f(x, y)$. Furthermore, by controlling the specific way how x and y simultaneously approach zero, one can make the limit of $f(x, y)$ be equal to pretty much any desired value. Thus 0^0 is undefined, if understood as a real power of a real value.

On the other hand, if we restrict to integer powers and consider the function $f(x, n) = x^n$, it doesn't make sense to speak of continuity in the second, integer, argument, and thereby it doesn't make sense to speak of simultaneous continuity in both arguments either. At the same time we could notice that $\lim_{x\to+0} x^0 = 0$ and thereby we could define $x^0 = 0$ for $x = 0$ by extension by continuity. This definition is useful e.g. for writing polynomials as $\sum a_n x^n$ where the term $a_0 x^0$ thereby reduces to simply a_0 at all values of x, including $x = 0$.

2.7 Comparisons

Until now we haven't actually defined a formal way to compare real numbers even for equality. It was fine, since we never actually even tried to compare the numbers yet. E.g. in (2.20) it might look as if we compare the two real numbers in the left- and right-hand sides respectively, however (2.20) is just defining an algorithm to compute the left-hand side by using the right-hand side, rather than comparing two real numbers with the purpose of establishing

whether they are equal. However we obviously need a way to compare two real numbers for equality as well as for one being less than the other, so let's try to figure out something.

Goal 2.4. Find a way to compare two real numbers for equality. *Hint:* real numbers are effectively defined as limits of sequences.

We have already come to the conclusion that infinitely close real numbers must be treated as equal. This is because we are effectively treating real numbers as limits (or in some other way equivalent to using limits) and limits cannot really distinguish between infinitely close numbers. E.g. the terms of the sequences $a_n = \frac{1}{n}$ and $b_n = \frac{2}{n}$ are getting infinitely close, but never equal to each other. However we agree that the limit of each of these sequences is zero and thereby both limits are equal to each other:

$$\lim_{n\to\infty} \frac{1}{n} = \lim_{n\to\infty} \frac{2}{n} = 0$$

▷ A more general manifestation of the same issue is the well-known theorem

$$a_n < b_n \implies \lim_{n\to\infty} a_n \le \lim_{n\to\infty} b_n \tag{2.21}$$

that is strict inequality of terms translates into nonstrict inequality of the limits.

Since we cannot distinguish between limits of sequences which are getting infinitely close to each other, it would be natural to come up with the idea that unequal numbers should be defined by sequences that are not getting infinitely close, and respectively equal numbers should be defined by sequences that are getting infinitely close:

$$a = b \overset{\Delta}{\iff} \lim_{n\to\infty} |a_n - b_n| = 0$$

or simply

$$a = b \overset{\Delta}{\iff} \lim_{n\to\infty} (a_n - b_n) = 0 \tag{2.22}$$

But

$$\lim_{n\to\infty} (a_n - b_n) = \lim_{n\to\infty} a_n - \lim_{n\to\infty} b_n$$

and thereby (2.22) simply means

$$a = b \overset{\Delta}{\iff} a - b = 0 \tag{2.23}$$

which is totally logical.

However did we reach anything by (2.22) and (2.23)? Actually, we did, because instead of comparing two real numbers we are now comparing a real number to zero, which is a rational number. Furthermore, since we define real numbers as limits of rational sequences, we just need to check whether the rational sequence $c_n = a_n - b_n$ (where a_n and b_n are rational sequences converging to a and b) converges to zero. In principle this can be done by using the definition (2.13a), so we already have a formal answer.

Goal 2.4 completed. ■

Goal 2.5. Find a way to do a "three-way comparison" of two real numbers a and b, with the three possible outcomes $a < b$, $a = b$ and $a > b$. *Hint:* by getting

into the details of the earlier designed approach to comparing two real numbers for equality, develop a way to do a "three-way comparison" between $a - b$ and 0.

We have found a way to compare two real numbers for equality, which ultimately formally boils down to finding out whether $c_n = a_n - b_n$ converges to zero, however our job can be further simplified by noticing that $\{c_n\}$ is already known to converge, we just don't know whether its limit c is zero or not.

Our intuitive idea would be, since the set of rational numbers is *densely ordered* everywhere (that is rational numbers are getting indefinitely close to each other), there would be a rational number c' between zero and c in case $c \neq 0$. However, formally this idea leads to a circular argument because instead of comparing c to zero we would need to compare c to c'. Nevertheless, the idea can work, if instead of comparing the real number c to $c' \neq 0$, we would be comparing the rational numbers c_n to the rational number c'. Indeed, if $c \neq 0$, then, starting with some n, all c_n will become either larger than c' (if $c > 0$) or smaller than c' (if $c < 0$) and thereby we not only addressed the question of whether $c = 0$ or not, but also whether c is greater or less than zero.

Thus, for a rational sequence $c_n \to c$,

(three-way comparison of a real number c to zero)

- if there is a rational number $c' > 0$ and a natural number N such that $c_n > c' > 0 \ \forall n \geq N$ then $c > 0$,
- if there is a rational number $c' < 0$ and a natural number N such that $c_n < c' < 0 \ \forall n \geq N$ then $c < 0$,
- otherwise the condition of (2.13a) is satisfied and $c = 0$.

Apparently, the same idea works for a comparison of a real number to any rational number.

▷ The condition $\exists b' \colon a_n > b' > b$ could be referred to as "a_n being uniformly greater than b", which thereby ensures that $\lim a_n > b$ (whereas a_n being simply greater than b could result in $\lim a_n \geq b$). The term "uniformly greater/less" is not truly standard, but describes the respective situations in a concise and intuitive manner.

So now that we are able to compare real numbers to zero, we can generalize the idea of (2.23) and postulate

(three-way comparison of real numbers)

$$a = b \overset{\Delta}{\iff} a - b = 0$$
$$a < b \overset{\Delta}{\iff} a - b < 0$$
$$a > b \overset{\Delta}{\iff} a - b > 0$$

▷ An alternative approach could be to use the previously mentioned possibility to compare a real number a to a rational number c by requiring the terms a_n of the respective sequence to be uniformly greater or less that c. This yields

$$a < b \overset{\Delta}{\iff} \exists c \in \mathbb{Q} \colon a < c < b \quad (\text{where } a, b \in \mathbb{R})$$

Yet another possibility is to insert two rational numbers between a and b defined by their respective sequences a_n and b_n:

$$a < b \overset{\Delta}{\iff} \Big(\exists c_1, c_2 \in \mathbb{Q},\ N \in \mathbb{N}\colon a_n \le c_1 < c_2 \le b_n\ \forall n \ge N\Big)$$

This avoids the problem created by (2.21), and respectively we can also use nonstrict inequalities $a_n \le c_1$ and $c_2 \le b_n$ (although it doesn't really matter).

Goal 2.5 completed. ■

2.8 Uncountability

It is a well-known fact that "there are more real numbers than rational numbers". In this section we shall look into details of that statement and some of its implications.

The idea of the "size of a set" is usually formalized as the concept of the "cardinality of a set", which, given a set S, is denoted as $|S|$. The cardinality of a finite set is defined simply as the count of elements of that set (thereby cardinality is a natural number), however this stops working one we reach sets with infinite counts of elements. In particular there arises a problem of comparing the sizes of infinite sets, as usually two infinities would be either defined to compare equal (although that would be a questionable definition in most contexts) or the comparison of infinities would be undefined. Neither approach provides "a sufficient comparison resolution" for infinite sets.

Therefore, instead of first "converting the size of a set into a natural number" and then comparing two natural numbers, there is an approach to directly compare the set sizes. The approach is based on trying to establish a 1:1 mapping between the elements of two sets A and B:

- if such mapping is successfully established and there are no unmapped elements left in either set, we define $|A| = |B|$
- if upon such an attempt there are unmapped elements left in A, we define $|A| > |B|$
- if upon such an attempt there are unmapped elements left in B, we define $|A| < |B|$.

Fig. 2.8 illustrates.

With this approach in mind, let's analyze the cardinalities of various number sets, such as $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$.

2.8.1 Countable sets

Starting with the set of natural numbers $\mathbb{N}$, we are going to take its cardinality as a reference value. We already mentioned that simply writing $|\mathbb{N}| = \infty$ is not going to provide "enough resolution", therefore we introduce a special symbol $\aleph_0$ (*aleph-null*) for the cardinality of $\mathbb{N}$:

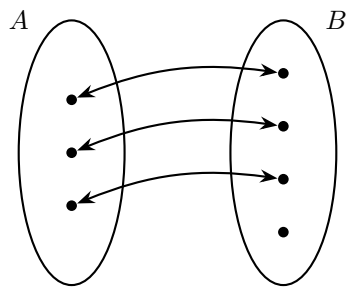


Figure 2.8: Comparing the cardinalities of two sets. The set B has an unmapped element, therefore $|A| < |B|$.

($\aleph_0$ is defined as the cardinality of $\mathbb{N}$)

$$\aleph_0 \triangleq |\mathbb{N}|$$

Now, the size of the set of integer numbers $\mathbb{Z}$ intuitively should be

$$|\mathbb{Z}| = 2 \cdot |\mathbb{N}| + 1 = 2\aleph_0 + 1$$

since for each natural number there is now its negative counterpart, plus there is also the zero. The above relationship is not wrong per se, but there's more to it. Actually

$$\begin{cases} \aleph_0 + n = \aleph_0 \\ n \cdot \aleph_0 = \aleph_0 \end{cases} \qquad \forall n \in \mathbb{N}$$

therefore

(cardinality of $\mathbb{Z}$ is the same as cardinality of $\mathbb{N}$)

$$|\mathbb{Z}| = 2\aleph_0 + 1 = \aleph_0 + 1 = \aleph_0$$

▷ To demonstrate that $\aleph_0 + n = \aleph_0$ consider the set $\mathbb{N}$ and the set A_n of integers which are strictly greater than $-n$, that is $A_n = \{m \in \mathbb{Z} \colon m > -n\}$. E.g. consider the set $A_2 = \{-1,\ 0,\ 1,\ 2,\ \ldots\}$. We claim that $|A_2| = |\mathbb{N}|$. Indeed, there is a 1:1 mapping $-1 \leftrightarrow 1$, $0 \leftrightarrow 2$, $1 \leftrightarrow 3$, $\ldots$, $m \leftrightarrow (m+n)$, $\ldots$ between A_2 and $\mathbb{N}$, thereby the sets have equal cardinalities.

Similarly, to demonstrate that $2\aleph_0 = \aleph_0$ consider the set $\mathbb{N}$ and the set $\mathbb{Z} \setminus \{0\}$ of all nonzero integers. There is a mapping $1 \leftrightarrow 1$, $-1 \leftrightarrow 2$, $2 \leftrightarrow 3$, $-2 \leftrightarrow 4$, $\ldots$ between $\mathbb{Z} \setminus \{0\}$ and $\mathbb{N}$, thereby the sets have equal cardinalities as well. By a similar argument, $n \cdot \aleph_0 = \aleph_0$.

The 1:1 mapping between some set A and the set of natural numbers $\mathbb{N}$ can be conveniently represented by arranging the elements of A into a sequence $\{a_n\}$. Thereby each natural number n is mapped to the respective term a_n of the sequence, and conversely each term a_n of the sequence is mapped to its own index n. Thus any set, whose elements can be arranged into a sequence, has cardinality $\aleph_0$. Such sets are referred to as *countable*.

The next topic will be the size of the set of rational numbers $\mathbb{Q}$. Intuitively

$$|\mathbb{Q}| = |\mathbb{Z}| \cdot |\mathbb{N}| = \aleph_0 \cdot \aleph_0 = \aleph_0^2$$

since we are having integer numerators against natural denominators. Although, maybe the above assessment is a bit too large, since it doesn't account for reducible fractions.

Let's first consider a general Cartesian product $A \times B$ of two countable sets A and B, whose elements thereby can be arranged into the sequences a_1, a_2, a_3, ... and b_1, b_2, b_3, ... Let's arrange the elements of $A \times B$ into a table:

(a_1, b_1)	(a_1, b_2)	(a_1, b_3)	...
(a_2, b_1)	(a_2, b_2)	(a_2, b_3)	...
(a_3, b_1)	(a_3, b_2)	(a_3, b_3)	...
$\vdots$	$\vdots$	$\vdots$	$\ddots$

Now, this table can be arranged into a sequence. First we take the element (a_1, b_1). Then we take all elements with indices of a_n and b_n not exceeding 2, except the (a_1, b_1) element which we have already taken. Then we take all elements with indices of a_n and b_n not exceeding 3, except those already taken. These element groups are shown by the following table:

1	2	3	...
2	2	3	...
3	3	3	...
$\vdots$	$\vdots$	$\vdots$	$\ddots$

That is, first we take all elements from group 1, then all elements from group 2, then all elements from group 3, etc., thereby arranging all elements of $A \times B$ into a sequence, thereby

$$\aleph_0^2 = \aleph_0 \cdot \aleph_0 = \aleph_0$$

By induction we obtain

$$\aleph_0^n = \aleph_0 \quad \forall n \in \mathbb{N}$$

We therefore have

$$|\mathbb{Q}| \leq |\mathbb{Z}| \cdot |\mathbb{N}| = \aleph_0^2 = \aleph_0$$

but also

$$|\mathbb{Q}| \geq |\mathbb{N}| = \aleph_0$$

since $\mathbb{N} \subseteq \mathbb{Q}$. It follows that

(cardinality of $\mathbb{Q}$ is the same as cardinality of $\mathbb{N}$)

$$|\mathbb{Q}| = \aleph_0$$

▷ Readers who are not 100% convinced by the above argument, may simply arrange the elements of the Cartesian product $\mathbb{Z} \times \mathbb{N}$ into a sequence and then remove all duplicates created by reducible fractions.

2.8.2 Cantor's diagonal argument

We are now going to show that $|\mathbb{R}| > \aleph_0$. We shall start by considering the open interval $S = (0, 1)$ which is a subset of $\mathbb{R}$. We claim that S is not countable and we shall demonstrate it by assuming the opposite.

Indeed, suppose S is countable. Then the elements of S can be arranged into a sequence $s_1, s_2, s_3, \ldots$. Let's write those elements one below the other as infinite decimal fractions, s_1 written in the first line, s_2 written in the second line, etc.:

$$\begin{aligned} s_1 &= 0.s_1^1 s_1^2 s_1^3 s_1^4 s_1^5 \ldots \\ s_2 &= 0.s_2^1 s_2^2 s_2^3 s_2^4 s_2^5 \ldots \\ s_3 &= 0.s_3^1 s_3^2 s_3^3 s_3^4 s_3^5 \ldots \\ &\ldots \end{aligned}$$

where s_n^m is the m-th digit of the fractional part of s_n. Since $s_n < 1\ \forall n$, the integer part of the infinite decimal fractions is zero. Also fractional parts of s_n shall contain at least one non-zero digit each (since otherwise s_n would be equal to zero, while, by construction, $0 \notin S$).

We can imagine the digits s_n^m making up a table:

s_1^1	s_1^2	s_1^3	s_1^4	s_1^5	$\ldots$
s_2^1	s_2^2	s_2^3	s_2^4	s_2^5	$\ldots$
s_3^1	s_3^2	s_3^3	s_3^4	s_3^5	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We are now going to construct another number x belonging to the same set S, which is not listed in the table. The integer part of x is going to be zero. The first fractional digit of x, which we denote as x^1, can be chosen arbitrarily, as long as it is different from each of the digits s_1^1, 0 and 9 (usually any digit different from s_1^1 would do, but we also want to avoid running into aliased decimal fraction pairs like $0.4999(9)\ldots$ and 0.5, which look different but denote one and the same number, therefore we forbid the digits 0 and 9). The second digit x^2 is taken to be different from s_2^2, 0 and 9. The third digit x^3 is taken different from s_3^3, 0 and 9, and so on. We thereby construct the number $x = 0.x^1x^2x^3x^4x^5\ldots$, which is different from each of the numbers s_n, since it differs from each of those numbers by at least one digit (and since we avoided aliased fractions). Apparently $x > 0$, since it doesn't consist of all zeros, and $x < 1$, since its integer part is zero and the remaining digits do not even contain any 9's, let alone being all equal to 9. Therefore $x \in S$. We thus ran into a contradiction with our original assumption that the table contains all numbers from the range S. This proof is known as *Cantor's diagonal argument* (since we constructed s by looking at the diagonal elements $s_1^1, s_2^2, s_3^3, \ldots$ of the table).

We thereby have shown that $|S| \neq \aleph_0$. On the other hand, there are subsets of S whose cardinality is $\aleph_0$. E.g. we can take a set of all numbers of the form $1/(n+1)$, where $n \in \mathbb{N}$. Therefore $|S|$ cannot be less than $\aleph_0$ and it must follow that $|S| > \aleph_0$. Such sets with a larger than countable cardinality (more precisely, the sets with the same cardinality as S) are referred to as *uncountable*. Thus, S is an uncountable set.

If we limited the length of decimal fractions in S to 1 digit after the decimal point, apparently we would have had only 10 possible different values 0.0, 0.1, 0.2, ..., 0.9. If we limited to 2 digits we would have had 100 different values 0.00, 0.01, 0.02, ..., 0.99. Limiting to n digits we obtain 10^n values. It follows that without limiting the number of digits, which we could describe as "limiting to $\aleph_0$" digits, we have $10^{\aleph_0}$ different values.[2] Thus $|S| = 10^{\aleph_0} > \aleph_0$.

> ▷ We didn't have to use base-10 fractions for Cantor's diagonal argument. We could have taken any other base. E.g. we could have used octal (base 8) fractions, or fractions base 4. Smaller bases would do as well, although there we would have additional complexities in avoiding aliased fractions. Thus $|S| = n^{\aleph_0} > \aleph_0$ where n is any natural number larger than 1.

Goal 2.6. Using the fact that the set S is uncountable, show that the set of all real numbers is uncountable as well: $|\mathbb{R}| = |S|$. *Hint:* build a 1:1 mapping between $\mathbb{R}$ and S.

The equality of cardinalities $|\mathbb{R}| = |S|$ can be shown by establishing a 1:1 mapping between $\mathbb{R}$ and S, which can be achieved by any invertible function mapping $(0,1)$ to $(-\infty, +\infty)$, e.g.

$$f(x) = \frac{1}{1-x} - \frac{1}{x} = 2\frac{x - \frac{1}{2}}{x(1-x)} \tag{2.24}$$

We have therefore established that $|\mathbb{R}| = |S| = n^{\aleph_0}$ where n is any natural number larger than 1. This cardinality value is denoted as $\mathfrak{c}$, thus

(cardinality of $\mathbb{R}$ is larger than cardinality of $\mathbb{N}$)

$$|\mathbb{R}| = \mathfrak{c} = n^{\aleph_0} > \aleph_0 \quad (n \in \mathbb{N},\ n > 1)$$

Goal 2.6 completed. ■

2.8.3 Uncountability implications

We have established that the set of real numbers is uncountable, while all the previous number sets from the natural numbers to the rational numbers were countable. But does it even matter? After all, all these sets have infinitely many numbers, is there any value in distinguishing between these infinities? It turns out that this distinction can lead to some interesting insights and probably the largest one of those is that the absolute majority of real numbers will not be ever "referred to". And by "ever" we mean "ever".

[2] We ignored the aliased fractions in this assessment, but intuitively they shouldn't affect the "final count". At any rate, there is only a countable number of aliased fractions, since all of those have finite decimal fraction representations and thereby correspond to rational numbers.

This is not the case with countable number sets. E.g. let's take natural numbers and imagine there is a person who is writing them down one after the other. It will take that person infinitely long time to write them all down, however any given natural number will be written down after a finite amount of time! E.g. say the person writes down one number in one second. Then the number 1000 will be written down after 1000 seconds, the number 1000000 will be written down after 1000000 seconds, etc. No matter how big any specific number is, it will be written down sooner or later. This "writing down" is one example of "referring to a number".

Apparently there is a similar situation with any countable number set. But with uncountable sets this is simply not the case. A person would be writing down numbers one after another, therefore the written down numbers must be a part of some countable subset of the uncountable set. All remaining numbers in that uncountable set won't be written down *ever*. It doesn't make a difference if we put many persons on the same job, the numbers written by them will still make a countable set, since $\aleph_0 \cdot n = \aleph_0$. It also won't help if we keep adding more and more persons with time, since, because $(\aleph_0)^2 = \aleph_0$, they will still write down only elements of a countable subset.

The event of "writing down" a number in the previous argument can be replaced with a more generic event of "referring to" a number, where "referring to" describes any action involving that specific number (writing down, reading in some text, saying it out loud, writing a computer program to generate that number, etc.) No matter what these events are, there is only a countable number of those.

▷ The example of writing a computer program to generate the number (e.g. by generating its decimal digits one after the other, where the program would need to run for an infinite time, but the program's text itself is supposed to be finite) could possibly the most generic description of "referring to a number". For that reason we might describe the numbers which can be "referred to" as *computable*. Since the number of different finite-length programs is countable, so is the number of computable numbers.

It seems, however, that there are also numbers which can be described, but cannot be generated by a computer program. This somewhat compromises the usage of the concept of computability as the most generic form of "being referenceable", but we won't get into details of that. Still, the number of possible finite-length descriptions (which use finite or at most countable alphabet) is countable, therefore our argument shouldn't be affected by that.

Let's now take a step back and recall how we arrived at the idea of real numbers. We just wanted some uniform notation to write down the values of irrational roots of 2nd and higher powers as well as of further expressions arising of their combinations with other numbers. Now the set of all possible roots of natural powers taken from natural numbers $\{\sqrt[m]{n}\}$ is countable, apparently its cardinality is $(\aleph_0)^2 = \aleph_0$, pretty much like for the set of all rational numbers $\frac{m}{n}$. It won't get any different after we extend the roots to allow rational arguments and powers. Or after we start computing logarithms. It doesn't take long to realize that, in fact, the set of all expressions that we are going to ever (even potentially) write down is also countable.

Therefore the set of numbers that we need because we want to compute square roots, logarithms, or whatever other operations we might need in the future, is still *countable*. Once again: the set of numbers that we are going to

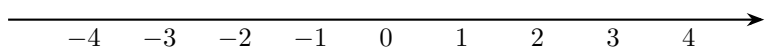
ever need, in order to represent the value of any expression we will potentially write or compute, is countable! We just don't need all those real numbers!

One might argue that once we start dealing with sequences and other infinite sets we arrive at uncountable sets of values. This is not exactly true. The sequence limit operation (or equivalently, the infinite decimal fractions) is delivering an uncountable set of potential results because the number of all possible sequences is uncountable. But we won't ever be able to "use" the majority of all those sequences, exactly because their total number is uncountable. We can't just "write down" an infinite sequence (or an infinite decimal fraction), because it takes infinite time. We either just write down a finite subpart of it, or we have to use some smart ways to deal with infinities. E.g. we write down a periodic decimal fraction by specifying the indefinitely repeating part of it: 0.1728(571428). Even though the fraction is infinitely long, we found a way to write it by a finite notation. In comparison, we can't fully write down the decimal fraction for $\sqrt{2}$ in the same fashion, it has no repeating part, but specifying that the square of that fraction must be equal to 2 (that's exactly what the notation $\sqrt{2}$ does) is sufficient to identify that number. A more complicated expression $\log_2 10 + \sqrt{5}$ also uniquely and completely identifies some infinite decimal fraction. Or we invent a special symbol such as π to denote yet another irrational number, etc. However the set of all such numbers is still countable!

In that sense, real numbers, however convenient (because they potentially cover any number we might need) are simultaneously a huge "overcommitment". Their uncountability can also complicate matters and lead to confusion in certain advanced situations. However it's just not easy (if possible at all) to come up with a smaller (countable) number set, which can be handled with the same amount of regularity and universal rules as the real numbers. Besides, as we shall see in Section 2.9, there is a different argument suggesting that, in a sense, all these "redundant" real numbers actually might exist "for real".

2.9 Continuity of the real line

We have discussed that real numbers are closely related to the the idea of function continuity, e.g. we had to postulate continuity of the arithmetic operations associated with real numbers. In this section we are going to talk about continuity of the real line itself (the line representing the real numbers):



Let's take a deeper look at the intervals on the real line. Actually, first let's talk about intervals on the "integer line", consisting only of integers. As there doesn't seem to be a commonly accepted notation for that, we are going to use the same notation as for real intervals. We are therefore going to introduce

- closed intervals: $[a, b] = \{x \in \mathbb{Z} \colon a \leq x \leq b\}$
- open intervals: $(a, b) = \{x \in \mathbb{Z} \colon a < x < b\}$
- semi-open intervals: $[a, b) = \{x \in \mathbb{Z} \colon a \leq x < b\}$ and $(a, b] = \{x \in \mathbb{Z} \colon a < x \leq b\}$

where $a, b \in \mathbb{Z}$.

Intuitively, we consider each integer number as if it was a "unit-wide pixel", that is $|[a, a)| = 1\ \forall a \in \mathbb{Z}$, where $|\cdot|$ denotes the length of the interval. It follows that the length of an integer interval is equal to the count of different numbers contained in that interval. Thereby the length of a (finite) integer interval is equal to the cardinality of the respective set. Thus, the length of a semi-open interval $[a, b)$ is simply $|[a, b)| = b - a$. The same holds for $|(a, b]|$, while the length of (a, b) is smaller by 1 and the length of $[a, b]$ is larger by 1.

We can also notice that the interval length of semi-open intervals of one and the same kind is thereby *additive*, that is the lengths of "connected intervals" add: $|[a, c)| = |[a, b)| + |[b, c)|$ for $a < b < c$ (this is actually what we can use to infer $|[a, b)| = b - a$ from $|[a, a)| = 1$). In comparison, the same doesn't hold e.g. for closed intervals, where we have $|[a, c]| = |[a, b]| + |[b, c]| - 1$, or for open intervals, where we have $|(a, c)| = |(a, b)| + |(b, c)| + 1$.

Following the same approach we will introduce real intervals

- closed intervals: $[a, b] = \{x \in \mathbb{R} \colon a \leq x \leq b\}$
- open intervals: $(a, b) = \{x \in \mathbb{R} \colon a < x < b\}$
- semi-open intervals: $[a, b) = \{x \in \mathbb{N} \colon a \leq x < b\}$ and $(a, b] = \{x \in \mathbb{R} \colon a < x \leq b\}$

where $a, b \in \mathbb{R}$. We will also define the length of a semi-open interval $[a, b)$ as $|[a, b)| = b - a$. Apparently, the length defined in such fashion is still additive, the lengths of "connected" intervals add: $|[a, c)| = |[a, b)| + |[b, c)|$ for $a < b < c$.

Now, intuitively, the lengths of the open interval (a, b) and the closed interval $[a, b]$ differ from the length of $[a, b)$ by an "infinitely small amount". However, starting with our interpretation of infinite decimal fractions, we have agreed that numbers differing by an "infinitely small amount" are equal. Therefore $|(a, b)| = |[a, b]| = b - a$.

Consider the closed interval $[a, a]$ on the real axis. Intuitively, this interval has an "infinitely small length". We will denote this "infinitely small length" as $+0$.[3] Then, since $[a, a]$ consists of a single point a, it follows that the "length" of a single point on the real line is $+0$.

Now consider an interval $[a, b]$ (where $a < b$). On one hand it does have a (non-zero) length $b - a$. On the other hand, it consists of an infinite number of points $x \in [a, b]$, and all these points have "infinitely small" lengths $+0$. As these points form a "connected chain of infinitely short intervals", by the additivity of length, the sum of all those infinitely small lengths must equal the finite amount $b - a$, kind of $\infty \cdot (+0) = b - a$. Here we can make an important observation: this sum cannot come from rational points alone, because the "total sum of their lengths" is zero.

Goal 2.7. Demonstrate that the "sum of the lengths" of rational points contained in the interval $[a, b]$ is zero. *Hint:* show that this sum is smaller than any finite number.

[3] This "+0" here is not a "full-featured number", but more a handwavy notation symbol for our intuitive discussion. Actually, the +0 notation is often used in calculus where it doesn't denote a proper number in the usual sense at all (it is not the same as 0).

The set of rational numbers contained in $[a, b]$ is countable. Let's arrange them into a sequence r_1, r_2, r_3, ... The "length" of a single point on a real line is smaller than any finite (positive) real number. Thus, the "length" of r_1 is definitely smaller than $\frac{1}{2}$. The length of r_2 is smaller than $\frac{1}{4}$. The length of r_n is smaller than $\frac{1}{2^n}$. It follows that the sum of the lengths of all r_n doesn't exceed

$$\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots = 1$$

(where one can use the formula for the sum of infinite geometric series, or it also would be sufficient for our purposes here to simply show that any partial sum of that series never exceeds 1, and thereby neither does the limit of those partial sums). That is, the sum of "lengths" of all r_n doesn't exceed 1. However instead of $\frac{1}{2^n}$ we could take $\frac{\lambda}{2^n}$ where $\lambda > 0$ is arbitrarily small positive number. It follows that the sum of "lengths" of r_n is smaller than λ. That is this sum is smaller than any positive number and thereby is equal to zero.[4]

In comparison, the same argument cannot be held for an uncountable set, such as the set of irrational numbers contained on $[a, b]$.

Goal 2.7 completed. ■

Thus, the contribution of rational numbers to the entire length $b - a$ of the interval $[a, b]$ is infinitely small, which just means "zero" (a fancy way of saying this is "the set of rational numbers contained within the interval $[a, b]$ has *measure zero*"). This means that the nonzero length of the interval $[a, b]$ must be due to contributions from some other numbers. However we are not aware of any other numbers on that interval besides other real numbers. Furthermore, Dedekind cuts suggest that there are no other numbers: no matter where we cut the line, we always cut it at some real number. It follows that the entire contribution to the nonzero length of the interval $[a, b]$ is due to the uncountable set of irrational numbers. This can be interpreted as irrational numbers existing "for real", and not just being some unnecessarily huge number set as our discussion in Section 2.8 might have suggested.

Thus, if we consider just a rational line, it is going to have "holes" at the irrational points, while the real line doesn't have holes and is therefore "continuous".

Summary of Chapter 2

- We have offered an intuitive definition of real numbers as the values denoted by infinite decimal fractions.
- By considering infinite decimal fractions as limits of sequences formed by truncated versions of these fractions, we had to look into questions of sequence convergence. We found out that our approach to define real numbers implicitly postulates convergence of Cauchy sequences. Following our approach, we had to define the arithmetic operations on real numbers as sequence limits, and thereby we had to postulate the continuity of arithmetic operations (as they were continuous on the set of rational numbers already anyway).

[4] Effectively, we have shown that the *Lebesgue measure* of the set of rational numbers contained in $[a, b]$ is zero.

- We had to resolve the difficulty in comparison of real numbers arising out of the limit-based approach and the property $a_n < b_n \implies \lim a_n \leq \lim b_n$.
- We looked into uncountability of real numbers and its implications and considered the arguments in favor and against dealing with reals as an uncountable set.

Chapter 3

Linearity everywhere

Defining linear spaces based on addition only (without scalar multiplication). Defining linearity based on addition only. Application of linearity to the analysis of the dot product and the rotation transformations. Application of linear algebra concepts for the analysis of nonlinear functions.

3.1 Linear spaces 63
3.2 Linear functions 69
3.3 Dot product 72
3.4 Rotations 74
3.4.1 Two dimensions 74
3.4.2 Multiple dimensions 76
3.4.3 Orthogonal transformations 77
3.4.4 Orthogonal transformations vs. rotations 79
3.5 Small scale picture 81
3.5.1 One-dimensional zoom-in 81
3.5.2 Multiple dimensions 84
3.6 Linearity and complex numbers 86
Summary of Chapter 3 91

In this chapter we shall review the foundational ideas of linear spaces and linearity, where we are going to try to get rid of the requirement to define scalar multiplication for the linear spaces as well as of the homogeneity requirement as a part of the linearity property. Then we shall proceed to the application examples of linearity, where we shall look at dot products, rotations and nonlinear functions.

3.1 Linear spaces

The idea of a linear space can be seen as a generalization of the Euclidean space $\mathbb{R}^N$. This generalization emerges by abstracting from defining the elements of the space as strings of real numbers and instead specifying just a number of abstract properties (axioms). These axioms are essentially describing addition of the linear space elements as well as their multiplication by a number.

This approach works efficiently in respect to keeping the set of axioms tidy, but somewhat lacks in intuitiveness. There is one operation too much: intuitively it should be sufficient just to specify the addition, and this is the idea that we shall investigate here.

In our discussion about linear spaces and the properties of their elements we are going to be drawing a strong analogy to the derivation of various number classes, which we therefore shall review now.

One can start just with the set of natural numbers $\mathbb{N}$. This set is closed under addition, multiplication and exponentiation,[1] but is not closed under the inverse operations: subtraction, division as well as taking the root and logarithm. By making, or at least by attempting to make, the set of numbers closed under the respective operations, the set $\mathbb{N}$ is extended to $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$:

- The set of integers $\mathbb{Z}$ is obtained by making the set $\mathbb{N}$ closed under subtraction. E.g. given two natural numbers 2 and 3, their difference is not in $\mathbb{N}$. However, given any two integer numbers m and n, their difference $m - n$ always exists and belongs to $\mathbb{Z}$.

- The set of rational numbers $\mathbb{Q}$ is obtained by making the set $\mathbb{Z}$ closed under division. Well, almost closed. There is one exception: division by zero. Given two integer numbers m and n, their ratio doesn't necessarily belong to $\mathbb{Z}$. However, given two rational numbers $p, q \in \mathbb{Q}$, their ratio p/q exists and belongs to $\mathbb{Q}$, with a single exception $q = 0$.

- The set of real numbers $\mathbb{R}$ is obtained by an attempt to make the set $\mathbb{Q}$ closed under the operations of taking the root and logarithm. This attempt is only half-way successful, e.g. we still cannot compute $\sqrt{-1}$. Also, the result that we settle with is that, instead of making the set of numbers closed under taking the root and logarithm, we rather make it closed under the operation of taking the limit of an infinite sequence.[2]

- The set of complex numbers $\mathbb{C}$ is obtained by completing our previous attempt (originally made with $\mathbb{R}$) to make the set of numbers closed under the operations of taking the root and logarithm.

As promised, we shall drop the idea of including multiplication into our original set of axioms about linear space. Instead we just start with the idea that our linear space V consists of elements (referred to as vectors) which can be added together. That is, formally, given $\vec{x}, \vec{y} \in V$, there exists another vector $\vec{z} = \vec{x} + \vec{y}$, and it exists not just "somewhere" but in exactly the same linear space, that is $\vec{z} \in V$. We also postulate that the vector addition is commutative:

$$\vec{x} + \vec{y} = \vec{y} + \vec{x} \quad \forall \vec{x}, \vec{y} \tag{3.1}$$

and associative

$$(\vec{x} + \vec{y}) + \vec{z} = \vec{x} + (\vec{y} + \vec{z}) \quad \forall \vec{x}, \vec{y}, \vec{z} \tag{3.2}$$

Let's see where does that get us.

[1] This is a fancy way of saying that it should be possible to apply the given operation to any numbers. E.g. given $m, n \in \mathbb{N}$, their sum $m + n$ is guaranteed to also belong to $\mathbb{N}$.

[2] More details on this are discussed in Chapter 2.

▷ In this book, *as a rule*, we will be using arrow notation $\vec{x}$ to highlight the geometric aspects of a vector as an entity. Boldface notation $\mathbf{x}$ will be more commonly used to denote a particular column matrix of vector coordinates.

Goal 3.1. Derive multiplication of a vector by a number from addition of two vectors. *Hint:* follow (approximately) the way how the set $\mathbb{N}$ was extected to $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$.

Goal 3.1a. Derive multiplication of a vector by a *natural* number from addition of two vectors. *Hint:* follow (approximately) the way how multiplication of two natural numbers is derived from their addition.

Definition of multiplication of a vector by a natural number,[3] can be done exactly in the same fashion how multiplication of one natural number by another is defined. E.g. given $m, n \in \mathbb{N}$, their product nm is defined as

$$nm = \underbrace{m + m + \ldots + m}_{n \text{ terms}}$$

Doing the same for vectors we obtain

$$n\vec{x} = \underbrace{\vec{x} + \vec{x} + \ldots + \vec{x}}_{n \text{ terms}} \tag{3.3}$$

One immediately obtains that the product (3.3) is distributive with respect to multiplication in each operand:

$$\begin{aligned} n(\vec{x} + \vec{y}) &= \underbrace{(\vec{x} + \vec{y}) + (\vec{x} + \vec{y}) + \ldots + (\vec{x} + \vec{y})}_{n \text{ terms}} \\ &= \underbrace{\vec{x} + \vec{x} + \ldots + \vec{x}}_{n \text{ terms}} + \underbrace{\vec{y} + \vec{y} + \ldots + \vec{y}}_{n \text{ terms}} = n\vec{x} + n\vec{y} \qquad (3.4a) \\ (n + m)\vec{x} &= \underbrace{\vec{x} + \vec{x} + \ldots + \vec{x}}_{n + m \text{ terms}} \\ &= \underbrace{\vec{x} + \vec{x} + \ldots + \vec{x}}_{n \text{ terms}} + \underbrace{\vec{x} + \vec{x} + \ldots + \vec{x}}_{m \text{ terms}} = n\vec{x} + m\vec{x} \qquad (3.4b) \end{aligned}$$

Goal 3.1a completed. ■

Goal 3.1b. Extend the definition of multiplication of a vector by a natural number to accommodate integers. *Hint:* introduce subtraction of vectors.

We have postulated that, given $\vec{x}, \vec{y} \in V$, there exists their sum $\vec{x} + \vec{y} \in V$ but this says nothing about the difference of two vectors. There is no problem defining the difference as addition's inverse: $\vec{x} - \vec{y}$ denotes a vector $\vec{z}$ such that $\vec{z} + \vec{y} = \vec{x}$:

$$\vec{z} = \vec{x} - \vec{y} \overset{\Delta}{\iff} \vec{z} + \vec{y} = \vec{x} \tag{3.5}$$

but there is no way we can be sure that such vector $\vec{z}$ exists in the first place.

This is very similar to promotion from natural numbers to integers. Originally, being familiar just with natural numbers and given two natural numbers

[3] In some conventions zero is considered to be a natural number, while in the others it isn't. For the purposes of the current discussion, $0 \notin \mathbb{N}$.

m and n, we could not be sure if their difference $m - n$ exists or not. By introducing integer numbers we have postulated that any two natural numbers m and n can be subtracted from each other and, furthermore, any two integer numbers can be subtracted from each other.

We now postulate the same thing for vectors: any two vectors $\vec{x}, \vec{y} \in V$ can be subtracted from each other, producing another vector $\vec{z} = \vec{x} - \vec{y}$ belonging to the same linear space (that is, $\vec{z} \in V$), such that $\vec{z} + \vec{y} = \vec{x}$.

> ▷ Notice that for natural numbers postulating that subtraction is closed only extends the set of natural numbers "in the opposite, negative direction". With vectors it doesn't explicitly extend the set per se, but it ensures that the set "extends equally in all directions". How many of these directions are there, obviously, depends on the dimensionality of the space.

Vector subtraction allows us to introduce the zero vector by first noticing that

$$\vec{x} - \vec{x} = \vec{y} - \vec{y} \quad \forall \vec{x}, \vec{y} \tag{3.6}$$

Indeed, by adding $\vec{x} + \vec{y}$ to both sides of that equality, the left-hand side is turned into

$$(\vec{x} - \vec{x}) + (\vec{x} + \vec{y}) \overset{(3.2)}{=} ((\vec{x} - \vec{x}) + \vec{x}) + \vec{y} \overset{(3.5)}{=} \vec{x} + \vec{y}$$

while the right-hand side is turned into

$$\begin{aligned}(\vec{y} - \vec{y}) + (\vec{x} + \vec{y}) &\overset{(3.1)}{=} (\vec{y} - \vec{y}) + (\vec{y} + \vec{x}) \overset{(3.2)}{=} ((\vec{y} - \vec{y}) + \vec{y}) + \vec{x} \\ &\overset{(3.5)}{=} \vec{y} + \vec{x} \overset{(3.1)}{=} \vec{x} + \vec{y}\end{aligned}$$

The property (3.6) allows to introduce the zero vector

$$0 = \vec{x} - \vec{x} \quad \forall \vec{x} \tag{3.7}$$

> ▷ Often the zero vector is denoted as $\vec{0}$ rather than simply 0, but mostly it should be clear from the context, whether we are talking of the zero vector or of the number zero, and in the cases where it isn't (e.g. $0\vec{x}$ in principle can be understood as a dot product $\vec{0} \cdot \vec{x}$) the distinction usually doesn't matter.

We also can derive the distributivity over subtraction from distributivity over addition, obtaining

$$n(\vec{x} - \vec{y}) = n\vec{x} - n\vec{y} \tag{3.8a}$$

from (3.4a) and

$$(n - m)\vec{x} = n\vec{x} - m\vec{y} \tag{3.8b}$$

from (3.4b), however with (3.8b) we have a problem if $n \leq m$, since we only can multiply vectors by natural numbers so far. Thus (3.8b) effectively provides motivation to define multiplication of a vector by zero as

$$0\vec{x} \triangleq 0 \quad \forall \vec{x} \tag{3.9}$$

so that

$$(n - n)\vec{x} \overset{(3.8b)}{=} n\vec{x} - n\vec{x} \overset{(3.7)}{=} 0$$

and by a negative number as

$$(-n)\vec{x} \stackrel{\Delta}{=} 0 - n\vec{x} \tag{3.10}$$

so that

$$n\vec{x} + (-n)\vec{x} \stackrel{(3.4b)}{=} (n + (-n))\vec{x} = 0\vec{x} \stackrel{(3.9)}{=} 0$$

Further properties of vector subtraction can be derived from those of vector addition.

Goal 3.1b completed. ■

Goal 3.1c. Extend the definition of multiplication of a vector by an integer number to accommodate rational numbers. *Hint:* introduce division of a vector by a number as an inverse operation to multiplication by a number.

Division of vector by a number can be defined as inversion of multiplication by a number:

$$\vec{y} = \vec{x}/\lambda \stackrel{\Delta}{\Longleftrightarrow} \vec{x} = \lambda\vec{y} \tag{3.11}$$

however nothing guarantees that such $\vec{y}$ will exist for given $\vec{x}$ and λ in the first place.

This is similar to promotion from integers to rational numbers. Originally, being familiar only with integer numbers and given two integer numbers m and n, we could not be sure if their ratio m/n exists or not. By introducing rational numbers, we have postulated that any two integer numbers m and n can be divided (except dividing by zero) and, furthermore, any two rational numbers can be divided.

We now postulate a similar thing for vectors. Actually, first we shall postulate it just for integer divisors: any vector $\vec{x}$ can be divided by any nonzero integer number. That is given any $\vec{x} \in V$ and any $n \in \mathbb{Z} \setminus \{0\}$, the expression $\vec{x}/n$ produces another vector $y = \vec{x}/n$ belonging to the same linear space (that is, $\vec{y} \in V$), such that $n\vec{y} = \vec{x}$.

This postulate allows us to just compute multiplication of a vector by a rational number as

$$\frac{m}{n}\vec{x} = \frac{m\vec{x}}{n} \tag{3.12}$$

where we now not only are able to compute $m\vec{x}$ for any $m \in \mathbb{Z}$, but also to further divide $m\vec{x}$ by any n.

Respectively, by (3.11) and (3.12), the division by a rational number is computed as

$$\vec{x} \Big/ \left(\frac{m}{n}\right) = \frac{n\vec{x}}{m} \tag{3.13}$$

so that

$$\frac{m}{n} \cdot \left(\vec{x} \Big/ \left(\frac{m}{n}\right)\right) \stackrel{(3.13)}{=} \frac{m}{n} \cdot \frac{n\vec{x}}{m} \stackrel{(3.12)}{=} \frac{m\frac{n\vec{x}}{m}}{n} \stackrel{(3.11)}{=} \frac{n\vec{x}}{n} \stackrel{(3.11)}{=} \vec{x}$$

thereby being compatible to (3.11).

The properties of division can be obtained from those of multiplication.

Goal 3.1c completed. ■

Goal 3.1d. Extend the definition of multiplication of a vector by a rational number to accommodate real numbers. *Hint:* introduce vector sequences and their convergence. This may require introduction of some vector norm.

We introduced the possibility to multiply vectors by rational numbers, but we still need to be able to multiply by real numbers. The real numbers are effectively defined as limits of rational number sequences[4], therefore we probably would need to do a similar thing for vectors: given $\vec{x} \in V$ and $\lambda \in \mathbb{R}$, we construct some rational number sequence $\lambda_n \to \lambda$ and define

$$\lambda\vec{x} \triangleq \lim_{n\to\infty} \lambda_n\vec{x} \tag{3.14}$$

However, there is a problem in (3.14): we don't know how to compute limits of sequences of vectors.

Now recall from our discussion in Chapter 2 that the existence of limits of sequences of rational numbers converging to real numbers cannot be really proven, it needs to be postulated directly (by stating the convergence of Cauchy sequences) or indirectly (by some other equivalent postulate). Either way we are effectively postulating the existence of real numbers, pretty much like we were postulating the existence of rational and integer numbers earlier by claiming that the set of numbers is closed under division or subtraction. In a similar way, we are going to postulate the convergence of Cauchy sequences of vectors.

In order to be able to tell whether a sequence of vectors is a Cauchy sequence, one needs to have some kind of metric in the linear space, or at least the idea of neighborhoods. In common formal treatment of linear space these ideas are usually separated from the pure idea of a linear space. However in our intuitive derivation it should be acceptable to e.g. postulate the idea of the length of a vector. After all, in geometry irrational numbers commonly appear from attempts to compute lengths, such as application of Pythagorean theorem, computation of trigonometric functions (as ratios of two lengths) or computation of the length of a circle.

Having introduced the idea of the length (or any kind of norm) of a vector, we can use it to check whether a particular sequence of vectors $\vec{x}_n$ is a Cauchy sequence: the distances $|\vec{x}_{n+m} - \vec{x}_n|$ should be bounded by ε for all $m \in \mathbb{N}$.

Having postulated the convergence of Cauchy sequences, we can employ (3.14) as the definition of the product of a vector with a real number. The division of a vector by a real number λ can be simply defined as multiplication by $1/\lambda$.

> ▷ By introducing vector length and Cauchy sequence convergence, we effectively made our linear space more specific by turning it into a Banach space. That can affect the overall structure of formal statements, but for the purposes of intuitive understanding we probably mostly shouldn't mind.

Goal 3.1d completed. ■

Multiplication by complex numbers will be dealt with in Section 3.6. This is where we are going to complete Goal 3.1.

[4] See the discussion in Chapter 2.

3.2 Linear functions

There are two different meanings of the term "linear functions":

1. functions defined as polynomial expressions with no terms of orders higher than 1, which generally means that they are having the form

$$f(x) = ax + b \tag{3.15}$$

2. functions satisfying the *linearity property*

$$f(\lambda x + \mu y) = \lambda f(x) + \mu f(y) \tag{3.16}$$

In this section we'll be interested in the latter.

> ▷ To avoid confusion with linear functions in the sense of (3.16), we could in principle refer to functions (3.15) as *affine* (although this might be a somewhat nonstandard terminology), since they define affine transformations.

The linearity property (3.16) is actually an even harder restriction than (3.15), because it results in functions of the form

$$f(x) = ax \tag{3.17}$$

which seem pretty boring. However normally the concept of linearity is used for functions whose arguments are not just numbers but vectors. Also their values are often vectors as well. In the latter case the constant a in (3.17) turns into a matrix and the function f turns into a linear transformation of the linear space of vectors. We thereby shall not restrict the argument x of the function $f(x)$ in (3.17) only to numbers, but rather consider it to be some kind of vector, and the same shall be done about the value of the function.

> ▷ We are not going to explicitly notate the argument of f as a vector, because it could be notated at least in two ways: $\vec{x}$ or $\mathbf{x}$, the former notation assuming some kind of basis-independent vector, while the latter might be explicitly suggesting the coordinate-based interpretation of a vector. Similarly, we are not going to replace the notation of the function $f(\cdot)$ with $\vec{f}(\cdot)$ or $\mathbf{f}(\cdot)$.
>
> The vector argument and the vector value of the function can also be more generic than just that. E.g. functions can be treated as vectors, and thereby can be supplied as the argument of the function f, in which case the function f becomes an operator (if it returns a function) or a functional (if it returns a number). Thus, we are talking of a huge set of various linear entities, all formally represented by the function f in (3.17) and (3.16).

The linearity property (3.16) can be equivalently written as two separate properties: additivity

$$f(x + y) = f(x) + f(y) \tag{3.18a}$$

and homogeneity

$$f(\lambda x) = \lambda f(x) \tag{3.18b}$$

Similarly to how we tried to define linear spaces based solely on addition in Section 3.1, we shall now try to define linearity based solely on the additivity property (3.18a).

Goal 3.2. Research the possibility to define linearity based solely on additivity.

Goal 3.2a. Derive homogeneity (3.18b) from additivity (3.18a) for multiplication by a natural λ. *Hint:* express multiplication by a natural number in terms of addition.

Assuming the function $f(x)$ satisfies the additivity property (3.18a) we can simply infer the homogeneity for the case of multiplication by natural numbers. Indeed,

$$f(nx) = f(\underbrace{x + x + \ldots + x}_{n \text{ terms}}) \overset{(3.18a)}{=} \underbrace{f(x) + f(x) + \ldots + f(x)}_{n \text{ terms}} = nf(x) \quad \forall n \in \mathbb{N} \tag{3.19}$$

Goal 3.2a completed. ■

Goal 3.2b. Derive homogeneity from additivity for multiplication by an integer λ. *Hint:* show that $f(0) = 0$ for an additive f, and then use subtraction.

We can further extend the homogeneity to include all integer numbers by still just using the additivity property (3.18a). In Section 3.1 to introduce multiplication by integers we first had to introduce the zero element, which was possible due to (3.6). The counterpart of that step in case of linearity property is to notice that

$$f(x) = f(0 + x) \overset{(3.18a)}{=} f(0) + f(x)$$

thereby $f(0) = f(x) - f(x) = 0$. That is we obtain the property

$$f(0) = 0 \tag{3.20}$$

which also implies

$$f(0x) = 0 \cdot f(x) \tag{3.21}$$

Combining (3.20) with (3.18a) we obtain

$$\begin{aligned} f(-nx) = f(0 - nx) &\overset{(3.18a)}{=} f(0) - f(nx) \overset{(3.20)}{=} 0 - f(nx) = -f(nx) \\ &\overset{(3.19)}{=} -nf(x) \end{aligned} \tag{3.22}$$

Combining (3.19), (3.21) and (3.22) we extend (3.19) to $n \in \mathbb{Z}$:

$$f(nx) = nf(x) \quad \forall n \in \mathbb{Z} \tag{3.23}$$

> ▷ Notice that, differently from Section 3.1, this time we didn't need to postulate the existence of the result. We already did that for the entire linear space in Section 3.1 and, assuming f is defined on the entire linear space, it thereby automatically applies to f.

Goal 3.2b completed. ■

Goal 3.2c. Derive homogeneity from additivity for multiplication by a rational λ. *Hint:* use division.

We continue extending the homogeneity to now include all rational numbers, still just from the additivity property (3.18a). By (3.23) we notice that

$$f(x) = f(n \cdot (x/n)) \overset{(3.23)}{=} nf(x/n)$$

thereby

$$f(x/n) = f(x)/n \quad \forall n \in \mathbb{Z} \setminus \{0\} \tag{3.24}$$

It further follows that

$$f\left(\frac{m}{n}x\right) = f(mx/n) = f((mx)/n) \overset{(3.24)}{=} f(mx)/n \overset{(3.23)}{=} \frac{mf(x)}{n} = \frac{m}{n}f(x)$$

Thereby

$$f\left(\frac{m}{n}x\right) = \frac{m}{n}f(x) \tag{3.25}$$

and it's not difficult to further obtain

$$f(x/\lambda) = f(\lambda^{-1}x) \overset{(3.25)}{=} \lambda^{-1}f(x) = f(x)/\lambda \quad \forall \lambda \in \mathbb{Q} \tag{3.26}$$

Goal 3.2c completed. ■

Goal 3.2d. Derive homogeneity from additivity for multiplication by a real λ. *Hint:* assume continuity of f.

In Section 3.1 to define multiplication by real numbers we represented the real numbers as limits of sequences of rational numbers in order to derive the multiplication of vectors by real numbers. We shall try the same approach here:

$$\begin{aligned} f(\lambda x) &= f\left(\lim_{n\to\infty} \lambda_n \cdot x\right) = f\left(\lim_{n\to\infty} \lambda_n x\right) = \lim_{n\to\infty} f(\lambda_n x) \\ &\overset{(3.25)}{=} \lim_{n\to\infty} \lambda_n f(x) = f(x) \lim_{n\to\infty} \lambda_n = \lambda f(x) \quad \text{where } \lambda_n \to \lambda \end{aligned}$$

However in doing so we switched the order of computing the limit and evaluating $f(\cdot)$ from $f(\lim \lambda_n x)$ to $\lim f(\lambda_n x)$. This is possible only if $f(\cdot)$ is continuous at λx and this is what we should require.

▷ This continuity requirement is the counterpart of postulating the convergence of Cauchy sequences of vectors that we did in Section 3.1.

Thus, additionally assuming the continuity of f (which is quite a reasonable assumption, holding in typical cases), we obtained

$$f(\lambda x) = \lambda f(x) \quad \forall \lambda \in \mathbb{R} \tag{3.27}$$

▷ It suffices to show or postulate the continuity at a single point. Continuity at all other points follows by additivity. E.g. given $f(x)$ continuous at $x = 0$, at any other point x_1 we obtain

$$\begin{aligned} \lim_{x\to x_1} f(x) &= \lim_{x\to 0} f(x + x_1) = \lim_{x\to 0}\big(f(x) + f(x_1)\big) = f(x_1) + \lim_{x\to 0} f(x) \\ &= f(x_1) + f(0) = f(x_1 + 0) = f(x_1) \end{aligned}$$

Goal 3.2d completed. ■

Homogeneity with respect to multiplication by complex numbers will be addressed in Section 3.6. This is where we are going to complete Goal 3.2.

3.3 Dot product

Linear functions can be defined by specifying their values just at a small number of points. In one dimension (a function from $\mathbb{R}$ to $\mathbb{R}$) it suffices to specify the function's value at a single nonzero point $x_0 \neq 0$. The values at all other points can be obtained by $f(x) = f(\frac{x}{x_0}x_0) = \frac{x}{x_0}f(x_0)$. In two dimensions (a function from $\mathbb{R}^2$ to $\mathbb{R}$) we would need to specify two different function values at two linearly independent argument values. E.g. we could take two basis vectors and express $f(\vec{x})$ as

$$f(\vec{x}) = f(x_1\vec{e}_1 + x_2\vec{e}_2) = x_1 f(\vec{e}_1) + x_2 f(\vec{e}_2)$$

that is it suffices to specify the values $f(\vec{e}_1)$ and $f(\vec{e}_2)$ and the function will be fully defined.

The dot product of two vectors $\vec{x}\cdot\vec{y}$ is a bilinear operation, that is it is linear in each argument:

$$\begin{aligned} \vec{x}\cdot(\vec{y}_1 + \vec{y}_2) &= \vec{x}\cdot\vec{y}_1 + \vec{x}\cdot\vec{y}_2 \\ (\vec{x}_1 + \vec{x}_2)\cdot\vec{y} &= \vec{x}_1\cdot\vec{y} + \vec{x}_2\cdot\vec{y} \end{aligned} \tag{3.28}$$

where we skipped the homogeneity aspect (3.18b) according to our discussion in Section 3.2. The property (3.28) alone (or, more precisely, together with continuity, in order to accommodate real numbers) should allow us to derive a coordinate-based expression for the dot product in a very simple manner and this is what we are going to do in this section, however before that let's justify the statement that the dot product is bilinear.

We shall start from the "geometric" definition of the dot product $\vec{x}\cdot\vec{y}$ of the two vectors $\vec{x}$ and $\vec{y}$ as the product of the vector lengths by the cosine of the angle between them:

$$\vec{x}\cdot\vec{y} = |\vec{x}|\cdot|\vec{y}|\cdot\cos\theta \tag{3.29}$$

▷ The main motivation for the definition (3.29) is probably coming from the need to compute projections. Although, for that specific purpose we actually do not need exactly the dot product $\vec{x}\cdot\vec{y}$, but rather the expression $\frac{\vec{x}\cdot\vec{y}}{|\vec{y}|}$ (unless we force one of the vectors to have unit length). However, the dot product is the bilinear part of that expression and it is convenient to consider it as a separate entity.

Some further motivations might come from physics, where e.g. the work of a force is computed as $|\vec{F}|\cdot|\Delta\vec{r}|\cdot\cos\theta$.

Goal 3.3. Given the geometric definition (3.29) of the dot product, demonstrate additivity in both arguments (3.28).

We will now use the projection interpretation of the dot product (3.29) to demonstrate bilinearity, where, as mentioned, we are going to demonstrate only the additivity aspect. Assume one of the vectors has unit length. In such case we can write the dot product as $\vec{x}\cdot\vec{n}$, where $\vec{n}$ is the unit vector. The expression $\vec{x}\cdot\vec{n}$ describes the projection of $\vec{x}$ onto the line $\vec{u}(t) = t\vec{n}$ where $t \in \mathbb{R}$. Such projection is apparently additive, as demonstrated by Fig. 3.1, which, in terms of the dot product, reads

$$(\vec{x} + \vec{y})\cdot\vec{n} = \vec{x}\cdot\vec{n} + \vec{y}\cdot\vec{n} \tag{3.30}$$

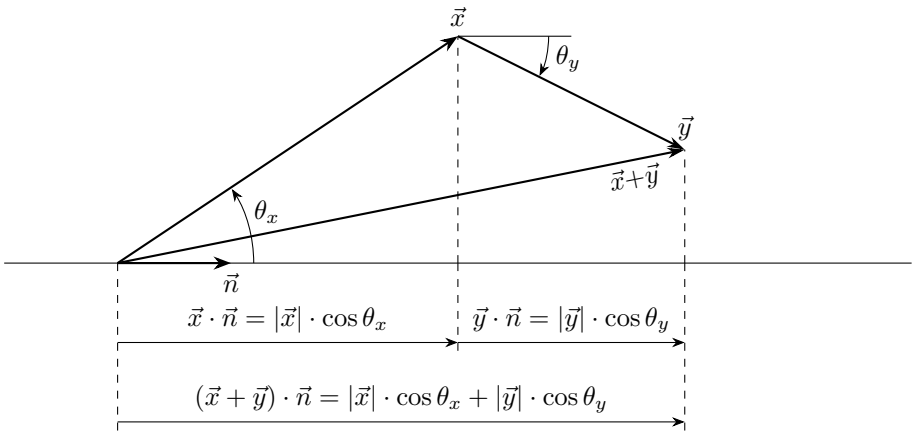


Figure 3.1: Additivity of projection.

By (3.29) the dot product is proportional to the vector length, thereby, given a vector $\vec{v}$ of a non-unit length, we can express it as $\vec{v} = |\vec{v}| \cdot \vec{n}$, where $\vec{n}$ is a unit-length vector. Multiplying (3.30) by $|\vec{v}|$ yields

$$(\vec{x}+\vec{y})\cdot\vec{v} = (\vec{x}+\vec{y})\cdot(|\vec{v}|\cdot\vec{n}) = \vec{x}\cdot(|\vec{v}|\cdot\vec{n}) + \vec{y}\cdot(|\vec{v}|\cdot\vec{n}) = \vec{x}\cdot\vec{v} + \vec{y}\cdot\vec{v}$$

(where formally we apply (3.29) to ensure that multiplication by $|\vec{v}|$ distributes over the sum). Therefore the dot product is additive in the first operand.

We can apply exactly the same argument to the second operand of the dot product, obtaining the additivity in the second operand.

$$\vec{v}\cdot(\vec{x}+\vec{y}) = \vec{v}\cdot\vec{x} + \vec{v}\cdot\vec{y}$$

(alternatively we could just use the commutativity of the dot product which follows from the definition (3.29)).

Goal 3.3 completed. ■

▷ We might also wish to demonstrate continuity, since the latter is required to infer linearity from the additivity alone. The continuity of dot product follows from the continuity of the respective factors in (3.29), and thereby dot product is linear in both arguments.

Given the bilinearity of the dot product we can simply notice that e.g. in two dimensions

$$\begin{aligned}\vec{x}\cdot\vec{y} &= (x_1\vec{e}_1 + x_2\vec{e}_2)\cdot(y_1\vec{e}_1 + y_2\vec{e}_2)\\ &\overset{(3.28)}{=} x_1y_1\cdot(\vec{e}_1\cdot\vec{e}_1) + x_2y_1\cdot(\vec{e}_2\cdot\vec{e}_1) + x_1y_2\cdot(\vec{e}_1\cdot\vec{e}_2) + x_2y_2\cdot(\vec{e}_2\cdot\vec{e}_2)\end{aligned}$$

Assuming an orthonormal basis yields

$$\begin{aligned}\vec{e}_1\cdot\vec{e}_1 &= |\vec{e}_1|\cdot|\vec{e}_1|\cdot\cos 0 = 1\cdot 1\cdot 1 = 1\\ \vec{e}_2\cdot\vec{e}_1 &= |\vec{e}_2|\cdot|\vec{e}_1|\cdot\cos 90^\circ = 1\cdot 1\cdot 0 = 0\end{aligned}$$

$$\vec{e}_1 \cdot \vec{e}_2 = |\vec{e}_1| \cdot |\vec{e}_2| \cdot \cos 90^\circ = 1 \cdot 1 \cdot 0 = 0$$
$$\vec{e}_2 \cdot \vec{e}_2 = |\vec{e}_2| \cdot |\vec{e}_2| \cdot \cos 0 = 1 \cdot 1 \cdot 1 = 1$$

and thereby

(dot product in an orthonormal 2D basis)

$$\vec{x} \cdot \vec{y} = x_1 y_1 \cdot 1 + x_2 y_1 \cdot 0 + x_1 y_2 \cdot 0 + x_2 y_2 \cdot 1 = x_1 y_1 + x_2 y_2 \tag{3.31}$$

Higher dimension counts can be addressed similarly.

▷ The formula (3.31) can be used to derive the formula for the cosine of a difference. Let

$$\vec{x} = (\cos\theta_x,\ \sin\theta_x) \qquad \vec{y} = (\cos\theta_y,\ \sin\theta_y)$$

be two unit vectors $\vec{x}$ and $\vec{y}$ expressed in terms of their components in an orthonormal basis in two dimensions. Then

$$\begin{aligned}\cos(\theta_x - \theta_y) &= |\vec{x}| \cdot |\vec{y}| \cdot \cos(\theta_x - \theta_y) = \vec{x} \cdot \vec{y} = x_1 y_1 + x_2 y_2 \\ &= \cos\theta_x \cos\theta_y + \sin\theta_x \sin\theta_y\end{aligned}$$

3.4 Rotations

We have just seen in the example of the dot product that a bilinear scalar function of two vector arguments needs to be specified at N^2 points to be fully defined (where N is the number of dimensions of the space). Similarly, a vector-valued linear function of a vector argument also needs just N^2 points. We shall demonstrate it using the rotational transformation.

3.4.1 Two dimensions

Let $\vec{x}\,' = \vec{f}(\vec{x})$ be a transformation of a 2-dimensional linear space, rotating all vectors in the space by the angle α. Similarly to our approach in Section 3.3 we shall start from geometric definitions. So we need a geometric definition of rotation. This is an interesting question per se. We could define rotation as a transformation of all vectors in the linear space which preserves all vector lengths and *signed* angles between vectors. This definition immediately implies that all dot products are preserved, but not vice versa: the preservation of dot products doesn't imply the preservation of the angle signs.

▷ The idea of a "signed angle" between two vectors makes sense only in 2 dimensions and is also something which we'll have to define axiomatically, but that could do for now. In a larger number of dimensions we will need to generalize the preservation of "signed angles" to preservation of relative orientation of a set of N vectors, where N is the number of space dimensions (the idea of the orientation of a set of vectors is strongly related to the materials of Chapter 5). This generalization may still look sufficiently simple formally, but, from the intuition perspective, may be getting cumbersome and at least require some further justifications.

A better approach in this respect might be to split the rotation by the angle α into a number of infinitely small rotations by infinitely small angles $\mathrm{d}\alpha$, all done

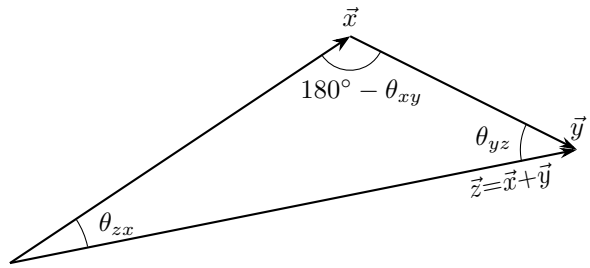


Figure 3.2: The triangle rule of vector addition and angles.

in one and the same direction. Such infinitely small rotations can be defined as infinitely small transformations of space, which preserve the vector lengths and angles (or, equivalently, the dot products). In this case we would be able to drop the requirement of preserving the orientations of vector sets, since an infinitely small transformation cannot change those orientations. We shall discuss this idea further in Section 7.10.

We now need to demonstrate that the rotational transformation $\vec{x}\,' = \vec{f}(\vec{x})$ satisfies the additivity property: if $\vec{z} = \vec{x} + \vec{y}$ before the transformation, then after the transformation $\vec{z}\,' = \vec{x}\,' + \vec{y}\,'$, that is $\vec{f}(\vec{z}) = \vec{f}(\vec{x}) + \vec{f}(\vec{y})$, or simply and shortly

$$\vec{f}(\vec{x} + \vec{y}) = \vec{f}(\vec{x}) + \vec{f}(\vec{y}) \tag{3.32}$$

Goal 3.4. Demonstrate the additivity (3.32) of a 2D rotational transformation. *Hint:* use geometric implications of preserving the lengths and angles.

Since $\vec{f}(\vec{\cdot})$ preserves all dot products, it preserves the vector lengths:

$$|\vec{x}\,'| = |\vec{x}\,| \qquad |\vec{y}\,'| = |\vec{y}\,| \qquad |\vec{z}\,'| = |\vec{z}\,|$$

and it preserves the absolute values of the angles between the vectors

$$\theta_{xy} = \theta_{x'y'} \qquad \theta_{yz} = \theta_{y'z'} \qquad \theta_{zx} = \theta_{z'x'}$$

(where θ_{xy} etc. stand for the absolute values, rather than signed angles). This means that the triangle formed by the vectors $\vec{x}$, $\vec{y}$, $\vec{z}$ (Fig. 3.2) is unchanged by the transformation and therefore the transformed vector $\vec{z}\,' = \vec{f}(\vec{z})$ is equal to the sum of vectors $\vec{x}\,'$ and $\vec{y}\,'$. Thus we have $\vec{z}\,' = \vec{x}\,' + \vec{y}\,'$ and thereby (3.32) holds.

▷ Notice that we demonstrated (3.32) solely based on the assumption of preservation of dot products. Therefore the transformation $\vec{f}(\vec{\cdot})$ is not restricted to rotations only, but can be any transformation which preserves the dot product, that is any *orthogonal transformation.*

Goal 3.4 completed. ■

▷ We have demonstrated the additivity of rotation (or, for that matter, of any orthogonal transformation). It remains to demonstrate the continuity, which simply follows from the preservation of vector lengths and angles. Indeed, small variations of the length of the vector argument of $\vec{f}(\vec{\cdot})$ produce small (actually the same)

variations of the length of the vector value of $\vec{f}(\vec{\cdot})$. Similarly, small variations of the direction of the vector argument of $\vec{f}(\vec{\cdot})$ produce small (actually of the same absolute amount) variations of the direction of the vector value of $\vec{f}(\vec{\cdot})$. Thereby $\vec{f}(\vec{\cdot})$ is continuous, and it follows that $\vec{f}(\vec{\cdot})$ is linear.

Goal 3.5. Use the linearity of rotation to derive the 2D rotation formula in the matrix form (including the values of the rotation matrix coefficients). *Hint:* use the same approach as in the derivation of the dot product formula (3.31).

Now we will use the linearity of rotation to derive the rotation formula, similarly to how we derived the dot product formula (3.31) from its bilinearity. By linearity of the rotational transformation $\vec{f}(\vec{\cdot})$,

$$\vec{f}(\vec{x}) = \vec{f}(x_1\vec{e}_1 + x_2\vec{e}_2) = x_1\vec{f}(\vec{e}_1) + x_2\vec{f}(\vec{e}_2) \tag{3.33}$$

that is we need to know the values of $\vec{f}(\vec{e}_1)$ and $\vec{f}(\vec{e}_2)$. Those are straightforward to obtain:

$$\vec{f}(\vec{e}_1) = (\cos\alpha,\ \sin\alpha)$$
$$\vec{f}(\vec{e}_2) = (-\sin\alpha,\ \cos\alpha)$$

and thus

$$\vec{x}\,' = \vec{f}(\vec{x}) = x_1 \cdot (\cos\alpha,\ \sin\alpha) + x_2 \cdot (-\sin\alpha,\ \cos\alpha)$$

or, in the matrix form,

$$\mathbf{x}' = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

or

$$\mathbf{x}' = \mathbf{R} \cdot \mathbf{x} \tag{3.34a}$$

where

(2D rotation matrix)

$$\mathbf{R} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \tag{3.34b}$$

Goal 3.5 completed. ■

3.4.2 Multiple dimensions

In two dimensions it was only possible to rotate in one and the same plane. In three dimensions we can choose the plane to rotate in, which is normally expressed as "choosing the rotation axis", where the rotation axis is simply a line perpendicular to the rotation plane.

▷ In order to be able to specify the rotation direction we actually need an *oriented plane* rather than just a plane, respectively the axis is not just a line, but a line with direction, which is often specified by a unit-length vector. The topic of plane orientation is strongly related to the materials of Chapter 5.

It is easiest to choose one of the coordinate planes as the rotation plane (and respectively one of the coordinate axes as the rotation axis), in which case we can simply temporarily ignore the third coordinate and apply the 2-dimensional rotation (3.34) to the remaining 2 coordinates. This yields e.g. for the xy coordinate plane (and z rotation axis):

(matrix of rotation in the xy plane in 3D)

$$\mathbf{R}_{xy} = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.35}$$

and rotations about the other 2 coordinate axes have similar forms.

▷ Intuitively, we should be able to construct an arbitrary rotation in 3D from the coordinate-plane rotations $\mathbf{R}_{xy}$, $\mathbf{R}_{yz}$, $\mathbf{R}_{zx}$. A common way to do this is by using the *Euler angles* (which we won't discuss in this text). The general statement about being able to express any rotation via coordinate plane rotations will be addressed in Subsection 3.4.4.

In four dimensions we proceed in a similar way. Assuming the 4 dimensions are labelled as x, y, z, w, a rotation in the coordinate plane xy is expressed by the matrix

(matrix of rotation in the xy plane in 4D)

$$\mathbf{R}_{xy} = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.36}$$

Notice that this time there is no rotation axis but a "plane about which we rotate". That is rotation *in* the xy coordinate plane as performed "*about* the zw coordinate plane".

Higher dimension counts are addressed similarly.

It is easily noticed that matrices (3.34b), (3.35) and (3.36) have a kind of "transpose symmetry". Specifically, if transposed, the component values $\sin\alpha$ and $-\sin\alpha$ are swapped. This can alternatively be seen as replacing α by $-\alpha$ in the entire matrix, since sine is odd and cosine is even, thereby transposed matrix simply describes the rotation in the opposite direction by the same angle and we have

$$\mathbf{R}^{\mathrm{T}} = \mathbf{R}^{-1} \tag{3.37}$$

3.4.3 Orthogonal transformations

Our previous results about rotations were mostly derived from the sole fact that the transformation preserves the dot product. It's only in the derivation of the specific values of the components of the matrices (3.34b), (3.35) and (3.36) that we used the extra knowledge that this is not just any dot product-preserving transformation, but specifically a rotation.

Let's therefore take a look at a more generic class of transformations which preserve the dot product but do not necessarily represent rotations. By the argument made earlier in Subsection 3.4.1, such transformations are linear, since we only used the property of dot product preservation to show the linearity. Also notice that this earlier argument is not restricted to 2 dimensions but holds in any other dimension count N. That is we can repeat the same expansion (3.33) in N dimensions obtaining

$$\vec{f}(\vec{x}) = \vec{f}\left(\sum_{n=1}^{N} x_n \vec{e}_n\right) = \sum_{n=1}^{N} x_n \vec{f}(\vec{e}_n) \tag{3.38}$$

and respectively (3.34) generalizes into

$$\mathbf{x}' = \mathbf{M} \cdot \mathbf{x} \tag{3.39a}$$

where

$$\mathbf{M} = \begin{bmatrix} f_1(\vec{e}_1) & f_1(\vec{e}_2) & \cdots & f_1(\vec{e}_N) \\ f_2(\vec{e}_1) & f_2(\vec{e}_2) & \cdots & f_2(\vec{e}_N) \\ \vdots & \vdots & \ddots & \vdots \\ f_N(\vec{e}_1) & f_N(\vec{e}_2) & \cdots & f_N(\vec{e}_N) \end{bmatrix} \tag{3.39b}$$

where $f_n(\vec{x})$ delivers the n-th component of the vector $\vec{x}\,' = \vec{f}(\vec{x})$.

Now, if $\vec{f}(\vec{\cdot})$ preserves dot products, then, in particular, it should preserve the orthonormality of the basis vectors, so if our original basis $\{\vec{e}_n\}$ is orthonormal, then

(orthonormality of columns of an orthogonal matrix)

$$\begin{aligned} \vec{f}(\vec{e}_n) \cdot \vec{f}(\vec{e}_n) &= 1 \qquad \forall n \\ \vec{f}(\vec{e}_n) \cdot \vec{f}(\vec{e}_m) &= 0 \qquad \forall m \neq n \end{aligned} \tag{3.40}$$

The dot products in (3.40) are the dot products of the columns of the matrix $\mathbf{M}$ defined by (3.39b). Thus (3.40) describes the mutual orthogonality and the "unit length" of the columns of $\mathbf{M}$. The matrix $\mathbf{M}$ is respectively referred to as *orthogonal matrix* while the transformation $\vec{f}(\vec{\cdot})$ itself is referred to as *orthogonal transformation* of the linear space.

We have established the property (3.37) for rotation matrices. It would be interesting to see what this property turns into for the wider class of orthogonal matrices. To do that let's take a more detailed look into the elements of the matrix $\mathbf{M}$ given by (3.39b). The value $f_m(\vec{e}_n)$ delivers the m-th coordinate of the vector $\vec{e}_n{}' = \vec{f}(\vec{e}_n)$. If $\{\vec{e}_n\}$ is an orthonormal basis, then these coordinates are simply given by scalar products: $f_m(\vec{e}_n) = \vec{e}_m \cdot \vec{e}_n{}'$ and thus

$$\mathbf{M} = \begin{bmatrix} \vec{e}_1 \cdot \vec{e}_1{}' & \vec{e}_1 \cdot \vec{e}_2{}' & \cdots & \vec{e}_1 \cdot \vec{e}_N{}' \\ \vec{e}_2 \cdot \vec{e}_1{}' & \vec{e}_2 \cdot \vec{e}_2{}' & \cdots & \vec{e}_2 \cdot \vec{e}_N{}' \\ \vdots & \vdots & \ddots & \vdots \\ \vec{e}_N \cdot \vec{e}_1{}' & \vec{e}_N \cdot \vec{e}_2{}' & \cdots & \vec{e}_N \cdot \vec{e}_N{}' \end{bmatrix} \tag{3.41a}$$

Respectively

$$\mathbf{M}^{\mathrm{T}} = \begin{bmatrix} \vec{e}_1{}' \cdot \vec{e}_1 & \vec{e}_1{}' \cdot \vec{e}_2 & \cdots & \vec{e}_1{}' \cdot \vec{e}_N \\ \vec{e}_2{}' \cdot \vec{e}_1 & \vec{e}_2{}' \cdot \vec{e}_2 & \cdots & \vec{e}_2{}' \cdot \vec{e}_N \\ \vdots & \vdots & \ddots & \vdots \\ \vec{e}_N{}' \cdot \vec{e}_1 & \vec{e}_N{}' \cdot \vec{e}_2 & \cdots & \vec{e}_N{}' \cdot \vec{e}_N \end{bmatrix} \tag{3.41b}$$

(where we used the commutativity of the dot product to swap the order of factors in each component).

Both matrices (3.41a) and (3.41b) have exactly the same form, except that the two sets of vectors $\{\vec{e}_n\}$ and $\{\vec{e}_n{}'\}$ are swapped. Now if $\{\vec{e}_n{}'\}$ was another orthonormal basis, then $\mathbf{M}^{\mathrm{T}}$ would have been describing another orthogonal transformation, which turns $\{\vec{e}_n{}'\}$ back into $\{\vec{e}_n\}$. But $\{\vec{e}_n{}'\}$ was obtained from $\{\vec{e}_n\}$ by the transformation $\vec{f}(\cdot)$, which preserves lengths and angles, therefore $\{\vec{e}_n{}'\}$ *is* an orthonormal set and can be used as another orthonormal basis! Therefore $\mathbf{M}^{\mathrm{T}}$ describes a transformation which is an inverse of $\mathbf{M}$ and

(formal definition of an orthogonal matrix)

$$\mathbf{M}^{\mathrm{T}} = \mathbf{M}^{-1} \quad \text{(or, equivalently, } \mathbf{M}\mathbf{M}^{-1} = \mathbf{M}^{-1}\mathbf{M} = \mathbf{I}\text{)} \tag{3.42}$$

same as we had for rotation matrices.

▷ We found that matrices describing orthogonal transformations satisfy (3.42). But does the inverse statement hold, that is does any matrix satisfying (3.42) preserve dot products? Rewriting (3.42) as $\mathbf{M}^{-1}\mathbf{M} = \mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$, we obtain that at least all basis vectors are still orthonormal after the transformation. Let $\mathbf{x} = \sum_i x_i \mathbf{e}_i$ be a vector, which after the transformation turns into

$$\mathbf{x}' = \mathbf{M} \cdot \mathbf{x} = \mathbf{M} \cdot \sum_i x_i \mathbf{e}_i = \sum_i x_i \mathbf{M}\mathbf{e}_i = \sum_i x_i \mathbf{e}_i'$$

Thus the coordinates of $\mathbf{x}$ are unchanged upon transformation (if the coordinates after the transformation are taken in the new basis), and it follows that all dot products are preserved.

Thus any matrix satisfying (3.42) describes an orthogonal transformation, which completes the justification of (3.42) as a formal definition of an orthogonal matrix.

▷ Apparently the inverse of $\vec{f}(\cdot)$ is also an orthogonal transformation, and so is the matrix $\mathbf{M}^{-1}$. By (3.42), the rows of $\mathbf{M}^{-1}$ are the columns of $\mathbf{M}$, and, vice versa, the rows of $\mathbf{M}$ are the columns of $\mathbf{M}^{-1}$. Combining this with (3.40) we obtain that the rows of $\mathbf{M}$ are also orthonormal and so are the rows of $\mathbf{M}^{-1}$. That is, both the rows and the columns of an orthogonal matrix are in an orthonormal relationship.

3.4.4 Orthogonal transformations vs. rotations

So, the class of orthogonal transformations is, at least by their definition (as the transformations preserving the dot products), wider than the class of rotational transformations. But is it really wider, or is it actually the same? That is, are there orthogonal transformations, which are not rotations?

Let's start with a randomly chosen N-dimensional orthogonal transformation, which has a corresponding matrix $\mathbf{M}$ describing that transformation. We

want to try to express this matrix as a product of rotation matrices, where each rotation matrix describes a rotation in a coordinate plane (since we at least gained some experience with those in Subsection 3.4.2). Actually it might be simpler to build this sequence of rotations for $\mathbf{M}^{-1}$, by trying to undo the effects of $\mathbf{M}$ one by one.

Goal 3.6. Attempt to undo the effect of $\mathbf{M}$ by a sequence of rotations within the coordinate planes. Derive a conclusion of whether a generic orthogonal matrix is representable by a sequence of coordinate-plane rotations, and, if not, find out the missing details.

We could perform coordinate-plane rotations to put the transformed basis vectors $\mathbf{e}'_n = \mathbf{M} \cdot \mathbf{e}_n$ back into their original position (where, as usual, we assume that the basis is orthonormal). Putting all vectors $\mathbf{e}'_n$ back to their original positions $\mathbf{e}_n$ one by one, we will construct a sequence of coordinate-plane rotations amounting to the transformation by the $\mathbf{M}^{-1}$ matrix.

So, let's start with the vector $\mathbf{e}'_1$. We can perform a series of rotations in the coordinate planes x_1x_2, x_1x_3, $\ldots$, x_1x_N (where x_1, x_2, $\ldots$, x_N are the coordinate axes). The rotation in the x_1x_2 plane will aim to zero the x_2 component, while ensuring that the x_1 component is positive. The rotation in the x_1x_3 plane will aim to zero the x_3 component, while ensuring that the x_1 component is positive. Etc. After we have done these rotations, the vector $\mathbf{e}'_1$ will have all its components x_2, x_3, $\ldots$, x_N zeroed, while its x_1 component will be equal to 1 (since $|\mathbf{e}'_1| = 1$ and the rotations didn't change the lengths of any vectors). That is we have put the $\mathbf{e}'_1$ vector back into its original position.

Now, since rotations preserve the orthogonality relationships between the basis vectors, all the remaining vectors $\mathbf{e}'_n$ for $n \geq 2$ will be orthogonal to $\mathbf{e}'_1$, therefore their x_1 coordinates are already zero. We therefore can take the vector $\mathbf{e}'_2$ (which we already partially "rotated back" during rotating back the vector $\mathbf{e}'_1$) and perform a series of rotations in the coordinate planes x_2x_3, x_2x_4, $\ldots$, x_2x_N aiming to bring $\mathbf{e}'_2$ back to its original position $\mathbf{e}_2$ (which we obviously can do). Notice that since we already "locked" the x_1 coordinate, these rotations do not destroy the already reverted position of $\mathbf{e}'_1$.

We repeat the same with $\mathbf{e}'_3$, $\mathbf{e}'_4$, etc., where we will have to stop after $\mathbf{e}'_{N-1}$, because all coordinate axes x_1, x_2, $\ldots x_{N-1}$ have been already locked and there are no more coordinate planes that we can use for rotations. At this point all vectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\ldots$, $\mathbf{e}_{N-1}$ have been reverted to their original positions. We haven't reverted $\mathbf{e}'_N$ yet, but, since it must be orthogonal to all other vectors $\mathbf{e}'_n$, there are only two options left: it either coincides with $\mathbf{e}_N$ or with $-\mathbf{e}_N$.

In the first case we have successfully undone the transformation described by $\mathbf{M}$, effectively constructing a factoring of $\mathbf{M}^{-1}$ into rotations in coordinate planes. The matrix $\mathbf{M}$ can be therefore factored into a sequence of opposite direction rotations. In the second case we have failed: there is an additional mirror transformation of the space, which cannot be undone by rotations.

▷ Apparently two mirror operations can be canceled by a rotation. E.g. inversions of the x_1 and x_2 axes in 3D can be undone by a 180° rotation about the x_3 axis. Therefore any even number of mirror operations can be canceled.

Goal 3.6 completed. ■

▷ By the previous discussion, an arbitrary orthogonal transformation is not necessarily decomposable into a sequence of coordinate plane rotations. But is an arbitrary single rotation (about a given axis or a hyperaxis) decomposable into such sequence? Apparently yes: we can first perform a series of rotations making the desired rotation plane coincide with, say, the coordinate plane x_1x_2, then rotate in that plane, and then perform the inverse series of rotations bringing the desired rotation plane into its original position.

Instead of performing this auxiliary sequence of rotations of the space, we could also rotate the basis, so that the rotation plane is coinciding with one of the coordinate planes. In that case the series of auxiliary rotations will be replaced by the basis transformation matrix (which effectively will consist of the same rotations, just done in the opposite order and directions).

The discussion of rotations will be continued in Section 7.10.

3.5 Small scale picture

We have obtained earlier in this chapter that, at least intuitively, a linear space is a space whose elements can be added together, and a linear function is a function which is additive in its argument. Still, it might seem that a large number, if not the absolute majority of the functions and expressions that we have to deal with, are nonlinear, and thereby linear algebra covers only a small subset of whatever we need to deal with.

This is not exactly true. On the contrary, linear algebra is applicable pretty much everywhere and this happens directly via Taylor series expansion. Also, contrarily to what one might think, this doesn't apply only to the "linear part" of the Taylor series expansion (that is, up to the 1st order term). No, linear algebra is actually heavily involved in the entire expansion.

3.5.1 One-dimensional zoom-in

We shall start with considering the idea of "zooming in" into a function. Specifically, imagine a graph of some nonlinear function, e.g. $f(x) = \sin x$ and let's start zooming in more and more (Fig. 3.3). As we arrive at higher zoom levels, we notice that the function turns almost into a straight line. That is, it can be well approximated as $ax + b$.

Therefore, if we take a function graph and zoom in around the point $x = 0$, then we should be getting closer and closer to the approximation

$$f(x) \approx f(0) + f'(0) \cdot x$$

Fig. 3.4 illustrates this for the exponential function $y = e^x$.

Instead of using a 1st order polynomial, we might be willing to use a 2nd order polynomial for our approximation, hoping for a better precision (or for a lower amount of the necessary "zoom"). If our "zoom" is still centered around $x = 0$, it makes sense to choose a 2nd-order polynomial, which matches all its derivatives up to the 2nd one with those of $f(x)$ (higher order derivatives of a 2nd-order polynomial are all zero and cannot be matched against those of $f(x)$). This is achieved by

$$f(x) \approx f(0) + f'(0) \cdot x + \frac{f''(0)}{2} \cdot x^2$$

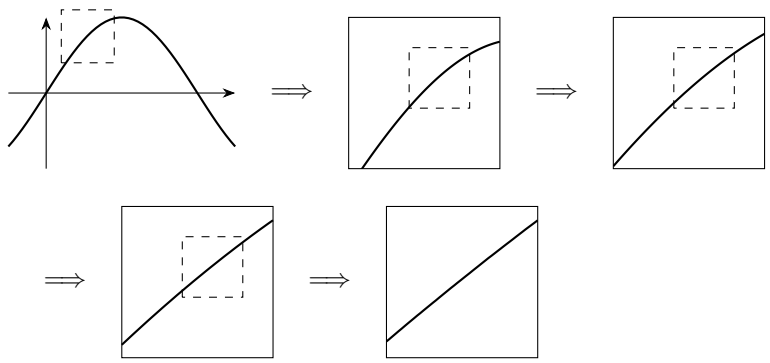


Figure 3.3: A graph becoming more and more linear upon zooming in.

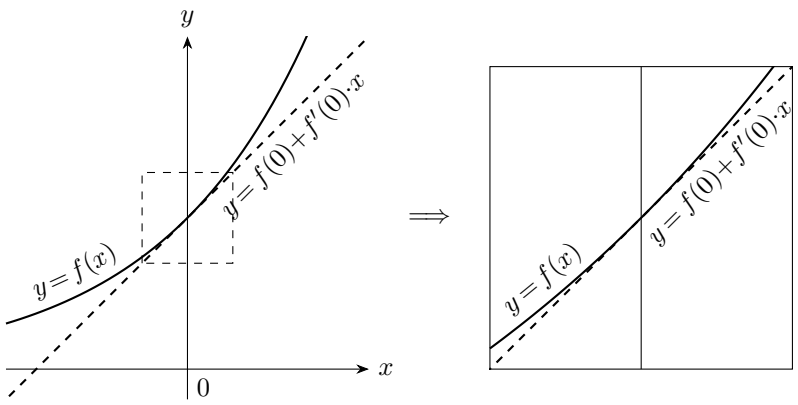


Figure 3.4: Zooming into the first order approximation of $y = e^x$.

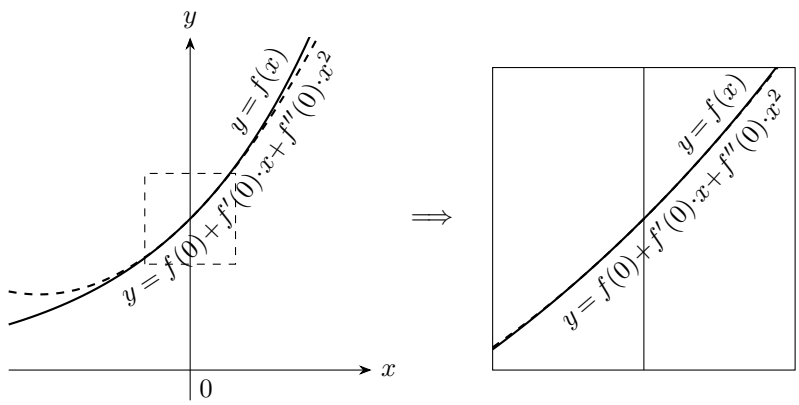


Figure 3.5: Zooming into the second order approximation of $y = e^x$.

(Fig. 3.5), which is easy to verify by differentiating the above expression.

Generally, for an N-th order polynomial, we obtain

(N-th partial sum of Taylor series expansion at $x = 0$)

$$f(x) \approx \sum_{n=0}^{N} \frac{f^{(n)}(0)}{n!} x^n \tag{3.43}$$

▷ The error of approximation by (3.43) can be assessed using the mean value theorem (discussed in Section 6.9). E.g. suppose we approximate $f(x)$ with a 0-th order polynomial:

$$f(x) \approx f(0)$$

The exact value of $f(x)$ can be written by adding another term to the above:

$$f(x) = f(0) + \int_0^x f'(t)\,\mathrm{d}t$$

The added term is exactly the error term. By the mean value theorem,

$$\int_0^x f'(t)\,\mathrm{d}t = f'(\xi) \cdot x$$

where ξ is some point on the interval between 0 and x. This allows to use the maximum of $|f'(x)|$ on that interval as the upper bound for the approximation error.

Similarly, for a 1st order approximation

$$f(x) \approx f(0) + f'(0) \cdot x$$

we also can write the exact value of $f(x)$ by adding another term:

$$\begin{aligned} f(x) &= f(0) + \int_0^x f'(t)\,\mathrm{d}t = f(0) + \int_0^x \left(f'(0) + \int_0^t f''(\tau)\,\mathrm{d}\tau \right) \mathrm{d}t \\ &= f(0) + f'(0) \cdot x + \int_0^x \int_0^t f''(\tau)\,\mathrm{d}\tau\,\mathrm{d}t \end{aligned}$$

This time the integration of $f''(\tau)$ is done in 2 dimensions over a triangle with the vertices $(0,0)$, $(x,0)$ and (x,x), thereby $f''(\tau)$ is getting averaged over this triangle. The area of the triangle is $x^2/2$ and thereby we obtain

$$\int_0^x \int_0^t f''(\tau)\,\mathrm{d}\tau\,\mathrm{d}t = f''(\xi)\frac{x^2}{2}$$

where ξ is some point on the interval between 0 and x.

For an N-th order approximation we obtain $f^{(N+1)}(\xi)\cdot x^{N+1}/(N+1)!$ as the error assessment.

▷ The mean value theorem works only "in one dimension". That is it works for a real-valued f, but not if the value of f is a complex number or a vector. However, this only affects the existence of the point ξ, where the necessary derivative value is attained. While we cannot claim the existence of the point ξ, where the expression $f^{(N+1)}(\xi)\cdot x^{N+1}/(N+1)!$ yields the exact error, we can still assess this error. Indeed, we could apply the integral form of the generalized triangle inequality $|\sum a_n| < \sum |a_n|$, and use the maximum value of $|f^{(N+1)}(\xi)|$ on the interval between 0 and x to assess the maximum possible error, as usual. E.g. for $N=2$:

$$\left|\int_0^x \int_0^t f''(\tau)\,\mathrm{d}\tau\,\mathrm{d}t\right| \le \int_0^x \int_0^t |f''(\tau)|\,\mathrm{d}\tau\,\mathrm{d}t \le \max|f''(\xi)|\cdot\frac{x^2}{2}$$

So by zooming in around $x=0$ and considering polynomial approximations of increasing orders we have arrived at Taylor series (3.43), but where is linearity? Somewhat paradoxically, this is easier to see in multiple dimensions.

3.5.2 Multiple dimensions

▷ This subsection expects familiarity with partial derivatives. Readers who are not familiar with the subject may consider checking Section 6.7 or other sources.

Let the function $f(\cdot)$, that we are approximating, be a function of a vector argument (while the value of the function is still just a number). We shall therefore write this function from now on as $f(\mathbf{x})$, suggesting that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

is a column matrix of vector coordinates (where N is the dimension of the space). The first-order Taylor approximation in this case looks as

$$f(\mathbf{x}) \approx f(0) + \sum_{n=1}^{N}\left(\frac{\partial f}{\partial x_n}(0)\cdot x_n\right) = f(0) + \mathbf{g}^{\mathrm{T}}(0)\cdot\mathbf{x} \tag{3.44}$$

where

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{bmatrix}$$

is the gradient of f. Notice that the dot product $\mathbf{g}^{\mathrm{T}}(0)\cdot\mathbf{x}$ is a linear function of $\mathbf{x}$ since

$$\mathbf{g}^{\mathrm{T}}(0)\cdot(\mathbf{x}+\mathbf{y}) = \mathbf{g}^{\mathrm{T}}(0)\cdot\mathbf{x} + \mathbf{g}^{\mathrm{T}}(0)\cdot\mathbf{y}$$

Adding the second-order term to (3.44) yields

$$\begin{aligned} f(\mathbf{x}) &\approx f(0) + \sum_{n=1}^{N}\left(\frac{\partial f}{\partial x_n}(0)\cdot x_n\right) + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\left(\frac{\partial^2 f}{\partial x_n \partial x_m}(0)\cdot x_n x_m\right) \\ &= f(0) + \mathbf{g}^{\mathrm{T}}(0)\cdot\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\cdot\mathbf{H}\cdot\mathbf{x} \end{aligned} \tag{3.45}$$

where

$$\mathbf{H}(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_N \partial x_N} \end{bmatrix}$$

is the Hessian of f. Notice that the second term $\mathbf{x}^{\mathrm{T}}\cdot\mathbf{H}\cdot\mathbf{x}$ is a *bilinear* function of $\mathbf{x}$ since

$$\mathbf{x}^{\mathrm{T}}\cdot\mathbf{H}\cdot(\mathbf{u}+\mathbf{v}) = \mathbf{x}^{\mathrm{T}}\cdot\mathbf{H}\cdot\mathbf{u} + \mathbf{x}^{\mathrm{T}}\cdot\mathbf{H}\cdot\mathbf{v}$$

and

$$(\mathbf{u}+\mathbf{v})^{\mathrm{T}}\cdot\mathbf{H}\cdot\mathbf{x} = \mathbf{u}^{\mathrm{T}}\cdot\mathbf{H}\cdot\mathbf{x} + \mathbf{v}^{\mathrm{T}}\cdot\mathbf{H}\cdot\mathbf{x}$$

The second-order term of the multidimensional second-order Taylor series expansion (3.45) is therefore a *bilinear form*, or more specifically a *quadratic form*, which is one of the standard topics of study of linear algebra.

Adding the third term to (3.45) yields

(3rd-order Tayor series expansion of a function of a vector argument at $\mathbf{x}=0$)

$$\begin{aligned} f(\mathbf{x}) \approx f(0) &+ \sum_{n=1}^{N}\left(\frac{\partial f}{\partial x_n}(0)\cdot x_n\right) + \frac{1}{2!}\sum_{n,m=1}^{N}\left(\frac{\partial^2 f}{\partial x_n \partial x_m}(0)\cdot x_n x_m\right) \\ &+ \frac{1}{3!}\sum_{n,m,l=1}^{N}\left(\frac{\partial^3 f}{\partial x_n \partial x_m \partial x_l}(0)\cdot x_n x_m x_l\right) \end{aligned} \tag{3.46}$$

The third term cannot be written in matrix notation, since a matrix has only two indices and we need three, however one can already see that the third-order term in (3.46) is trilinear.

▷ Trilinear forms might not be a standard topic in linear algebra, but it is still a particular case of multilinearity. The index notation of tensor algebra (which is, essentially, just a somewhat advanced area of basic linear algebra and is discussed in Chapter 4) might be of help here, rewriting (3.46) as

$$f(x^k) \approx f(0) + \frac{\partial f}{\partial x^n}(0)\cdot x^n + \frac{\partial f^2}{\partial x^n \partial x^m}(0)\cdot\frac{x^n x^m}{2!} + \frac{\partial f^3}{\partial x^n \partial x^m \partial x^l}(0)\cdot\frac{x^n x^m x^l}{3!}$$

and for an m-th order expansion

$$f(x^k) \approx \sum_{n=0}^{m}\frac{\partial f^n}{\partial x^{i_1}\ldots\partial x^{i_n}}(0)\cdot\frac{x^{i_1}\ldots x^{i_n}}{n!}$$

Thus

1. "At high resolution" (if we "zoom in far enough") the functions will become linear (plus a constant offset), unless we are zooming in at a breakpoint, or some other kind of a special point.

2. The terms of Taylor series expansion are all multilinear forms.

This suggests that linear algebra has quite a few more areas of application than one might originally think.

3.6 Linearity and complex numbers

In Sections 3.1 and 3.2 we progressed from multiplication of vectors by natural numbers to multiplication by real numbers, leaving the complex numbers out of scope. Now it's time to fill that gap. Let's start with linear spaces.

Complex linear spaces are commonly encountered in Fourier theory and quantum mechanics, where they originate from treating complex-valued sequences or functions as vectors. Thereby there is already an assumed coordinate system (the sequence terms and the function values being treated as vector coordinate components in that system). It would still be instructive to try to generalize the results of Sections 3.1 to complex numbers, so that we would be able to construct a complex linear space in abstract terms without an a priori given coordinate basis. Therefore we would like to define multiplication of a vector by a complex number in abstract terms, same as we did with other kinds of numbers in Section 3.1, thereby completing Goal 3.1.

Apparently it would suffice to define multiplication of a vector by the imaginary unit, multiplication by an arbitrary complex number $a + ib$ can be then obtained as $(a + ib)\vec{x} = a\vec{x} + b \cdot i\vec{x}$. Now recall that the imaginary unit is introduced as a solution to the equation $x^2 = -1$ and in order to introduce "complex vectors" we would need some kind of parallel to this equation. Our first idea could be using the dot product $(\vec{x})^2 = \vec{x} \cdot \vec{x} = -1$ to obtain a vector of an "imaginary unit length", but this is not going to work too well.

Dot product

In case of complex vector coordinates, attempting to employ the usual definition of the dot product

$$\vec{x} \cdot \vec{y} = \sum_{n=1}^{N} x_n y_n \tag{3.47}$$

(assuming an orthonormal basis) leads to awkward vector lengths, if we want to keep using $|\vec{x}| = \sqrt{\vec{x} \cdot \vec{x}}$ as the definition of vector length. In particular

1. Vector lengths now attain complex values.

2. The value of $|\vec{x}| = \sqrt{\vec{x} \cdot \vec{x}}$ is no longer uniquely defined, since we cannot take an arithmetic square root of a complex number. Even if we somehow extend the definition of arithmetic square root to complex numbers, there are different options for such extension.

3. Being essentially complex, the lengths cannot be compared for "greater" or "smaller". Although we might attempt to amend that by introducing the "absolute magnitude of a vector length" as $|\vec{x}| = |\sqrt{\vec{x} \cdot \vec{x}}|$, this idea doesn't seem to make things much better.

4. Some non-zero vectors are going to have zero lengths. E.g. the length of a 2D vector $\vec{x} = (1, i)$ is $|\vec{x}| = \sqrt{\vec{x} \cdot \vec{x}} = \sqrt{1^2 + i^2} = 0$.

5. Cauchy–Schwarz inequality is not applicable to complex lengths and it's not going to hold in many cases where the lengths and dot products are still real. E.g. for $\vec{x} = (1, i)$, $\vec{y} = (1, -i)$ we have $|\vec{x}| = |\vec{y}| = 0$ but $\vec{x} \cdot \vec{y} = 2$, thus $|\vec{x} \cdot \vec{y}| > |\vec{x}| \cdot |\vec{y}|$.

6. Triangle inequality gets broken in a similar manner. E.g. for the same $\vec{x} = (1, i)$, $\vec{y} = (1, -i)$ we have $|\vec{x}| + |\vec{y}| = 0 + 0 = 0$, but $|\vec{x} + \vec{y}| = 1$, therefore $|\vec{x}| + |\vec{y}| < |\vec{x} + \vec{y}|$.

Giving up any of those properties of the dot product and the associated vector length would be too much, so maybe we should modify the definition of the dot product instead. Especially if our modification is going to be a true generalization of the old definition (3.47), to which it then should automatically fall back in case of real numbers.

We can arrive at a reasonable generalization of (3.47) by considering a 1D case. In 1D case the length of a vector is simply equal to the absolute magnitude of its coordinate: $|\vec{x}| = |x_1|$ (assuming an "orthonormal" basis). In case of real numbers and (3.47) this works by $|\vec{x}| = \sqrt{\vec{x} \cdot \vec{x}} = \sqrt{(x_1)^2} = |x_1|$. In case of complex numbers we can recall the formula for the absolute magnitude of a complex value: $|z| = \sqrt{z^* \cdot z}$, where the asterisk denotes complex conjugation. Respectively for a 1D space with complex coordinates we should have $|\vec{x}| = \sqrt{\vec{x} \cdot \vec{x}} = \sqrt{(x_1)^* \cdot x_1} = |x_1|$. This suggests the *sesquilinear* ("one and a half-" rather than bi-linear) formula

(dot product in complex vector space)

$$\vec{x} \cdot \vec{y} = \sum_{n=1}^{N} x_n^* y_n \tag{3.48}$$

as the definition of the dot product in case of complex coordinates. Thereby

$$|\vec{x}| = \sqrt{\sum_{n=1}^{N} x_n^* x_n} = \sqrt{\sum_{n=1}^{N} |x_n|^2}$$

always yields nonnegative real results, also yielding nonzero results if any of the components is nonzero. Notice that for real x_n and y_n (3.47) and (3.48) yield identical results, that is (3.48) is a generalization of (3.47).

Having switched the definition of the dot product, we also need to switch the procedure of matrix transposition. E.g. usually in an orthonormal basis we could express the dot product of vectors $\vec{x}$ and $\vec{y}$ as $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ where $\mathbf{x}$ and $\mathbf{y}$ are column matrices containing the coordinates of vectors $\vec{x}$ and $\vec{y}$. However this no longer matches the formula (3.48). This problem is commonly solved by using

Hermitian transposition instead, which, in addition to turning columns into rows and vice versa, also performs complex conjugation of the matrix components. The expression $\mathbf{x}^*\mathbf{y}$ (where the asterisk denotes Hermitian transposition) used instead of $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ provides the result matching the formula (3.48).

▷ Apparently, it doesn't matter if we conjugate x_n or y_n in (3.48), this is simply a matter of convention and different conventions are in use. We prefer to conjugate x_n, because then the dot product $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ is replaced by $\mathbf{x}^*\mathbf{y}$.

▷ While the dot product defined by (3.47) is bilinear (linear in each of the arguments), after switching to (3.48) the linearity in the first argument doesn't hold anymore, because for the linear case we would have needed $(a\vec{x})\cdot\vec{y} = a\cdot(\vec{x}\cdot\vec{y})$, but instead we obtain the so-called *antilinear* ("conjugate-linear") behavior $(a\vec{x})\cdot\vec{y} = a^*\cdot(\vec{x}\cdot\vec{y})$. The two-argument function defined by (3.48) is referred to as *sesquilinear* ("1.5-linear").

Hermitian transposition $\mathbf{A}^*$ commonly replaces almost all usages of matrix transposition $\mathbf{A}^{\mathrm{T}}$ if complex numbers are involved, so that the "dot products" of matrix rows by matrix columns, which are emerging during multiplication of matrices, respect the formula (3.48). E.g. the definition of an orthogonal matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$ is replaced by the definition of a unitary matrix $\mathbf{A}^*\mathbf{A} = \mathbf{I}$. The definition still claims the mutual orthonormality of the matrix's columns (or rows, if the order of factors is swapped), however it now uses (3.48) to assess this orthonormality.

Complex linear spaces

After switching the dot product definition (3.47) to (3.48) we notice that the idea of using the equation $\vec{x}\cdot\vec{x} = -1$ to "get a hold of complex vectors" works even worse than before: now the value of the expression $\vec{x}\cdot\vec{x}$ is nonnegative real $\forall\vec{x}$ and the equation $\vec{x}\cdot\vec{x} = -1$ has no solutions at all. Therefore we need a different idea. How about we take some unit-length vector $\vec{x}$ and look for a unit-length vector $\vec{y}$, such that $\vec{x}\cdot\vec{y} = i$? Indeed, an "imaginary" unit vector $\vec{y} = i\vec{x}$ should yield

$$\vec{x}\cdot\vec{y} = \vec{x}\cdot i\vec{x} = i\cdot(\vec{x}\cdot\vec{x}) = i\cdot|\vec{x}|^2 = i$$

Also, no other unit vectors should yield the same result.

▷ Indeed, suppose there is another vector $\vec{y}\,'$ such that $\vec{x}\cdot\vec{y}\,' = i$. Then

$$0 = i - i = \vec{x}\cdot\vec{y}\,' - \vec{x}\cdot\vec{y} = \vec{x}\cdot(\vec{y}\,' - \vec{y})$$

and also

$$(\vec{y}\,' - \vec{y})\cdot\vec{x} = \big(\vec{x}\cdot(\vec{y}\,' - \vec{y})\big)^* = 0^* = 0$$

that is vector $\vec{y}\,' - \vec{y}$ is orthogonal to $\vec{x}$, and also it has nonzero length since $\vec{y}\,' \neq \vec{y}$. But then by the Pythagorean theorem the length of of $\vec{y}\,'$ must be larger than the length of $\vec{y}$:

$$\begin{aligned}
|\vec{y}\,'|^2 &= |\vec{y} + (\vec{y}\,' - \vec{y})|^2 = \big(\vec{y} + (\vec{y}\,' - \vec{y})\big)\cdot\big(\vec{y} + (\vec{y}\,' - \vec{y})\big) \\
&= \vec{y}\cdot\vec{y} + (\vec{y}\,' - \vec{y})\cdot\vec{y} + \vec{y}\cdot(\vec{y}\,' - \vec{y}) + (\vec{y}\,' - \vec{y})\cdot(\vec{y}\,' - \vec{y}) \\
&= |\vec{y}|^2 + (\vec{y}\,' - \vec{y})\cdot i\vec{x} + i\vec{x}\cdot(\vec{y}\,' - \vec{y}) + |\vec{y}\,' - \vec{y}|^2 \\
&= |\vec{y}|^2 + i\big((\vec{y}\,' - \vec{y})\cdot\vec{x}\big) + (-i)\big(\vec{x}\cdot(\vec{y}\,' - \vec{y})\big) + |\vec{y}\,' - \vec{y}|^2
\end{aligned}$$

$$= |\vec{y}|^2 + i \cdot 0 + (-i) \cdot 0 + |\vec{y}\,' - \vec{y}|^2 = |\vec{y}|^2 + |\vec{y}\,' - \vec{y}|^2 > |\vec{y}|^2 = 1$$

and thus $\vec{y}\,'$ cannot be of unit length.

Therefore, similarly to how we introduced vector length in order to be able to define multiplication of vectors by real numbers, we introduce the dot product in order to be able to define imaginary vectors.

▷ Since we already introduced the length, it would suffice to introduce the angle between vectors. Then the dot product can be simply defined as

$$\vec{x} \cdot \vec{y} = |\vec{x}| \cdot |\vec{y}| \cdot \cos\theta$$

Having $\vec{x} \cdot \vec{y} = i$ where $|\vec{x}| = |\vec{y}| = 1$ implies $\cos\theta = i$.

If we really want to find the value of θ, we can express the cosine via the exponential function:

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} = i$$
$$e^{i\theta} + e^{-i\theta} = 2i$$
$$\left(e^{i\theta}\right)^2 - 2ie^{i\theta} + 1 = 0$$

Using (1.4),

$$e^{i\theta} = i \pm \sqrt{-1-1} = i(1 \pm \sqrt{2})$$
$$i\theta = \operatorname{Log} i(1 \pm \sqrt{2}) = \ln(1 \pm \sqrt{2}) + i\frac{\pi}{2}$$

Now, if we take the "+" sign, we get

$$i\theta = \operatorname{Log} i(1 + \sqrt{2}) = \operatorname{Log}(1 + \sqrt{2}) + \operatorname{Log} i = \ln(1 + \sqrt{2}) + i\frac{\pi}{2}$$

If we take the "−" sign, we get

$$\begin{aligned} i\theta &= \operatorname{Log} i(1 - \sqrt{2}) = \operatorname{Log}(-i)(\sqrt{2} - 1) = \operatorname{Log}(\sqrt{2} - 1) + \operatorname{Log}(-i) \\ &= \ln(\sqrt{2} - 1) - i\frac{\pi}{2} = -\ln(1 + \sqrt{2}) - i\frac{\pi}{2} \end{aligned}$$

That is we just obtained two opposite values

$$\theta = \pm\left(\frac{\pi}{2} - i\ln(1 + \sqrt{2})\right)$$

corresponding to the cosine being an even function.

▷ Same as we have restricted our linear space to be a Banach space in order to derive multiplication with a real number, this time, by additionally introducing the dot product, we further restrict it to be a Hilbert space.

Goal 3.1 completed. ■

Linear functions

We now also wish to infer homogeneity from additivity for functions of complex argument (so that we can infer linearity), thereby completing Goal 3.2. For functions of real argument we could do this under an additional assumption of continuity, but while real numbers can be obtained as limits of rational sequences, complex or even imaginary numbers cannot. Thus, continuity doesn't help much.

As a bare minimum we would like to infer the property $f(ix) = if(x)$, the rest will automatically follow from the already inferred homogeneity with respect to real factors as well as the function's additivity. Recall that we just solved the problem of multiplication of vectors by the imaginary unit by introducing the dot product, which, intuitively, allowed us to measure the angles between the vectors. This may lead to the following related idea: in order to obtain the property $f(ix) = if(x)$, we could require the transformation by $f(x)$ to "preserve the angles in the complex plane", that is we require the function to define a conformal map, that is f should be an analytic function.

Being analytic, f would preserve the angles locally. In particular at the origin point (where we already know that $f(0) = 0$ by the earlier discussion in Section 3.2) we are going to have

$$f(i \cdot \mathrm{d}x) = i \cdot f(\mathrm{d}x) \tag{3.49a}$$

where $\mathrm{d}x$ intuitively is an infinitely small value. Equation (3.49a) reads that rotating the function's argument by 90° in the complex plane rotates the function's value by the same 90°. By additivity of f,

$$\begin{aligned} f(i \cdot x) &= f\left(i \cdot \int_0^x \mathrm{d}t\right) = f\left(\int_0^x i \cdot \mathrm{d}t\right) \overset{(3.18a)}{=} \int_0^x f(i \cdot \mathrm{d}t) \overset{(3.49a)}{=} \int_0^x i \cdot f(\mathrm{d}t) \\ &= i \int_0^x f(\mathrm{d}t) \overset{(3.18a)}{=} i \cdot f\left(\int_0^x \mathrm{d}t\right) = i \cdot f(x) \end{aligned} \tag{3.49b}$$

Respectively, multiplication with a generic complex number $a + ib$ yields

$$\begin{aligned} f\big((a+ib) \cdot x\big) &= f\big(ax + ibx\big) \overset{(3.18a)}{=} f(ax) + f(ibx) \overset{(3.49b)}{=} f(ax) + if(bx) \\ &\overset{(3.27)}{=} af(x) + ibf(x) = (a+ib)f(x) \end{aligned}$$

That is

$$f(\lambda x) = \lambda f(x) \quad \forall \lambda \in \mathbb{C} \tag{3.50}$$

▷ Same as with continuity, it suffices for the function to be analytic at one point. Analyticity at other points follows by additivity.

▷ To define a complex linear space we turned a linear space equipped just with the length into a linear space equipped with a dot product (which is a stronger requirement, since dot product implies the existence of the length). Similarly, here we turned the requirement of function's continuity into a stronger requirement of the function being analytic.

The argument made above to infer homogeneity from analyticity applies for functions of a scalar complex argument $x \in \mathbb{C}$. If we want x to be a vector, we need to generalize the idea of analytic function to a function of a vector argument. Analytic functions can be defined as those which can be expanded into 1st order Taylor series of the form

$$f(x_0 + \mathrm{d}x) = f(x_0) + f'(x_0)\,\mathrm{d}x \tag{3.51a}$$

which is simply another way to express the differentiability of $f(x)$. Replacing the scalars in (3.51a) by vectors we obtain

$$\mathbf{f}(\mathbf{x}_0 + \mathrm{d}\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{f}'(\mathbf{x}_0) \cdot \mathrm{d}\mathbf{x} \tag{3.51b}$$

where $\mathbf{f}'(\mathbf{x}_0)$ is Jacobian matrix[5] consisting of partial derivatives $\partial f_n/\partial x_m$, that is we need existence of partial derivatives of f_n in each of the arguments x_m.

Goal 3.2 completed. ■

Summary of Chapter 3

- We have looked into defining linear spaces and the linearity property based on addition/additivity only, where we managed to have quite some success by following how number classes are defined, although we had to take some additional assumptions.
- We have used the linearity property to help us get insights into the dot product and rotations.
- We have seen that at small scale functions become almost linear, while at larger scales Taylor series provide an expansion into a series of multilinear terms.
- We have discussed the implications of promoting real linear spaces to complex ones.

[5]Some discussion of Jacobian matrix can be found in Subsection 6.8.5.

Chapter 4

Beyond the basis

Basis-independent (covariant-form) equations. Linear basis-independent entities (tensors) as generalizations of the ideas of vectors, linear operators, bilinear forms etc. Tensor index notation and its relationship to the matrix notation. Basic operations on tensors. Review of some of the linear algebra topics from the tensor algebra perspective. Dual space. Tensor-like entities.

4.1 Covariant-form equations . . . 94
4.2 Matrix notation . . . 95
4.3 Basis change . . . 97
4.4 Linear forms . . . 101
4.5 Co- and contra-variance . . . 102
4.6 Tensor index notation . . . 104
4.7 Tensor components . . . 108
4.8 Kronecker delta symbol . . . 110
4.9 Metric tensor . . . 111
4.9.1 Dot product . . . 111
4.9.2 Index lowering . . . 113
4.9.3 Index rasing . . . 114
4.10 Tensor product . . . 115
4.11 Matrix trace . . . 117
4.12 Decomposable tensors . . . 120
4.13 Dual basis . . . 122
4.14 Differentiation . . . 124
4.15 Linear transformations of space . . . 128
4.15.1 Equivalence to basis change . . . 128
4.15.2 Pseudotensors . . . 130
4.16 Tensors and complex numbers . . . 131
Summary of Chapter 4 . . . 132

Linear algebra courses are typically based on matrix notation. However, the matrix notation, despite being a powerful notational tool, fails to highlight a number of aspects of crucial importance. This results in a number of common confusions among linear algebra students, such as

- A string of N numbers is often thought to be the same as a vector in N dimensions. E.g. $(10, 5, -1)$ is seen as a 3D vector. Linear algebra textbooks will usually explicitly make a reservation that this idea is limited (if not plain wrong in many contexts), but, lacking further detailed explanation and practice, this reservation is often eventually ignored and forgotten by students.

- Since basis changes are mathematically equivalent to linear transformations of the linear space, students often feel tempted to "simplify matters" and just concentrate on the linear transformations (as those seem to be more intuitive anyway).

- Matrix diagonalization topics, often being presented in an abstract way and lacking context, may be perceived as "an exercise for the sake of exercise".[1]

A good linear algebra course will strive to minimize those confusions. The problem is, however, that the matrix notation, which is lying at the core of the linear algebra presentation, is not a too well suitable framework to express certain key ideas. These ideas look way more natural in tensor terms, but tensors are typically omitted from introductory-level linear algebra courses. The result is that linear algebra often, due to a lack of proper understanding, turns into some kind of ritualized process.

Conversely, in a number of courses tensors may be directly introduced in the form of tensor calculus, skipping the tensor algebra part. One may question this latter approach, since reinterpreting the matrix notation-based linear algebra knowledge in terms of tensor algebra can solidify the understanding of the linear algebra a lot and simultaneously serve as a better foundation for diving into tensor calculus topics.

In this chapter we will attempt to amend the mentioned problems by providing an introduction into the key ideas of tensor algebra.

4.1 Covariant-form equations

At the root of the idea of tensors (as basis-independent linear entities) is the desire to write equations in a basis-independent (a.k.a. *covariant*) form. Thinking of it, the idea is fully natural. A linear space basis is an artificial construct, introduced to allow expressing the entities like vectors in a numerical form (by writing out their coordinates with respect to the basis). In fact, e.g. in physics we already have a habit of writing the laws of nature without making any assumptions of a linear space basis. E.g. the second Newton's law

$$\vec{F} = m\vec{a}$$

equates two vectors to each other without being specific about which basis should be used.

Besides, imagine what would have happened if formulas had been correct only in one specific basis each. Then a formula requiring one basis would be

[1] Matrix diagonalization topics will be discussed in Chapter 7, but that discussion will be partially based on the materials of the present chapter.

incompatible with another formula, requiring a different basis, unless we include an explicit transformation from one basis to the other. The whole thing could quickly get unwieldy.

A straightforward way to ensure that equations are in a basis-independent form is to make sure that equations only use basis-independent entities and basis-independent operations. A general concept expressing the idea of a linear basis-independent entity, which has a set of basis-independent operations defined on this entity, is the concept of a *tensor*. Respectively, if we write equations exclusively using tensors and tensor operations, we would thereby ensure that our equations are in a covariant form.

The simplest example of a tensor is a vector, provided the vector is seen from the perspective of being a basis-independent object and not as a string of its coordinates (as such string would be explicitly or implicitly assuming some basis). It helps to intuitively visualize such vector as a direction which has a length. This direction exists "in the real world", without us choosing any specific set of basis vectors. In other words, regarding of whatever basis we may use (or not use any basis at all), this vector denotes always one and the same direction and length.

> ▷ Assuming that vector directions and lengths exist without reference to any other objects effectively means that we treat linear space as an absolute space, and this is how we should conceptually treat it. Even if we treat space as being relative (like in physics), the reference objects do not automatically constitute a linear basis, as we would still need to define the coordinate axes.

Motivational starter for Chapter 4

Develop a formal approach/framework to support basis-independent aspects of linear algebraic entities.

4.2 Matrix notation

We will use matrix notation as the starting point towards introducing a different notation, namely *tensor index notation*. Before staring the departure from the matrix notation, we shall review some of its aspects.

The most important thing to understand about matrix notation is that this notation doesn't really convey any special semantics, it's merely a way to write certain commonly occurring expression patterns in a more concise form. With this in mind, let's look at the few following examples.

A dot product of two vectors is usually expressed in matrix notation as a matrix multiplication of a row-vector by a column-vector:

$$\mathbf{x}^{\mathrm{T}} \cdot \mathbf{y} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3 \tag{4.1}$$

It is often left unnoticed that the same kind of matrix product but with the opposite order of factors is also making sense from the matrix notation perspec-

tive:

$$\mathbf{x} \cdot \mathbf{y}^{\mathrm{T}} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cdot \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & x_1 y_3 \\ x_2 y_1 & x_2 y_2 & x_2 y_3 \\ x_3 y_1 & x_3 y_2 & x_3 y_3 \end{bmatrix} = \mathbf{M} \tag{4.2}$$

Indeed, the number of columns in the first matrix in (4.2) is 1 and the number of rows in the second matrix in (4.2) is also 1, therefore the number of rows in the first matrix is equal to the number of columns in the second matrix, and therefore both matrices can be multiplied yielding a third matrix which is a square 3×3 matrix. Equation (4.2) is actually the formula of the *outer product* of two vectors, which is usually a part of more advanced linear algebra courses.

> ▷ The outer product of vectors is sometimes referred to as *tensor product* of vectors. In some contexts both terms are equivalent, but in some others, especially if complex numbers are involved, they are not. In particular, in quantum mechanics, component-wise both of these products look similar, but the result of an outer product is a matrix like in (4.2), with an additional conjugation of the components of the transposed vector, since Hermitian transpose is used, while in a tensor product this matrix is flattened into a single "column vector" (or a "nested column vector" to be more precise) and there is no conjugation.
>
> The outer product also should not be confused with the *exterior product*, which we will discuss in Chapter 5.

Now the next thing which we are going to do is to *stop thinking of vectors as column matrices* or generally thinking of vectors as objects "consisting of their coordinates". Instead we shall think of them as opaque objects, not having any "internal structure". An intuitive geometric interpretation of a vector as a "quantified direction" in 3D space (that is, direction which has a length) might work much better in this respect. Thinking of vectors as some fully abstract entities could do as well, the main point being to disconnect the idea of a vector from the idea of its coordinates. Coordinates are something which exists only in relationship to a particular basis, but, as we mentioned earlier, a vector can exist in the absence of any basis, or, vice versa, in the presence of multiple bases. A vector per se therefore doesn't have anything to do with the basis and therefore doesn't have anything to do with the coordinates either.

We are going to use the arrow notation for vectors (e.g. $\vec{x}$) to highlight their geometric and/or abstract aspects. Boldface lowercase will be used to denote column matrices consisting of coordinates of vectors in some basis. E.g., given basis vectors $\vec{e}_1$, $\vec{e}_2$ and $\vec{e}_3$, we can express $\vec{x}$ in this basis as

$$\vec{x} = x_1 \vec{e}_1 + x_2 \vec{e}_2 + x_3 \vec{e}_3 \tag{4.3}$$

while boldface $\mathbf{x}$ denotes a column of the components of the same vector:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{4.4}$$

Respectively, row matrices will be denoted using transposition, e.g.

$$\mathbf{y}^{\mathrm{T}} = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}$$

Now remember what we said: matrix notation is merely a formal thing. This, in particular, means that matrix elements do not necessarily have to be

numbers. In particular, we could form a row matrix of 3 elements, where each element is not a number, but a basis vector $\vec{e}_i$:

$$\vec{\mathbf{e}}^{\,\mathrm{T}} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \tag{4.5}$$

We could multiply that matrix row with the column matrix $\mathbf{x}$ from (4.4) obtaining another form of writing (4.3):

$$\vec{x} = x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3 = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{x} \tag{4.6}$$

4.3 Basis change

The next thing we shall do is review the transformation of components of vectors and matrices under a change of basis. This is an absolutely key topic for understanding the idea of tensors.

Let $\{\vec{e}_i\}$ be our old basis (in which (4.6) is specified) and let $\{\vec{e}_i{}'\}$ be the new basis. Expressing each of the vectors $\vec{e}_i{}'$ in terms of vectors $\vec{e}_i$, since each of $\vec{e}_i{}'$ (as well as any other vector in the same space) can be expressed as a linear combination of $\vec{e}_i$, we have

$$\begin{aligned} \vec{e}_1{}' &= T_{11}\vec{e}_1 + T_{21}\vec{e}_2 + T_{31}\vec{e}_3 \\ \vec{e}_2{}' &= T_{12}\vec{e}_1 + T_{22}\vec{e}_2 + T_{32}\vec{e}_3 \\ \vec{e}_3{}' &= T_{13}\vec{e}_1 + T_{23}\vec{e}_2 + T_{33}\vec{e}_3 \end{aligned} \tag{4.7}$$

where T_{ij} are the coefficients of those linear combinations. Or, writing each line of (4.7) in matrix notation similarly to (4.6):

$$\vec{e}_i{}' = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \begin{bmatrix} T_{1i} \\ T_{2i} \\ T_{3i} \end{bmatrix} \tag{4.8}$$

where i goes from 1 to 3. We could combine all 3 different equations arising for 3 different values of i in (4.8) into a single matrix equation:

$$\begin{bmatrix} \vec{e}_1{}' & \vec{e}_2{}' & \vec{e}_3{}' \end{bmatrix} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix} \tag{4.9}$$

Let's denote the matrix on the right of the dot in the right-hand side of (4.9) by $\mathbf{T}$. Then (4.9) becomes

(transformation of basis vectors upon a change of basis)

$$\vec{\mathbf{e}}'^{\,\mathrm{T}} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{T} \tag{4.10}$$

Goal 4.1. Obtain the formula for the transformation of vector components under a change of basis. *Hint:* use (4.6) and (4.10).

In the new basis $\{\vec{e}_i{}'\}$ the formula (4.6) should look as

$$\vec{x} = x_1'\vec{e}_1{}' + x_2'\vec{e}_2{}' + x_3'\vec{e}_3{}' = \begin{bmatrix}\vec{e}_1{}' & \vec{e}_2{}' & \vec{e}_3{}'\end{bmatrix} \cdot \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \vec{\mathbf{e}}'^{\,\mathrm{T}} \cdot \mathbf{x}' \tag{4.11}$$

where x_i' are the coordinates of vector $\vec{x}$ in the new basis $\{\vec{e}_i{}'\}$ and $\mathbf{x}'$ is the column matrix of those coordinates.

> ▷ Notice that the vector $\vec{x}$ in the left-hand side of (4.11) is not primed, since the vector $\vec{x}$ itself (as a geometric or abstract object) does not change. It is the very same vector, just this time we're expanding it in a different basis.

Since both (4.6) and (4.11) expand one and the same vector, we can equate the expansions:

$$\vec{x} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{x} = \vec{\mathbf{e}}'^{\,\mathrm{T}} \cdot \mathbf{x}' \tag{4.12}$$

If we substitute (4.10) into (4.12) we obtain

$$\vec{x} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{x} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{T} \cdot \mathbf{x}' \tag{4.13}$$

immediately suggesting

$$\mathbf{x} = \mathbf{T} \cdot \mathbf{x}' \tag{4.14}$$

Since an expansion of a vector in a basis is unique, there are no other possibilities for the component values of $\mathbf{x}'$ except the ones satisfying (4.14).

There is apparently a restriction that the matrix $\mathbf{T}$ cannot be singular, because otherwise basis vectors $\{\vec{e}_i{}'\}$ would be linearly dependent and won't be actually forming a basis. Since $\mathbf{T}$ is non-singular, we can multiply left- and right-hand sides of (4.14) by $\mathbf{T}^{-1}$ obtaining

$$\mathbf{T}^{-1} \cdot \mathbf{x} = \mathbf{x}'$$

or, resolving for $\mathbf{x}'$,

(transformation of vector coordinates under a change of basis)

$$\mathbf{x}' = \mathbf{T}^{-1} \cdot \mathbf{x} \tag{4.15a}$$

or, writing out the components of $\mathbf{x}$ and $\mathbf{x}'$ explicitly

$$\begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \mathbf{T}^{-1} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{4.15b}$$

Goal 4.1 completed. ■

Goal 4.2. Obtain the formula for the transformation of the components of a square matrix under a change of basis. *Hint:* there will be different formulas depending on the matrix semantics.

Goal 4.2a. Obtain the formula for the transformation of the components of a square matrix under a change of basis, given that the matrix represents a linear transformation of vector space.

Let $\mathbf{A}$ be a matrix representing a linear transformation of space:

$$\mathbf{y} = \mathbf{A} \cdot \mathbf{x} \tag{4.16}$$

that is $\mathbf{A}$ is a matrix representation of a *linear operator*. By (4.14) and (4.15a)

$$\begin{aligned}\mathbf{y}' &\overset{(4.15a)}{=} \mathbf{T}^{-1} \cdot \mathbf{y} \overset{(4.16)}{=} \mathbf{T}^{-1} \cdot \mathbf{A} \cdot \mathbf{x} \overset{(4.14)}{=} \mathbf{T}^{-1} \cdot \mathbf{A} \cdot \mathbf{T} \cdot \mathbf{x}' = (\mathbf{T}^{-1} \cdot \mathbf{A} \cdot \mathbf{T}) \cdot \mathbf{x}' \\ &= \mathbf{A}' \cdot \mathbf{x}'\end{aligned}$$

where

(transformation of a linear operator matrix under a change of basis)

$$\mathbf{A}' = \mathbf{T}^{-1} \cdot \mathbf{A} \cdot \mathbf{T} \tag{4.17}$$

The matrix transformation from $\mathbf{A}$ to $\mathbf{A}'$ defined by (4.17) is referred to as *similarity transformation* of matrix $\mathbf{A}$.

Goal 4.2a completed. ■

Goal 4.2b. Obtain the formula for the transformation of the components of a square matrix under a change of basis, given that the matrix represents a bilinear form. *Hint:* the value returned by the form doesn't depend on the choice of basis.

Let $\mathbf{B}$ be a matrix representing a bilinear form:

$$u = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{B} \cdot \mathbf{y} \tag{4.18}$$

that is u is a function of $\vec{x}$ and $\vec{y}$, which is linear in each of the two arguments.

Let's for a second return to (4.16). The linear transformation of space is essentially a linear, vector-valued function $\vec{y} = \vec{A}(\vec{x})$ of a vector argument $\vec{x}$, represented in the matrix form as (4.16). In (4.17) we have established that the function is represented in the new basis by a matrix equation of the same form $\mathbf{y}' = \mathbf{A}' \cdot \mathbf{x}'$, but the matrix components in the new basis are different, as given by (4.17).

Similarly, a bilinear form is a bilinear scalar-valued function $u = B(\vec{x}, \vec{y})$, defined for a pair of vector arguments. Since the result of the function is a scalar, it should be independent of the basis.

> ▷ This is actually the definition of a scalar: it is a numeric value existing independently of the basis (pretty much like a vector $\vec{v}$ exists independently of the basis).

Substituting (4.14) into (4.18) we obtain

$$\begin{aligned}u &\overset{(4.18)}{=} \mathbf{x}^{\mathrm{T}} \cdot \mathbf{B} \cdot \mathbf{y} \overset{(4.14)}{=} (\mathbf{T} \cdot \mathbf{x}')^{\mathrm{T}} \cdot \mathbf{B} \cdot (\mathbf{T} \cdot \mathbf{y}') = \mathbf{x}'^{\mathrm{T}} \cdot \mathbf{T}^{\mathrm{T}} \cdot \mathbf{B} \cdot \mathbf{T} \cdot \mathbf{y}' \\ &= \mathbf{x}'^{\mathrm{T}} \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{B} \cdot \mathbf{T}) \cdot \mathbf{y}' = \mathbf{x}'^{\mathrm{T}} \cdot \mathbf{B}' \cdot \mathbf{y}'\end{aligned}$$

where

(transformation of a bilinear form matrix under a change of basis)

$$\mathbf{B}' = \mathbf{T}^{\mathrm{T}} \cdot \mathbf{B} \cdot \mathbf{T} \tag{4.19}$$

The matrix transformation from $\mathbf{B}$ to $\mathbf{B}'$ defined by (4.19) is referred to as *congruence transformation* of matrix $\mathbf{B}$.

Goal 4.2b completed. ■

Goal 4.2 completed. ■

Let's review our results (4.10), (4.15a), (4.17) and (4.19), which we for convenience summarize here once again:

$$\vec{\mathbf{e}}'^{\,\mathrm{T}} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{T} \tag{4.20a}$$

$$\mathbf{x}' = \mathbf{T}^{-1} \cdot \mathbf{x} \tag{4.20b}$$

$$\mathbf{A}' = \mathbf{T}^{-1} \cdot \mathbf{A} \cdot \mathbf{T} \tag{4.20c}$$

$$\mathbf{B}' = \mathbf{T}^{\mathrm{T}} \cdot \mathbf{B} \cdot \mathbf{T} \tag{4.20d}$$

These results were obtained from the idea that $\mathbf{x}$, $\mathbf{A}$, $\mathbf{B}$ represent entities which exist independently of the basis $\vec{\mathbf{e}}^{\,\mathrm{T}}$. Specifically we started with the idea that the vectors exist independently of the basis choice. We also realized that numbers (at least some of them) exist independently of the basis choice, these numbers being referred to as *scalars*. Then we also had a couple of different function kinds, where the function arguments and values were vectors and/or scalars, therefore such functions also exist independently of the basis choice.

Respectively the column matrix $\mathbf{x}$ and the square matrices $\mathbf{A}$ and $\mathbf{B}$ are representations of these basis-independent entities (a scalar being its own representation, therefore not requiring yet another representation). These representations have different components under different bases, and the relationships between these components are given by (4.20).

> ▷ It was actually not granted that an entity (a linear operator) representable by matrix multiplication (4.16) in one basis would be representable by matrix multiplication of the same form $\mathbf{y}' = \mathbf{A}'\mathbf{x}'$ in another basis. Similarly, it wasn't granted that an entity (a bilinear form) representable by (4.18) in one basis would be representable by an equation of the same form $u = \mathbf{x}'^{\mathrm{T}}\mathbf{B}'\mathbf{y}'$ in another basis.
>
> In this regard notice how in obtaining the formulas (4.17) and (4.19) we also implicitly derived that the respective entities are representable in the same matrix form in the new basis. We'll return to this topic once more in Section 4.7.

There seems to be some pattern in (4.20), although maybe a not 100% clear one yet. At least the transformations seem to involve multiplications with matrices $\mathbf{T}$, $\mathbf{T}^{-1}$ and $\mathbf{T}^{\mathrm{T}}$. Let's try to find out some more details about this pattern.

The relationship between the equations (4.20a) and (4.20b) is actually pretty straightforward. By multiplying their left- and right-hand sides we obtain

$$\vec{x} = \vec{\mathbf{e}}'^{\,\mathrm{T}} \cdot \mathbf{x}' = \vec{\mathbf{e}}^{\,\mathrm{T}}\mathbf{T} \cdot \mathbf{T}^{-1}\mathbf{x} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{T}\mathbf{T}^{-1} \cdot \mathbf{x} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{I} \cdot \mathbf{x} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{x} \tag{4.21}$$

where we can see how the matrices $\mathbf{T}$ and $\mathbf{T}^{-1}$ cancel each other in the middle.

However before attempting to apply the same idea to (4.20c) and (4.20d) (which is going to be less straightforward) let's construct a simpler situation first.

4.4 Linear forms

Equations (4.20c) and (4.20d) define the component transformation for the linear operator $\vec{A}(\vec{x})$ and the bilinear form $B(\vec{x}, \vec{y})$ respectively. We can notice that in each case we deal with functions defined on the vector space. In the first case we are having a vector-valued linear function of a single vector argument, in the second case we are having a scalar-valued function of two vector arguments, also linear in each of the arguments.

Maybe we should start by looking at a simpler class of linear functions defined on the vector space, namely, scalar-valued linear functions of a single vector argument $y = h(\vec{x})$. Rewriting the equation (4.16) in the component form:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{31} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

we can notice that if we want a scalar rather than vector result, it would suffice to take just the first row of the matrix $\mathbf{A}$:

$$y_1 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

or, renaming y_1 to y and a_{1i} to h_i

$$y = \begin{bmatrix} h_1 & h_2 & h_3 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{x} = h(\vec{x}) \tag{4.22}$$

The entity $h(\vec{\cdot})$ defined by (4.22) is referred to as a *linear form*. Respectively the row matrix $\mathbf{h}^{\mathrm{T}}$ is the representation of $h(\vec{\cdot})$ in a basis, pretty much like $\mathbf{x}$ is a representation of $\vec{x}$ in a basis.

▷ Linear forms are standard linear algebraic entities, pretty much like the linear operators and bilinear forms. However they are often omitted from basic linear algebra courses. Therefore we took some effort to construct a motivation to introduce them. They are going to play one of the key roles in our subsequent discussion.

▷ Another way to introduce linear forms could be taking a bilinear form and fixing one of the arguments to a particular value, therefore obtaining a linear function of a single argument. Yet another (and probably formally the most correct) way to introduce them is to simply define them as linear functions of a vector argument. The latter idea will be reviewed in Section 4.7.

We can immediately notice a strong similarity between (4.6) and (4.22), just instead of a row matrix of $\vec{e}_i$ we are having a row matrix of h_i. This suggests that upon a basis change $\mathbf{x}$ will be transformed by (4.20b), as any vector, and $\mathbf{h}^{\mathrm{T}}$ will be transformed, similarly to $\vec{\mathbf{e}}^{\,\mathrm{T}}$, by (4.20a):

(transformation of linear form components under a change of basis)

$$\mathbf{h}'^{\mathrm{T}} = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{T} \tag{4.23}$$

Let's verify our conjecture (4.23).

Goal 4.3. Obtain (4.23) in a manner similar to (4.15a), (4.17) and (4.19). *Hint:* the value of $h(\vec{x})$ must behave as a scalar.

Since $y = h(\vec{x})$ must be a scalar, we have

$$y = h(\vec{x}) = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{x} = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{T}\mathbf{x}' = \mathbf{h}^{\mathrm{T}}\mathbf{T} \cdot \mathbf{x}' = \mathbf{h}'^{\mathrm{T}} \cdot \mathbf{x}'$$

where $\mathbf{h}'^{\mathrm{T}} = \mathbf{h}^{\mathrm{T}}\mathbf{T}$.

Goal 4.3 completed. ■

Respectively, in place of (4.21) we obtain

$$y = \mathbf{h}'^{\mathrm{T}} \cdot \mathbf{x}' = \mathbf{h}^{\mathrm{T}}\mathbf{T} \cdot \mathbf{T}^{-1}\mathbf{x} = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{T}\mathbf{T}^{-1} \cdot \mathbf{x} = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{I} \cdot \mathbf{x} = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{x} \tag{4.24}$$

where, again, $\mathbf{T}$ and $\mathbf{T}^{-1}$ are cancelling each other.

4.5 Co- and contra-variance

Comparing (4.21) and (4.24) we can see that in both cases in the left-hand side we are having basis-independent entities $\vec{x}$ and y respectively. In both cases this basis-independent entity is being equal to a dot product of a row matrix and a column matrix. Now in the first case (4.21), upon the basis change, the row matrix was multiplied by $\mathbf{T}$ (according to the definition of the basis change), which caused the column matrix to get multiplied by $\mathbf{T}^{-1}$ to keep the dot product unchanged. In the second case (4.24), upon the basis change, the column matrix is multiplied by $\mathbf{T}^{-1}$ (as we just obtained by (4.21)), which causes the row matrix to get multiplied by $\mathbf{T}$ to keep the dot product unchanged.

This suggests that there is the same kind of duality between $\mathbf{h}^{\mathrm{T}}$ and $\mathbf{x}$ as between $\vec{\mathbf{e}}^{\mathrm{T}}$ and $\mathbf{x}$:

$$\begin{aligned} \vec{\mathbf{e}}'^{\mathrm{T}} &= \vec{\mathbf{e}}^{\mathrm{T}} \cdot \mathbf{T} \\ \mathbf{h}'^{\mathrm{T}} &= \mathbf{h}^{\mathrm{T}} \cdot \mathbf{T} \end{aligned} \qquad \mathbf{x}' = \mathbf{T}^{-1} \cdot \mathbf{x} \qquad \begin{aligned} \vec{\mathbf{e}}'^{\mathrm{T}} \cdot \mathbf{x}' &= \vec{\mathbf{e}}^{\mathrm{T}}\mathbf{T} \cdot \mathbf{T}^{-1}\mathbf{x} = \vec{\mathbf{e}}^{\mathrm{T}} \cdot \mathbf{x} \\ \mathbf{h}'^{\mathrm{T}} \cdot \mathbf{x}' &= \mathbf{h}^{\mathrm{T}}\mathbf{T} \cdot \mathbf{T}^{-1}\mathbf{x} = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{x} \end{aligned}$$

Let's rewrite the same equations in terms of matrix components:

$$\begin{aligned} \vec{e}'_j &= \sum_i \vec{e}_i T_{ij} \\ h'_j &= \sum_i h_i T_{ij} \end{aligned} \qquad x'_j = \sum_i x_i T'_{ji} \qquad \begin{aligned} \sum_i \vec{e}'_i x'_i &= \sum_i \vec{e}_i x_i \\ \sum_i h'_i x'_i &= \sum_i h_i x_i \end{aligned}$$

where T_{ij} and T'_{ij} are the components of $\mathbf{T}$ and $\mathbf{T}^{-1}$ respectively, the first index always identifying the matrix row, the second index always identifying the matrix column:

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{31} & T_{33} \end{bmatrix} \qquad \mathbf{T}^{-1} = \begin{bmatrix} T'_{11} & T'_{12} & T'_{13} \\ T'_{21} & T'_{22} & T'_{23} \\ T'_{31} & T'_{31} & T'_{33} \end{bmatrix}$$

We can see that whenever T_{ij} is involved, there is always summation over the first index of T_{ij}. Conversely, when T'_{ij} is involved, there is a summation over the second index of T'_{ij}. Obviously this corresponds to multiplication by $\mathbf{T}$ from the right or multiplication by $\mathbf{T}^{-1}$ from the left respectively. We will refer to the first kind of summation, corresponding to multiplication by $\mathbf{T}$ from the right (like it happens for the basis vectors $\vec{e}_i$) as *covariant transformation*. We will refer to the second kind of summation, corresponding to multiplication by $\mathbf{T}^{-1}$ from the left, thereby being kind of an "opposite" of covariant transformation, as *contravariant transformation*. In the dot product $\vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{x}$ or in the dot product $\mathbf{h}^{\mathrm{T}} \cdot \mathbf{x}$ these transformations are cancelling each other.

> ▷ The term *covariant transformation* shouldn't be confused against the term *covariant-form equation*. The former is referring to a specific kind of transformation under a change of basis, the latter is referring to the fact that the equation's form is preserved (the equation looks the same) under a change of basis.

Rewriting (4.17) in the component form we obtain

$$A'_{kl} = \sum_{i,j} A_{ij} T_{jl} T'_{ki}$$

We can immediately notice that the first index of A_{ij} participates in the contravariant transformation, while the second index of A_{ij} participates in the covariant transformation. We can also notice that covariant transformation seems to correspond to the "argument-receiving end" of a linear function: in each case of $h(\vec{\cdot})$ and $\vec{A}(\vec{\cdot})$, a covariant transformation is applied to the index corresponding to the function's argument. Conversely, contravariant transformation seems to be each time applied to the "vector-valued outcome": previously it was applied to the only index of the vector $\vec{x}$, this time it is being applied to the index corresponding to the vector value of the function $\vec{A}(\vec{\cdot})$. Apparently, the covariant transformation, corresponding to the function's argument, will result in cancellation of $\mathbf{T}$ and $\mathbf{T}^{-1}$ in the dot product $\mathbf{A} \cdot \mathbf{x}$, similarly how it happens in $\mathbf{h}^{\mathrm{T}} \cdot \mathbf{x}$:

$$\mathbf{A}' \cdot \mathbf{x}' = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} \cdot \mathbf{T}^{-1}\mathbf{x} = \mathbf{T}^{-1} \cdot \mathbf{A}\mathbf{x}$$

Likewise, the contravariant transformation, applied to the function value, will result in cancellation of $\mathbf{T}$ and $\mathbf{T}^{-1}$ in the dot product $\vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{A}\mathbf{x}$, similarly to how it happens in $\vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{x}$:

$$\vec{\mathbf{e}}'^{\,\mathrm{T}} \cdot \mathbf{A}'\mathbf{x}' = \vec{\mathbf{e}}^{\,\mathrm{T}}\mathbf{T} \cdot (\mathbf{T}^{-1}\mathbf{A}\mathbf{T} \cdot \mathbf{T}^{-1}\mathbf{x}) = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{A}\mathbf{x}$$

Likewise, rewriting (4.19) in the component form yields

$$B'_{kl} = \sum_{i,j} B_{ij} T_{ik} T_{jl}$$

where we notice that covariant transformations are being applied to both indices of B_{ij}, corresponding to the two arguments of the function $B(\vec{\cdot}, \vec{\cdot})$. These transformations will be cancelled in the respective dot products with the function's arguments:

$$\mathbf{x}'^{\mathrm{T}} \cdot \mathbf{B}' \cdot \mathbf{y}' = (\mathbf{T}^{-1}\mathbf{x})^{\mathrm{T}} \cdot \mathbf{T}^{\mathrm{T}}\mathbf{B}\mathbf{T} \cdot \mathbf{T}^{-1}\mathbf{y} = \mathbf{x}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}})^{-1} \cdot \mathbf{T}^{\mathrm{T}}\mathbf{B}\mathbf{T} \cdot \mathbf{T}^{-1}\mathbf{y} = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{B} \cdot \mathbf{y}$$

▷ The multiplication by $\mathbf{T}^{\mathrm{T}}$ from the left in (4.19) is also a covariant transformation. It is exactly the same transformation as multiplication by $\mathbf{T}$ from the right, just written differently due to the limitations of matrix notation. But if really wanted, these limitations can also be bypassed by rewriting (4.19) as

$$\mathbf{B}' = \mathbf{T}^{\mathrm{T}}\mathbf{B}\mathbf{T} = \Big((\mathbf{T}^{\mathrm{T}} \cdot \mathbf{B}\mathbf{T})^{\mathrm{T}}\Big)^{\mathrm{T}} = \Big((\mathbf{B}\mathbf{T})^{\mathrm{T}} \cdot \mathbf{T}\Big)^{\mathrm{T}}$$

where we have two multiplications by $\mathbf{T}$ from the right.

We will now start referring to indices participating in covariant transformations (emerging in case of a basis change) as *covariant indices* and indices participating in contravariant transformations as *contravariant indices*.

4.6 Tensor index notation

We now introduce the *tensor index notation* which means that all covariant indices are to be written as subscripts and all contravariant indices are to be written as superscripts. Thus, we are going to write $\vec{e}_i$, x^i, h_i, A^i_j and B_{ij} respectively. Using this index placement we also have

$$\text{by (4.6):} \qquad \vec{x} = \vec{\mathbf{e}}^{\,\mathrm{T}} \cdot \mathbf{x} = \sum_i \vec{e}_i x^i \tag{4.25a}$$

$$\text{by (4.22):} \qquad y = h(\vec{x}) = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{x} = \sum_i h_i x^i \tag{4.25b}$$

$$\text{by (4.16):} \qquad \vec{y} = \vec{A}(\vec{x}) \iff \mathbf{y} = \mathbf{A} \cdot \mathbf{x} \iff y^j = \sum_i A^j_i x^i \tag{4.25c}$$

$$\text{by (4.18):} \qquad u = B(\vec{x}, \vec{y}) = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{B} \cdot \mathbf{y} = \sum_i \sum_j B_{ij} x^i y^j \tag{4.25d}$$

In (4.25) we can notice that usually contravariant indices correspond to matrix row indices and covariant indices correspond to matrix column indices. The only exception is the row index of B_{ij} which is also covariant, but it's happening due to the limitations of matrix notation, we are having two covariant indices here, but only one of those can correspond to a column index.

▷ We mentioned that matrix notation is, generally speaking, merely a concise form of writing certain kinds of expression and per se doesn't imply any specific semantics. In that sense row indices do not have to have contravariant semantics, nor column indices do have to have covariant semantics. The rule "row indices are contravariant, column indices are covariant" is not really a rule, but rather a convention, which also cannot even be held 100% of the time.

We are going to refer to the basis-independent linear entities denoted by x^i, h_i, A^i_j and B_{ij} as *tensors*. Notice that in this case x^i, h_i, A^i_j and B_{ij} refer to the entire entities rather than to their individual components. At the same time, in formulas like (4.25) the same notation refers to the individual components of the tensors in a particular implied basis. The basis-independent entity $\vec{e}_i$ is formally treated like a tensor, but, since its components are vectors rather than numbers, it's not really a tensor in the full sense.

We can also notice that in (4.25) each summation is always done over a pair of a covariant index and a contravariant index (denoted by the same letter). By

our earlier discussion, this shouldn't be surprising. Such summations correspond to dot products between two matrices (also explicitly written in (4.25)), where the additional $\mathbf{T}$ and $\mathbf{T}^{-1}$ matrices arising in case of basis transformation cancel each other, therefore such dot products themselves are unaffected by the basis change (these dot products are invariant to the basis change).

We will now introduce *Einstein summation convention*, which is commonly used together with tensor index notation. Einstein summation convention drops the explicit summation signs, the summation being implicitly understood each time the same index letter is occurring once as a superscript, once as a subscript:

(Einstein summation convention examples)

$$\vec{x} = \vec{e}_i x^i \tag{4.26a}$$

$$y = h_i x^i \tag{4.26b}$$

$$y^j = A_i^j x^i \tag{4.26c}$$

$$u = B_{ij} x^i y^j \tag{4.26d}$$

The summation implicitly denoted by this kind of index repetition is referred to as *contraction of tensor indices*. E.g. in (4.26b) the subscript i of h_i is contracted with the superscript i of x^i. By our previous discussion, contraction of tensor indices is an operation which is invariant to the basis change, due to the mutual cancellation of transformation matrices $\mathbf{T}$ and $\mathbf{T}^{-1}$.

▷ Any other situations of index letter repetitions (within one expression) are forbidden in Einstein summation convention. In particular it is not allowed to have the same letter occurring more than two times (e.g. $a_{ii}b^i$) or having the same letter appearing in the same position, e.g. both times as a subscript or both times as a superscript. If necessary, explicit deviations from Einstein summation convention need to be made (e.g. by using explicit summation) to allow such index repetitions.

▷ For the time being we will allow index contraction only between two indices belonging to two different tensors. Technically, there is nothing stopping us from also contracting the indices of one and the same tensor, such as a_i^i (after all, it's just an implicit summation). However such contraction would no longer correspond to a dot product of two matrices, where $\mathbf{T}$ and $\mathbf{T}^{-1}$ would be inserted into the middle of the dot product upon a basis change. Thus we don't know yet, whether such tensor index contraction would be still a *covariant* operation, that is whether it produces a basis-independent entity as its result. The same question arises if we contract more than one index pair between two tensors e.g. $a^{ij}b_{ij}$. These cases of tensor index contraction will be addressed in Section 4.11.

Conversely, for the time being we will only allow to multiply tensors together, if there is also a contraction between their indices. Tensor products such as e.g. $a^i b_j$ are not allowed yet, because we don't even know what that means. This will be dealt with in Section 4.10.

We would now like to rewrite the transformations (4.20) plus the transformation (4.23) in tensor index notation. This also requires to sort the indices of T_{ij} and T'_{ij} according to their co- or contra-variance. We can notice that all of (4.20) and (4.23) can be formally written in tensor index notation if T_{ij} is rewritten as T_j^i and T'_{ij} is rewritten as T'^i_j, that is each time the row index turns into a superscript and column index turns into a subscript:

(transformation laws under a change of basis, written in tensor index notation)

$$\vec{e}_i{}' = \vec{e}_j T_i^j \tag{4.27a}$$

$$h_i' = h_j T_i^j \tag{4.27b}$$

$$x'^i = x^j T'^i_j \tag{4.27c}$$

$$A_l'^k = A_j^i T'^k_i T_l^j \tag{4.27d}$$

$$B_{kl}' = B_{ij} T_k^i T_l^j \tag{4.27e}$$

The entities T_j^i and T'^i_j are not really tensors, because their components cannot be written "in a given basis". Being essentially the transformation matrices, these entities refer to two different bases at once. Respectively, even though T_j^i and T'^i_j look pretty much the same as the operator A_j^i, it doesn't make sense to try to write the transformation law like (4.27d) for either of them.

▷ Actually the transformation law (4.27d) can be written for T_j^i, but the contravariant transformation of the index i will be done with respect to the "source basis of T_j^i" while the covariant transformation of the index j will be done with respect to the "destination basis of T_j^i". A similar consideration holds for T'^i_j, the index j referring to the "source basis" and the index i referring to the "destination basis".

Let's now review the so far introduced features of tensor index notation.

- The component indices of a tensor are written either as subscripts or as superscripts, depending on whether they are transforming by the co- or the contra-variant law. We can also use sub-/superscript tensor index notation for indices which do not exactly transform by the co-/contravariant laws, but behave as co-/contravariant indices in all other aspects (the tensor-like entities T_j^i and T'^i_k are one example of that).

- The order of factors within one term doesn't matter, e.g. $h_i x^i$ and $x^i h_i$ are equivalent notations.

- Within boundaries of one term, each index letter should be used exactly once, except where indices are contracted. E.g. the term $a_i b_i$ is incorrect.

- Each index letter implicitly and independently takes all possible values from 1 to N (where N is the number of space dimensions). E.g. the equation $A_i^j x^i + 2y^j = z^j$ is understood in the sense that the equality holds for all possible values of the uncontracted contravariant index j, which is present in all 3 terms $A_i^j x^i$, $2y^j$ and z^j, while the implicit summation over the contracted pair of indices i also goes over all possible values of i.

- Index contraction occurs between one superscript and one subscript index denoted by the same letter. If there is more than one contracted pair of indices within one term, the implied summations can be done in any order, since this order doesn't affect the result.

- Mismatched sub-/superscript placement of indices across different terms of the same sum or of the same equation are forbidden, because transformation laws do not match. E.g. the result of the sum $x^i + y_i$ does neither transform co-variantly nor contra-variantly, and respectively we even cannot denote this result as $z^i = x^i + y_i$ or $z_i = x^i + y_i$. Similarly, an equation $x^i = y_i$ is not a covariant-form equation, since it cannot hold in all bases due to mismatched transformation laws.

- Mismatched indices across terms of sums and/or equations are forbidden for the same reason. E.g. the results of the sums $x^i + y_j$ and $x^i + A^j_k$ neither transform co-variantly nor contra-variantly.

- If necessary, we can also use indices which are exempt from tensor index notation rules, provided we are explicitly aware of such special indices and of the implications of their usage. E.g. instead of a single tensor A^i_j we might want to consider a generic sequence of tensors A^i_{jk} where the letter k is used to index the sequence elements, in which case the index k is neither co- nor contra-variant.

So, we have introduced the concept of tensors and also introduced tensor index notation, but what did we achieve by that? Actually we achieved quite a bit:

- The transformation laws of tensors are immediately visible from their notation (superscript indices transform contravariantly, subscript indices transform covariantly). As we have seen, matrix notation is not 100% capable of achieving that, as the rule "rows transform covariantly, columns transform contravariantly" cannot hold 100% of the time.

- Tensor index contraction is guaranteed to produce a basis-invariant result (tensor), while not every dot product of two matrices is basis-invariant.

- Tensor notation can handle the case of more than 2 indices, while matrix notation is restricted to the maximum of 2 indices.[2]

The total number of tensor indices is referred to as the tensor's *rank*. Thereby x^i and h_i are tensors of rank 1, while A^i_j and B_{ij} are tensors of rank 2. Scalars (basis-independent numbers), having no indices, are tensors of rank 0.

Even though x^i and h_i have the same rank, they are clearly different tensors and are not compatible to each other (they cannot be compared to each other or added to each other). The same holds for A^i_j and B_{ij}. A more specific qualification than the tensor rank is tensor type, which explicitly lists the numbers of contra- and co-variant indices (in that order). So a vector x^i is a $(1,0)$-type tensor, a linear form h_u is a $(0,1)$-type tensor, a linear operator A^i_j is a $(1,1)$-type tensor and a bilinear form B_{ij} is a $(0,2)$-type tensor. By the previous discussion we can realize that the number of covariant indices of a tensor is equal to the number of vector arguments of a function, that the tensor represents. The number of contravariant indices is respectively 0 if the function is scalar-valued and 1 if the function is vector valued. Tensors with more than one contravariant index will be discussed in Section 4.10.

[2]There is an extension of matrix notation which in principle allows to have more than 2 indices, but it's rather unwieldy.

So the operations that we introduced for tensors produce other tensors. These operations so far include index contraction across two tensors within one term, as well as addition and subtraction of tensors and multiplication of a tensor by a scalar. This leads to one of the most important properties and one one the most important reasons why we introduced tensor notation: *equations written in tensor notation are automatically in covariant form.*

> ▷ Tensor equations will be automatically in covariant form only provided they contain exclusively tensors and not other tensor-like entities, like the transformation matrices T^i_j and T'^i_j or the covariant Kronecker delta symbol δ_{ij} introduced in Section 4.8.

4.7 Tensor components

We shall now try to get a little bit more insight into what the tensor components are equal to.

Goal 4.4. Given a linear form $h(\vec{x})$ and a basis $\{\vec{e}_i\}$, find out the relationship between $h(\vec{x})$ and its components in this given basis. *Hint:* use the linearity of $h(\vec{x})$.

We are going to be using the same technique that we have been using in Chapter 3 in the derivation of the dot product formula (3.31) as well as in (3.33), which we have been using for the derivation of the rotation matrix components. Expanding $\vec{x}$ in the basis $\{\vec{e}_i\}$ and substituting into $h(\vec{x})$ yields

$$\begin{aligned} h(\vec{x}) &= h(x^1\vec{e}_1 + x^2\vec{e}_2 + x^3\vec{e}_3) = x^1 h(\vec{e}_1) + x^2 h(\vec{e}_2) + x^3 h(\vec{e}_3) \\ &= x^1 h_1 + x^2 h_2 + x^3 h_3 \end{aligned}$$

(where we used the linearity of $h(\vec{\cdot})$). Or, in Einstein summation convention

$$h(\vec{x}) = h(x^i\vec{e}_i) = x^i h(\vec{e}_i) = x^i h_i$$

Or, in matrix form

$$h(\vec{x}) = \begin{bmatrix} h(\vec{e}_1) & h(\vec{e}_2) & h(\vec{e}_3) \end{bmatrix} \cdot \begin{bmatrix} x^1 \\ x^2 \\ x^3 \end{bmatrix} = \begin{bmatrix} h_1 & h_2 & h_3 \end{bmatrix} \cdot \begin{bmatrix} x^1 \\ x^2 \\ x^3 \end{bmatrix} = \mathbf{h}^{\mathrm{T}} \cdot \mathbf{x}$$

Therefore

(linear form components)

$$h_i = h(\vec{e}_i) \tag{4.28}$$

Goal 4.4 completed. ■

Goal 4.5. Given a bilinear form $B(\vec{x}, \vec{y})$ and a basis $\{\vec{e}_i\}$, find out the relationship between $B(\vec{x}, \vec{y})$ and its components in this given basis. *Hint:* do the same as for $h(\vec{x})$.

Using Einstein summation convention, by the linearity of $B(\vec{\cdot},\vec{\cdot})$ in each of the arguments,

$$B(\vec{x},\vec{y}) = B(x^i\vec{e}_i, y^j\vec{e}_j) = x^i y^j B(\vec{e}_i,\vec{e}_j) = x^i y^j B_{ij}$$

therefore

(bilinear form components)

$$B_{ij} = B(\vec{e}_i,\vec{e}_j) \tag{4.29}$$

Goal 4.5 completed. ■

Goal 4.6. Given a linear operator $\vec{A}(\vec{x})$ and a basis $\{\vec{e}_i\}$, find out the relationship between $\vec{A}(\vec{x})$ and its components in this given basis. *Hint:* do the same as for $h(\vec{x})$, but mind that the function is now vector-valued.

By the linearity of $\vec{A}(\vec{\cdot})$,

$$\begin{aligned}\vec{y} = \vec{A}(\vec{x}) = \vec{A}(x^i\vec{e}_i) = x^i\vec{A}(\vec{e}_i) &= x^1\vec{A}(\vec{e}_1) + x_2\vec{A}(\vec{e}_2) + x_3\vec{A}(\vec{e}_3)\\ &= \begin{bmatrix}\vec{A}(\vec{e}_1) & \vec{A}(\vec{e}_2) & \vec{A}(\vec{e}_3)\end{bmatrix}\cdot\begin{bmatrix}x^1\\x^2\\x^3\end{bmatrix}\end{aligned}$$

Now let a_1^j be the components of $\vec{A}(\vec{e}_1)$ (in the basis $\{\vec{e}_i\}$), a_2^j be the components of $\vec{A}(\vec{e}_2)$ and a_3^j be the components of $\vec{A}(\vec{e}_3)$, that is

(linear operator components)

$$\vec{A}(\vec{e}_i) = a_i^j\vec{e}_j \tag{4.30}$$

It follows that

$$\vec{y} = x^i\vec{A}(\vec{e}_i) = x^i a_i^j\vec{e}_j$$

and respectively

$$y^j = x^i a_i^j$$

It follows that a_i^j given by (4.30) (that is a_i^j being the components of the vector $\vec{A}(\vec{e}_i)$ for each given i) are the components of $\vec{A}(\vec{\cdot})$.

▷ In the matrix form we are having $\mathbf{y} = \mathbf{A}\cdot\mathbf{x}$ written in the component form as

$$\begin{bmatrix}y^1\\y^2\\y^3\end{bmatrix} = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3\end{bmatrix}\cdot\begin{bmatrix}x^1\\x^2\\x^3\end{bmatrix}$$

where $\mathbf{a}_i$ are column matrices representing the components of $\vec{A}(\vec{e}_i)$ in the basis $\{\vec{e}_i\}$.

Goal 4.6 completed. ■

4.8 Kronecker delta symbol

In matrix notation $\mathbf{I}$ represents an identity matrix. It might be useful to have a notational counterpart of that in the tensor index notation. Before attempting to find this counterpart, we can realize that the notation $\mathbf{I}$ doesn't specify co- and contra-variance properties of its indices. The rank of the corresponding tensor must be 2, but the tensor type can be $(2,0)$, $(1,1)$ or $(0,2)$.

Let's start with the $(1,1)$-type tensor, representing a linear operator. The corresponding matrix equation is $\mathbf{y} = \mathbf{I} \cdot \mathbf{x} = \mathbf{x}$. Denoting this tensor as δ^i_j we are having

$$y^i = \delta^i_j x^j$$

where the components of the tensor δ^i_j are identical to the components of the identity matrix $\mathbf{I}$:

(Kronecker delta tensor, $(1,1)$***-type)***

$$\delta^i_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{4.31}$$

Similarly, for a $(0,2)$-type tensor counterpart of $\mathbf{I}$ we are having a matrix equation $u = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{I} \cdot \mathbf{y}$, which we can write in tensor form as

$$u = \delta_{ij} x^i y^j$$

where

(Kronecker delta symbol, $(0,2)$***-type)***

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{4.32}$$

There is, however, one problem with δ_{ij}. Recalling the transformation law of the bilinear forms (4.20d) (or, in tensor form, (4.27e)) we should realize that in another basis the components of δ_{ij} will not necessarily be given by (4.32) anymore, since applying (4.20d) to an identity matrix doesn't necessarily produce an identity matrix:

$$\mathbf{T}^{\mathrm{T}} \cdot \mathbf{I} \cdot \mathbf{T} = \mathbf{T}^{\mathrm{T}} \cdot \mathbf{T} \neq \mathbf{I}$$

▷ If $\mathbf{T}$ is an orthogonal matrix, $\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T} = \mathbf{I}$, therefore the form (4.32) of δ_{ij} will be preserved by orthogonal transformations between bases.

It follows that the entity δ_{ij} with components given by (4.32) (in the sense that (4.32) is supposed to hold in any basis) is not a tensor, since its components do not transform by the covariant law (4.27e). For that reason δ_{ij} is considered to be a tensor-like symbol, rather than a tensor. More specifically, δ_{ij} is referred to as *Kronecker delta symbol*.

On the other hand, the linear operator form δ^i_j of the Kronecker delta symbol given by (4.31) is a true tensor. Indeed, upon a basis change the components

of a linear operator transform by the similarity transformation law (4.20c) (or, in tensor form, (4.27d)), and applying (4.20c) to an identity matrix $\mathbf{I}$ yields

$$\mathbf{T}^{-1} \cdot \mathbf{I} \cdot \mathbf{T} = \mathbf{T}^{-1} \cdot \mathbf{T} = \mathbf{I}$$

or, in tensor form, applying (4.27d) to δ^i_j yields

$$\delta^i_j T'^k_i T^j_l = \delta^i_j T'^k_i \cdot T^j_l = T'^k_j \cdot T^j_l = \delta^k_l$$

The $(1,1)$-type Kronecker delta symbol δ^i_j is thereby a true tensor.

The $(2,0)$-type Kronecker delta symbol is given by

(Kronecker delta symbol, $(2,0)$***-type)***

$$\delta^{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{4.33}$$

Similarly to δ_{ij}, it's not difficult to realize that δ^{ij} is not a tensor either.

Now that we have a tensor notation counterpart of the identity matrix let's take a look at its usage example.

Goal 4.7. Rewrite (4.24) in tensor index notation, using Kronecker delta symbol.

Taking (4.24) and writing it in tensor index notation we have

$$h'_i \cdot x'^i = h_j T^j_i \cdot x^k T'^i_k = h_j x^k \cdot T^j_i T'^i_k = h_j x^k \cdot \delta^j_k = h_j \delta^j_k \cdot x^k = h_k x^k$$

where we used (4.27b) and (4.27c).

Goal 4.7 completed. ■

4.9 Metric tensor

4.9.1 Dot product

Consider a dot product of two vectors $\vec{x} \cdot \vec{y}$. Commonly the dot product of vectors is computed as the dot product of their matrix representations $\mathbf{x}^{\mathrm{T}} \cdot \mathbf{y}$, let's find a tensor notation counterpart of that.

The rule "row indices are contravariant, column indices are covariant" suggests that the row matrix $\mathbf{x}^{\mathrm{T}}$ (within which the column index takes values from 1 to N where N is the number of dimensions) should be written as x_i, while the column matrix $\mathbf{y}$ (where the row index takes different values) should be written as y^i and the expression $\mathbf{x}^{\mathrm{T}} \cdot \mathbf{y}$ would be written as $x_i y^i$. There is a problem with that: the notation x_i implies that x is a linear form (whose components are transforming by the covariant law), while actually $\vec{x}$ is a vector and its components are transforming by the contravariant law. We cannot simply replace the correct notation x^i with x_i.

> ▷ The other way to look at this is to notice that the result of the expression $x_i y^i$ is not going to be basis-invariant (be a scalar), if the components of x_i transform by contra- rather than co-variant law, since we are going to have two contravariant transformations, which cannot mutually cancel in the contraction.

▷ This is one large benefit of tensor notation: we can immediately notice that the expression $x_i y^i$ is incorrect due to incorrect index placement on x (where x, being a vector, must be indexed by a superscript). Similarly, we can immediately notice that the expression $x^i y^i$ is incorrect, because it violates the index repetition rule of tensor index notation. The actual meaning of the formal incorrectness the expressions $x_i y^i$ and $x^i y^i$ is that they are not covariant-form expressions, therefore their results are not basis-invariant.

Apparently, the source of the problem is that the expression $\mathbf{x}^{\mathrm{T}} \cdot \mathbf{y}$, as a general formula for the dot product of two vectors, is incorrect. Let's quickly obtain a correct formula for $\vec{x} \cdot \vec{y}$.

Goal 4.8. Obtain a correct formula for the dot product of two vectors.

We can redo the steps taken in the derivation of the formula (3.31), where this time we don't make assumptions about the basis being orthonormal. We have (using tensor index notation)

(dot product of two vectors in tensor notation)

$$\vec{x} \cdot \vec{y} = x^i \vec{e}_i \cdot y^j \vec{e}_j = x^i y^j \cdot (\vec{e}_i \cdot \vec{e}_j) = x^i y^j \cdot g_{ij} = g_{ij} x^i y^j \tag{4.34a}$$

where

$$g_{ij} = \vec{e}_i \cdot \vec{e}_j \tag{4.34b}$$

The tensor g_{ij} given by (4.34b) is known as the *metric tensor*. Since dot product of two vectors commutes, we have $\vec{e}_i \cdot \vec{e}_j = \vec{e}_j \cdot \vec{e}_i$ and $g_{ij} = g_{ji}$, that is g_{ij} is a *symmetric tensor*.

Thereby the dot product of two vectors is given by the bilinear form (4.34a) built on the tensor g_{ij}. The components of this bilinear form are $\vec{e}_i \cdot \vec{e}_j$. In the matrix form g_{ij} corresponds to the *Gram matrix*

$$\mathbf{G} = \begin{bmatrix} \vec{e}_1 \cdot \vec{e}_1 & \vec{e}_1 \cdot \vec{e}_2 & \vec{e}_1 \cdot \vec{e}_3 \\ \vec{e}_2 \cdot \vec{e}_1 & \vec{e}_2 \cdot \vec{e}_2 & \vec{e}_2 \cdot \vec{e}_3 \\ \vec{e}_3 \cdot \vec{e}_1 & \vec{e}_3 \cdot \vec{e}_2 & \vec{e}_3 \cdot \vec{e}_3 \end{bmatrix} \tag{4.34c}$$

where we have

$$\vec{x} \cdot \vec{y} = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{G} \cdot \mathbf{y} \tag{4.34d}$$

as a generic formula for the dot product of two vectors.

In an orthonormal basis we have $g_{ij} = \delta_{ij}$ and $\mathbf{G} = \mathbf{I}$.

▷ It might be more appropriate to refer to the dot product in this context (especially with a non-identity matrix $\mathbf{G}$) as the *inner product* and notate it as $\langle \vec{x}, \vec{y} \rangle$. However the term "inner product" intuitively might feel too abstract, as in principle it can denote lots of different bilinear operations, which can thereby be relatively randomly chosen. To highlight the intuitive geometric interpretation of the operation $\vec{x} \cdot \vec{y}$ as a product of lengths and a cosine (rather than highlighting a range of abstract possibilities), we will keep using the term "dot product" and the notation $\vec{x} \cdot \vec{y}$ here.

Goal 4.8 completed. ■

4.9.2 Index lowering

The previous discussion suggests that matrix transposition is not really a "correct" operation. E.g. transposing a column vector $\mathbf{x}$ into the row vector $\mathbf{x}^{\mathrm{T}}$ in tensor notation looks like replacing x^i with x_i, which changes the expected transformation properties of the tensor x. Similarly, transposing a linear operator matrix $\mathbf{A}$ in tensor notation looks like swapping the covariant properties of the first and second indices of the tensor A from $A^i{}_j$ into $A_i{}^j$.

> ▷ We left some space before the index j in $A^i{}_j$ and $A_i{}^j$ to explicitly highlight that j is the second index. Most of the time, when we don't manipulate the covariant properties of the indices, the same indices always stay in the same positions and we can simply write A^i_j, since we can always tell, which index is which. However in a situation, where we are moving the indices around vertically, we must rely on the horizontal ordering of the indices, therefore we need an extra space before j to make sure it is positioned to the right of i. The letters themselves can be freely chosen (e.g. we could have written $A^k{}_l$) and cannot be used as a means to identify specific indices of a given tensor.

Strictly speaking, matrix transposition per se cannot be "incorrect", since matrix notation per se doesn't assume any particular semantics. However, with respect to the rule "row indices are contravariant, column indices are covariant", matrix transposition is an incorrect operation, since it changes the covariant properties of the index or indices. From this perspective, we can ask the question: "is there a "correct" way to transpose a matrix?" In application to the dot product of two vector $\vec{x} \cdot \vec{y}$, our question would mean "is there a correct way to "transpose" the column matrix $\mathbf{x}$ to obtain the "correct" counterpart of $\mathbf{x}^{\mathrm{T}} \cdot \mathbf{y}$?"

Goal 4.9. Define "a correct way to transpose a column vector", so that the dot product $\vec{x} \cdot \vec{y}$ can be written as a dot product of a row matrix (obtained from the column matrix $\mathbf{x}$) with the column matrix $\mathbf{y}$.

Starting with (4.34a) we can rewrite its right-hand side as

$$\vec{x} \cdot \vec{y} = g_{ij} x^i y^j = g_{ij} x^i \cdot y^j = h_j y^j$$

where

$$h_j = g_{ij} x^i$$

The expression $h_i y^j$ can be written as $\mathbf{h}^{\mathrm{T}} \cdot \mathbf{y}$ since $h(\vec{\cdot})$ is represented by a row matrix $\mathbf{h}^{\mathrm{T}}$. Noticing that h_i is pretty much the x_i which we wanted to use a little bit earlier in $x_i y^i$ (but didn't know how to do that without breaking the covariance of the expression), we are going to introduce notation

(tensor index lowering)

$$x_j = g_{ij} x^i \tag{4.35}$$

for the linear form $x(\vec{\cdot})$ obtained from the vector $\vec{x}$. Respectively we have

$$\vec{x} \cdot \vec{y} = x(\vec{y}) = x_i y^i$$

that is we represented the dot product of $\vec{x}$ with $\vec{y}$ as application of the linear form $x(\vec{\cdot})$ to the vector $\vec{y}$. The transformation of a contravariant index into a

covariant one by multiplication with the metric tensor, such as we did in (4.35), is referred to as *index lowering*.

▷ Since $g_{ij} = g_{ji}$, we do not need to be particularly careful with which of the two indices of g_{ij} we pick for index contraction. We can also write (4.35) as

$$x_i = g_{ij} x^j$$

In an orthonormal basis, $g_{ij} = \delta_{ij}$, and respectively (4.35) turns into $x_j = \delta_{ij} x^i = x^j$ or simply $x_j = x^j$, this means that we can simply lower (or raise) the indices at will. However the equation $x_j = x^j$ is not in a covariant form (which can be clearly seen from mismatched index positioning on the two sides of the equation) and thus doesn't hold in all bases.

This suggests that one can freely transpose column (and row) vectors in an orthonormal basis. In a nonorthonormal basis one would likely need to involve the Gram matrix in such transformation.

Goal 4.9 completed. ■

4.9.3 Index rasing

If there is the operation of index lowering (4.35), we can also define its inverse, which would be the *index raising* operation.

Goal 4.10. Construct the index raising operation, which is the inverse of index lowering.

Rewriting $h_i = g_{ij} x^j$ in the matrix form we obtain

$$\mathbf{h}^{\mathrm{T}} = (\mathbf{G} \cdot \mathbf{x})^{\mathrm{T}} = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{G}^{\mathrm{T}} = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{G}$$

where $\mathbf{G}$ is the Gram matrix and where we used the fact that $\mathbf{G} = \mathbf{G}^{\mathrm{T}}$. Respectively the inverse operation is achieved by multiplication by $\mathbf{G}^{-1}$:

$$\mathbf{x} = \left(\mathbf{h}^{\mathrm{T}} \cdot \mathbf{G}^{-1}\right)^{\mathrm{T}} = \left(\mathbf{G}^{-1}\right)^{\mathrm{T}} \cdot \left(\mathbf{h}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{G}^{-1} \cdot \left(\mathbf{h}^{\mathrm{T}}\right)^{\mathrm{T}}$$

(where we noticed that, since $\mathbf{G}$ is symmetric, so is $\mathbf{G}^{-1}$). In tensor notation the latter formula can be written as

(tensor index raising)

$$x^j = g^{ij} x_i \tag{4.36}$$

where the components of the entity denoted as g^{ij} are identical to the components of $\mathbf{G}^{-1}$.

▷ Thereby g_{ij} and g^{ij} denote two different entities, which can be told apart by the placement of their indices.

Since $\mathbf{G}$, being the matrix of the bilinear form $\vec{x} \cdot \vec{y} = \mathbf{x}^{\mathrm{T}} \mathbf{G} \mathbf{y}$, transforms by (4.19) as $\mathbf{G}' = \mathbf{T}^{\mathrm{T}} \mathbf{G} \mathbf{T}$, its inverse $\mathbf{G}^{-1}$ must transform by

$$\mathbf{G}'^{-1} = \mathbf{T}^{-1} \mathbf{G}^{-1} (\mathbf{T}^{-1})^{\mathrm{T}}$$

It's not too difficult to guess that, if we rewrite $\mathbf{T}^{-1}\mathbf{G}^{-1}(\mathbf{T}^{-1})^{\mathrm{T}}$ in the tensor form, we are going to obtain two contravariant transformations for each of the indices of g^{ij} (the readers may do this as an exercise, if they wish). Therefore g^{ij} is a $(2,0)$-type tensor. It is referred to as *inverse metric tensor* and can be used for index raising as in (4.36).

Goal 4.10 completed. ■

Index raising and lowering can be also applied to tensors of ranks higher than 1, where we need to choose which of the indices we want to raise or lower, applying contraction with the metric tensor or inverse metric tensor to this chosen index. As a somewhat artificial example, we could attempt to raise one of the indices of the metric tensor by evaluating the expression $g^{ki}g_{ij}$. Since the components of g^{ki} and g_{ij} are represented by mutually inverse matrices,

(contraction of metric tensor with inverse metric tensor)

$$g^{ki}g_{ij} = \delta^k_j$$

If we continue by raising the other index we are going to have

$$g^{lj}\delta^k_j = g^{lk}$$

that is by raising both of the indices of g_{ij} we obtain g^{ij}. Thereby the notation of the inverse metric tensor as g^{ij} is consistent with the index raising/lowering convention, which expects that identically named tensors differing only by sub-/superscript placement of some indices can be obtained from each other by index raising/lowering.

Conversely, by lowering both indices of g^{ij} we obtain g_{ij}. In that sense, the δ^i_j tensor appearing "midway" during this raising or lowering can be also denoted as g^i_j.

4.10 Tensor product

Let x^i and y^i be two vectors. In principle we can write their product $a^{ij} = x^i y^j$ with two independent indices i and j. Being uncontracted, both indices run independently from 1 to N (where N is the number of space dimensions), therefore the components of a^{ij} form an $N \times N$ matrix.

Since both x^i and y^i transform accordingly to the contravariant law (4.27c), the product a^{ij} transforms as

$$a'^{ij} = x'^i y'^j \overset{(4.27c)}{=} x^k T'^i_k \cdot y^l T'^j_l = x^k y^l \cdot T'^i_k T'^j_l = a^{kl} \cdot T'^i_k T'^j_l$$

(notice the similarities to the bilinear form transformation law (4.27e), except that this time both transformations are contravariant). That is both indices of a^{ij} transform according to the contravariant law and a^{ij} is a $(2,0)$-type tensor.

> ▷ Differently from tensors with multiple covariant indices, which represent multilinear functions of vector arguments, there is no simple semantics for tensors with multiple covariant indices. They can be, however, understood as "joint vector arguments" of multilinear functions. E.g. a tensor a^{ij} can be contracted in both indices with a bilinear form b_{ij} producing a scalar value $u = a^{ij}b_{ij}$.

This allows more possibilities to specify bilinear form arguments. Indeed, usually we just have $B(\vec{x}, \vec{y}) = b_{ij}x^i y^j$, which corresponds to $b_{ij}a^{ij}$ where $a^{ij} = x^i y^j$. However we could for example let $a^{ij} = x^i y^j + v^i w^j$, in which case the product $b_{ij}a^{ij}$ cannot be expressed as a bilinear form of two separate arguments, these arguments have been joined or "melted" together. This is closely related to the idea of *decomposable tensors*, discussed in Section 4.12.

We should also remind the reader that contraction of more than one index pair between two tensors (such as in $b_{ij}a^{ij}$ where we are having contractions both in the i indices and in the j indices) is not really allowed yet, as we only demonstrated the covariance of contraction of a single index pair between two tensors, but this is going to be addressed in Section 4.11.

Similarly, given two linear forms x_i and y_j we can build a product $a_{ij} = x_i y_j$. Since both x_i and y_i transform according to the covariant law (4.27b), the product a_{ij} transforms as

$$a'_{ij} = x'_i y'_j \overset{(4.27b)}{=} x_k T^k_i \cdot y_l T^l_j = x_k y_l \cdot T^k_i T^l_j = a_{kl} \cdot T^k_i T^l_j$$

that is both indices of a^{ij} transform according to the covariant law, and the whole a_{ij} transforms by (4.27e), which means that a_{ij} is a $(0, 2)$-type tensor, that is a bilinear form.

Similarly, given a vector x^i and a linear form y_j we can build a product $a^i_j = x^i y_j$. Since both x_i transforms by the contravariant law (4.27c) and and y_j transforms according to the covariant law (4.27b), the product a^i_j transforms as

$$a'^i_j = x'^i y'_j = x^k T'^i_k \cdot y_l T^l_j = x^k y_l \cdot T'^i_k T^l_j = a^k_l \cdot T'^i_k T^l_j$$

that is the superscript index of a^i_j transforms according to the contravariant law, and the subscript index of a^i_j transforms according to the covariant law, and the whole a^i_j transforms by (4.27d), which means that a^i_j is a $(1, 1)$-type tensor, that is a linear operator.

The above examples should convince us that uncontracted tensor products such as $a^{ij} = x^i y^j$, $a_{ij} = x_i y_j$ and $a^i_j = x^i y_j$ produce tensors and therefore should be a valid tensor operation. Apparently, the idea of an uncontracted tensor product is not limited to products of tensors of rank 1, but can be applied to tensors of any rank, e.g. $A^i_j B_{kl} = C^i_{jkl}$.

An uncontracted product of tensors is referred to simply as a *tensor product*. We have already seen the matrix notation counterpart of a tensor product of two rank 1 tensors in the beginning of this chapter, which is the vector outer product (4.2).

In "indexless notation" the tensor product is denoted using the "$\otimes$" symbol. E.g. $x^i y^j$ can be respectively written as $\vec{x} \otimes \vec{y}$.

▷ While in tensor index notation the tensor product "commutes" in the sense that $x^i y^j = y^j x^i$, the "indexless" tensor product doesn't commute, since this time we do not have index letters which we can use to match the indices. Instead we have to rely on the ordering of the indices, knowing which index is the first one and which is the second one. Therefore changing the order changes the result and

$$\vec{x} \otimes \vec{y} \neq \vec{y} \otimes \vec{x}$$

which in the tensor index notation can be expressed as

$$x^i y^j \neq y^i x^j$$

It is not difficult to realize that tensor product is bilinear and associative.

4.11 Matrix trace

If tensor notation counterparts of an $N \times N$ matrix $\mathbf{A}$ are tensors a^{ij}, a_{ij} and a^i_j, then the counterpart of the matrix trace $\operatorname{tr} \mathbf{A}$ should be the contraction of the tensor indices a^i_i, which is thereby applicable only to $(1,1)$-type tensors. Earlier in this chapter we, however, noted that we allow index contraction only between two indices belonging to two different tensors. This restriction was solely due to the fact that we were not sure whether contraction between indices of the same tensor is a covariant (basis-independent) operation. Since a covariant tensor contraction should produce a scalar, we need to demonstrate that

$$a'^i_i = a^i_i \tag{4.37}$$

where a'^i_i is obtained from a^i_i by the basis change transformation law (4.27d).

▷ Since index contraction "cancels" the indices, they don't need to be matched between different terms, and we can use different letters in the left- and right-hand sides of (4.37). E.g. we could write $a'^i_i = a^j_j$ instead.

In matrix form the transformation law (4.27d) looks like a similarity transformation of the operator matrix $\mathbf{A}$, respectively the matrix notation counterpart of (4.37) is

(matrix trace is invariant under similarity transformation)

$$\operatorname{tr} \mathbf{A}' = \operatorname{tr}(\mathbf{T}^{-1}\mathbf{A}\mathbf{T}) = \operatorname{tr} \mathbf{A} \tag{4.38}$$

that is trace of an operator matrix is invariant under basis change.

Goal 4.11. Demonstrate (4.37), that is that contraction of two indices of one and the same tensor (of type $(1,1)$) is a covariant operation. Notice that (4.38) will follow automatically. *Hint:* Start with the $(1,1)$-type tensors representable as a tensor product of two rank 1 tensors.

Suppose $a^i_j = x^i y_j$. Then $a^i_i = x^i y_i$. But we already know that $x^i y_i$ is a scalar (a basis-invariant value), and therefore so is a^i_i.

▷ In matrix notation terms the same is expressible as

$$\operatorname{tr}(\mathbf{x} \cdot \mathbf{y}^{\mathrm{T}}) = \sum_i x^i y_i = \mathbf{y}^{\mathrm{T}} \cdot \mathbf{x}$$

that is dot product is the trace of the outer product.

Now suppose a^i_j is not representable by a tensor product $x^i y_j$. Then, it should be still representable as a sum or a linear combination of such products:

$$a^i_j = \sum_\nu \lambda_\nu x^i_\nu y_{\nu j} \tag{4.39}$$

where the index ν is not a part of tensor index notation. This index also doesn't run over the number of dimensions of the linear space, instead running over the number of terms of the linear combination.

It is probably simplest to let each of the tensors x^i_ν and $y_{\nu j}$ have all components equal to zero except for a single component, which is equal to 1, so that their product $x^i_\nu y_{\nu j}$ will also have all components equal to zero, except a single component emerging as the product of the equal to 1 components of x^i_ν and $y_{\nu j}$. E.g., assuming 2 dimensions for simplicity, we can have

$$x^i_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad y_{1j} = \begin{bmatrix} 1 & 0 \end{bmatrix} \qquad x^i_1 y_{1j} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

In the same fashion

$$x^i_2 y_{2j} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad x^i_3 y_{3j} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \qquad x^i_4 y_{4j} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

and the coefficients λ_ν of the linear combination (4.39) simply should be chosen equal to the values of the components of a^i_j.

Applying index contraction to the left-hand side of (4.39), we notice that this contraction distributes across the terms of the linear combination:

$$a^i_i = \sum_\nu \lambda_\nu x^i_\nu y_{\nu i}$$

or, writing the summation explicitly

$$\sum_i a^i_i = \sum_i \left(\sum_\nu \lambda_\nu x^i_\nu y_{\nu i} \right) = \sum_\nu \lambda_\nu \left(\sum_i x^i_\nu y_{\nu i} \right)$$

and since the contractions $x^i_\nu y_{\nu i}$ produce scalars (basis-independent values), the linear combination of these scalar values is also a scalar. That is the contraction a^i_i produces a scalar. We thereby demonstrated (4.37) and its matrix-notation version (4.38).

Goal 4.11 completed. ■

We now might conjecture that the previous result generalizes to higher-rank tensors (provided they still have at least one co- and at least one contra-variant index). We are not going to demonstrate it in a general form, but provide a (hopefully convincing) example.

Goal 4.12. Provide an example, illustrating that contraction of two indices of one and the same tensor of a rank higher than 2 is also a covariant operation. *Hint:* Use the same approach as for $(1,1)$-type tensors.

Suppose we have a tensor a^i_{jk} and we perform the contraction $b_k = a^i_{ik}$. We need to demonstrate that b_k transforms as a linear form.

Similarly to (4.39), let's expand a^i_{jk} into a linear combination of products:

$$a^i_{jk} = \sum_\nu \lambda_\nu x^i_\nu y_{\nu j} z_{\nu k} \tag{4.40}$$

Performing contraction,

$$b_k = a^i_{ik} = \sum_\nu \lambda_\nu x^i_\nu y_{\nu i} z_{\nu k} = \sum_\nu (\lambda_\nu x^i_\nu y_{\nu i}) z_{\nu k}$$

we notice that the result is a linear combination of linear forms $z_{\nu k}$ (indexed by ν), the linear combination coefficients being the scalar values $\lambda_\nu x^i_\nu y_{\nu i}$. Thus, the result is also a linear form.

Goal 4.12 completed. ■

It follows that contracting more than one index pair across the same pair of tensors is also a covariant operation. We are also going to provide a (hopefully convincing) example demonstration.

Goal 4.13. Provide an example, illustrating that contracting more than one index pair across the same pair of tensors is a covariant operation.

Let $u = b_{ij}a^{ij}$. We can represent this expression by first contracting only one pair of indices, obtaining some intermediate tensor $c^k_j = b_{ij}a^{ik}$, and then contracting the indices of this intermediate tensor: $u = c^j_j$. By the previous results, each of these two contractions is a covariant operation and therefore so is their combination $u = b_{ij}a^{ij}$.

Goal 4.13 completed. ■

> ▷ The just obtained results suggest that there is no difference between the contraction happening across two different tensors and the contraction happening within one tensor. E.g. instead of contracting $x^i y_i$ we could first compute $z^i_j = x^i y_j$ and then perform the contraction z^i_i.
>
> This in turn suggests a more uniform understanding of contraction as always happening within one tensor. That is we shall understand $x^i y_i$ as first computing an intermediate tensor $z^i_j = x^i y_j$ and then contracting its indices. The same idea also applies to tensors of other types and higher-rank tensors.

We have already seen that index contraction distributes across the terms of a linear combination. We also know that two different contractions commute (since the implicit summations over two different indices can be done in any order). The covariance of index contraction now allows us to more freely use index contraction in the expressions which are not solely employing tensor index notation, but also contain index-abstract tensor entities, e.g. vectors notated as $\vec{x}$, or linear forms notated as $h(\vec{\cdot})$, where we are now able not only to use contraction as a formal summation shortcut, but also make some basis-abstract transformations around the contraction operation.

E.g. suppose we wish to obtain the formula for the components of a vector in a non-orthogonal basis. Starting from $\vec{x} = x^i \vec{e}_i$ we dot-multiply both sides by $\vec{e}_j$ obtaining

$$\vec{x} \cdot \vec{e}_j = (x^i \vec{e}_i) \cdot \vec{e}_j$$

Since contraction is a covariant (basis-independent) operation, we can delay it until a later point in time by changing the position of the parentheses in the right-hand side of the above equation ("apply the associativity of multiplication"):

$$\vec{x} \cdot \vec{e}_j = x^i \cdot (\vec{e}_i \cdot \vec{e}_j) = x^i g_{ij}$$

Multiplying both sides by the inverse metric tensor and contracting again, we obtain

$$(\vec{x} \cdot \vec{e}_j) \cdot g^{jk} = x^i g_{ij} g^{jk} = x^i \delta^k_i = x^k$$

that is

(computation of the components of $\vec{x}$ in arbitrary basis)

$$x^k = (\vec{x} \cdot \vec{e}_j) \cdot g^{jk} \tag{4.41}$$

Formula (4.41) provides a way to compute the components x^k of $\vec{x}$ in any basis. Notice that if the basis is singular, the inverse metric tensor g^{jk} doesn't exist and the formula is not applicable, corresponding to the fact that an expansion of an arbitrary vector in a singular basis doesn't exist.

4.12 Decomposable tensors

Recall how in (4.39) we have been expanding a $(1,1)$-type tensor into a linear combination of tensor products. Intuitively we realized that not every $(1,1)$-type tensor is representable as a tensor product of a vector and a linear form. Let's look at this in more detail. We also found a guaranteed way to represent a $(1,1)$-type tensor by a linear combination of N^2 tensor products (where N is the dimension of the space), however it feels like we could do with a sum of a smaller number of tensor products.

Goal 4.14. Provide an example of a rank 2 tensor which is not representable by a single tensor product of a vector with a linear form. *Hint:* look at the matrix rank of the tensor component matrix.

Suppose $a^i_j = x^i y_j$. Then

$$a^i_j = \begin{bmatrix} x^1y_1 & x^1y_2 & x^1y_3 \\ x^2y_1 & x^2y_2 & x^3y_3 \\ x^3y_1 & x^3y_2 & x^3y_3 \end{bmatrix}$$

We can notice that all columns of the matrix are multiples of $\mathbf{x} = [x^1\ x^2\ x^3]^{\mathrm{T}}$ and all rows of the matrix are multiples of $[y_1\ y_2\ y_3]$. That is all columns of the matrix are linearly dependent and all rows of the matrix are linearly dependent. Therefore the determinant of the matrix is zero. Furthermore, the rank[3] of the matrix is 1.

Similar considerations apply to any other rank 2 tensor. Therefore we just found that a rank 2 tensor, representable as a product of two rank 1 tensors, has a matrix representation of rank 1. It follows that any rank 2 tensors, whose matrix representation has matrix rank higher than 1, will not be representable in the form $a^i_j = x^i y_j$. E.g. the matrix representation of tensor δ^i_j is the identity matrix $\mathbf{I}$. The rank of the identity matrix is equal to the number of space dimensions, therefore δ^i_j is not representable as a product of a vector and a linear form.

Goal 4.14 completed. ■

Tensors representable as a product of rank 1 tensors are called *decomposable tensors*. We have just found out that decomposable tensors of rank 2 have matrix representations of matrix rank 1. It is an interesting question, whether any rank 2 tensor which has matrix representation of rank 1 is decomposable.

[3] Here we mean the matrix rank, not the rank of a tensor, which is a very different concept.

Goal 4.15. Find out whether any rank 2 tensor, which has matrix representation of rank 1, is decomposable.

Let $\mathbf{A}$ be the matrix of components of a rank 2 tensor (in some basis) and suppose $\operatorname{rank}\mathbf{A} = 1$. Matrix rank 1 means that all rows of $\mathbf{A}$ are multiples of each other (and so are its columns). Let $\mathbf{y}^{\mathrm{T}}$ denote one of the rows of $\mathbf{A}$. Without loss of generality we can assume $\mathbf{y}^{\mathrm{T}} \neq 0$ (unless the entire matrix is zero, in which case it's trivially decomposable). Then the n-th row of $\mathbf{A}$ can be written as $x^n\mathbf{y}^{\mathrm{T}}$, where x^n is some number. Combining all x^n into a column matrix $\mathbf{x}$ we obtain

$$\mathbf{A} = \mathbf{x} \cdot \mathbf{y}^{\mathrm{T}}$$

that is $\mathbf{A}$ is an outer product of $\mathbf{x}$ and $\mathbf{y}$, which is the matrix notation counterpart of a tensor product. Thus, the corresponding tensor is decomposable.

We thereby found that a rank 2 tensor is decomposable iff its matrix representation has rank 1.

Goal 4.15 completed. ■

The next question we could ask is: "what is the minimum number of tensor product terms needed to represent a tensor?"

Goal 4.16. Find out a way to determine the minimum number of tensor product terms needed to represent a rank 2 tensor. *Hint:* First study the case of the maximum possible number of needed terms.

We already earlier described a way to split an $N \times N$ matrix into a linear combination of N^2 outer product terms. It's also not difficult to realize how we can split any $N \times N$ matrix into N outer products. E.g. in 3 dimensions

$$\begin{bmatrix} a^1_1 & a^1_2 & a^1_3 \\ a^2_1 & a^2_2 & a^2_3 \\ a^3_1 & a^3_2 & a^3_3 \end{bmatrix} = \begin{bmatrix} a^1_1 \\ a^2_1 \\ a^3_1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} a^1_2 \\ a^2_2 \\ a^3_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} a^1_3 \\ a^2_3 \\ a^3_3 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \quad (4.42)$$

where we effectively split the matrix into N columns. Instead we also could split it into N rows in a similar manner. This suggests that N terms would be sufficient for any $N \times N$ matrix. But can we still do with less terms?

Suppose the rightmost matrix column in (4.42) consists solely of zeros. Then, respectively, we can drop the third term in the right-hand side of (4.42), so we need only 2 terms. The next step would be to notice that the rank of such matrix will be 2 rather than 3 (or even less than 2, but we assume this is not the case) and to realize that the matrix rank is invariant under basis transformations, more precisely it is invariant under similarity transformations of the matrix. Indeed, assuming the matrix has linear operator semantics, the dimension of the image of this operator (which is equal to the matrix rank) is basis-independent. Thus a tensor represented by a 3×3 rank 2 matrix in one basis will also be represented by a 3×3 rank 2 matrix in all other bases.

Conversely, given a 3×3 rank 2 matrix we can always by a similarity transformation convert it to a matrix with a third column consisting of all zeros (by interpreting the similarity transformation as a basis change and choosing the third basis vector within the nullspace of the linear operator). In this basis (4.42) can use just 2 terms.

▷ We could also do the argument without a switch of basis. A rank 2 matrix has two linearly independent columns, the third column being a linear combination of the two. Without loss of generality we can assume $\mathbf{a}_3 = \lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2$, where $\mathbf{a}_n$ is the n-th column of the matrix. Then (4.42) can be rewritten as

$$\begin{bmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ a_1^3 & a_2^3 & a_3^3 \end{bmatrix} = \begin{bmatrix} a_1^1 \\ a_1^2 \\ a_1^3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & \lambda_1 \end{bmatrix} + \begin{bmatrix} a_2^1 \\ a_2^2 \\ a_2^3 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & \lambda_2 \end{bmatrix}$$

This rank 2 matrix example should be convincing us that generally it suffices to have the number of outer product terms equal to the rank of the matrix. But can we do even less than that?

We already know that a rank of a matrix, which is representable as an outer product of two vectors, is equal to 1. We might conjecture that the rank of a matrix representable as a sum of M outer products of two vectors is equal to M. Although, we probably need to correct that to the rank being less than or equal to M (e.g. if some of the terms are identical or otherwise linearly dependent):

$$\operatorname{rank}\left(\sum_{i=1}^{M} \mathbf{x}_i\mathbf{y}_i^{\mathrm{T}}\right) \leq M \tag{4.43}$$

To demonstrate that the rank cannot exceed M we can apply a similarity transformation to (4.43),[4] converting to the basis where the first M basis vectors coincide with $\mathbf{x}_i$. Then the matrix $\sum \mathbf{x}_i\mathbf{y}_i^{\mathrm{T}}$ will have the first M rows nonzero (potentially), while all following rows will be zero, and respectively the rank of the matrix cannot exceed M. E.g. in 3 dimensions for $M = 2$ we will have

$$\sum_{i=1}^{2} \mathbf{x}_i\mathbf{y}_i^{\mathrm{T}} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} y_{11} & y_{12} & y_{13} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} y_{21} & y_{22} & y_{23} \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

So, we have found that the rank of a sum of M outer products cannot exceed M. This means that less than M terms won't do to represent a matrix of rank M. We also just found a bit earlier that M terms are sufficient to represent a matrix of rank M. Thus, the minimum possible number of outer product terms to represent a matrix of rank M is M. Respectively, the minimum possible number of tensor product terms to represent a rank 2 tensor is equal to the rank of the matrix of tensor components.

Goal 4.16 completed. ■

We won't be looking into decomposability of tensors of higher ranks.

4.13 Dual basis

Assuming the basis $\vec{e}_i$, any vector $\vec{x}$ representable by a single-column matrix $\mathbf{x} = \begin{bmatrix} x^1 & x^2 & x^3 \end{bmatrix}^{\mathrm{T}}$ can be expanded in that basis as

$$\vec{x} = x^1\vec{e}_1 + x^2\vec{e}_2 + x^3\vec{e}_3 = x^i\vec{e}_i$$

[4] More precisely, we apply a contravariant transformation to $\mathbf{x}_i$ and a covariant transformation to $\mathbf{y}_i^{\mathrm{T}}$.

The basis vectors $\vec{e}_i$ in the component form can be represented by the column matrices

$$\mathbf{e}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix} \qquad \mathbf{e}_2 = \begin{bmatrix}0\\1\\0\end{bmatrix} \qquad \mathbf{e}_3 = \begin{bmatrix}0\\0\\1\end{bmatrix} \tag{4.44}$$

Then we could imagine that there is also the basis for the linear forms, consisting of the basis forms $f^i(\vec{\cdot})$, whose components are represented by row matrices

$$\mathbf{f}^{1\mathrm{T}} = \begin{bmatrix}1 & 0 & 0\end{bmatrix} \qquad \mathbf{f}^{2\mathrm{T}} = \begin{bmatrix}0 & 1 & 0\end{bmatrix} \qquad \mathbf{f}^{3\mathrm{T}} = \begin{bmatrix}0 & 0 & 1\end{bmatrix} \tag{4.45}$$

▷ The reason why we put the basis linear form index as a superscript in $f^i(\vec{\cdot})$ and $\mathbf{f}^i$ is the same one why the basis vector index is put as a subscript in $\vec{e}_i$ and $\mathbf{e}_i$: later we will want to contract that superscript index with the subscript index of the linear form components.

The idea is that any linear form $h(\vec{\cdot})$ representable by a row matrix $\mathbf{h}^\mathrm{T} = \begin{bmatrix}h_1 & h_2 & h_3\end{bmatrix}$ can be expanded in that basis as

$$h(\vec{\cdot}) = h_1 f^1(\vec{\cdot}) + h_2 f^2(\vec{\cdot}) + h_3 f^3(\vec{\cdot}) = h_i f^i(\vec{\cdot}) \tag{4.46}$$

which in matrix form simply corresponds to

$$\begin{aligned}\mathbf{h}^\mathrm{T} &= \begin{bmatrix}h_1 & h_2 & h_3\end{bmatrix} = h_1 \cdot \begin{bmatrix}1 & 0 & 0\end{bmatrix} + h_2 \cdot \begin{bmatrix}0 & 1 & 0\end{bmatrix} + h_3 \cdot \begin{bmatrix}0 & 0 & 1\end{bmatrix}\\ &= h_1 \cdot \mathbf{f}^{1\mathrm{T}} + h_2 \cdot \mathbf{f}^{2\mathrm{T}} + h_3 \cdot \mathbf{f}^{3\mathrm{T}} = h_i \cdot \mathbf{f}^{i\mathrm{T}}\end{aligned} \tag{4.47}$$

So, $f^i(\vec{\cdot})$ are indeed basis linear forms.

It would be nice to be able to describe these linear forms in abstract terms, without resorting to their component representation.

Goal 4.17. Construct a component-free, abstract definition of $f^i(\vec{\cdot})$. *Hint:* describe their relationships with $\vec{e}_i$.

It is easily noticed by direct evaluation that

$$f^i(\vec{e}_j) = \delta^i_j \tag{4.48}$$

E.g.

$$f^1(\vec{e}_1) = \begin{bmatrix}1 & 0 & 0\end{bmatrix} \cdot \begin{bmatrix}1\\0\\0\end{bmatrix} = 1 + 0 + 0 = 1$$

By the linearity of linear forms it follows that

$$\begin{aligned}f^i(\vec{x}) &= f^i(x^1\vec{e}_1 + x^2\vec{e}_2 + x^3\vec{e}_3) = x^1 f^i(\vec{e}_1) + x^2 f^i(\vec{e}_2) + x^3 f^i(\vec{e}_3)\\ &= x^j f^i(\vec{e}_j) = x^j \delta^i_j = x^i\end{aligned} \tag{4.49}$$

Intuitively, $f^i(\vec{\cdot})$ simply selects the i-th component from its argument.

Now, apparently (4.49) uniquely identifies $f^i(\vec{\cdot})$, but (4.49) is in turn uniquely defined by (4.48), therefore (4.48) uniquely identifies $f^i(\vec{\cdot})$.

Goal 4.17 completed. ■

By (4.46), the linear forms $f^i(\vec{\cdot})$ serve as the basis for all other linear forms. Linear forms thereby form their own linear space with the basis $f^i(\vec{\cdot})$. As in any

linear space, the basis choice is not unique, and there exist infinitely many other choices. However the specific basis $f^i(\vec{\cdot})$ that we just constructed is related to the specific basis $\vec{e}_i$ of our primary linear space of vectors via (4.48). We say that the basis $f^i(\vec{\cdot})$ is the dual basis of $\vec{e}_i$. We also refer to the space of the linear forms as the dual space of our primary linear space of vectors.

Conversely, the linear space of vectors is the dual of the space of linear forms. In principle those two spaces are equivalent and neither is "better than the other". Their duality is established by the relationships (4.48), which thereby (using the linearity property) allow to combine any element of one space with any element of the other one, producing a scalar value:

$$h(\vec{x}) = h_i f^i(\vec{x}) = h_i f^i(x^j \vec{e}_j) = h_i x^j f^i(\vec{e}_j) = h_i x^j \delta^i_j = h_i x^i$$

So, it's only that we have chosen one of these two linear spaces as our primary linear space (by originally starting considering vectors in that space), which makes one of those two spaces our primary one, there is no other distinction. The asymmetry of notation between $\vec{x}$ and $h(\vec{\cdot})$ is also solely due to that distinction, and in principle we could have introduced some other, more symmetric, notation for combining these two entities. E.g. we could have notated $\vec{x}$ as x and $h(\vec{\cdot})$ as h and notate $h(\vec{x})$ as $x \cdot h$ instead.

▷ Similarly to how we realized that linear forms also form a linear space (dual to the primary one), tensors of any given type also form a linear space. Implicitly we already demonstrated this in Section 4.11 for rank 2 tensors by providing an expansion of a rank 2 tensor into a sum of N^2 tensor products (N being the dimension of the primary linear space), where all components of each tensor product were zero except a single unity component. It is not difficult to realize that a basis for such tensor spaces can be given by tensor products of $\vec{e}_i$ and $f^i(\vec{\cdot})$, where there would be a factor of $\vec{e}_i$ for each contravariant index and a factor of $f^i(\vec{\cdot})$ for each covariant index. E.g. the linear space of $(1,1)$-type tensors admits $\vec{e}_i \otimes f^j(\vec{\cdot})$ as a basis.

4.14 Differentiation

Gradients are vectors pointing in the direction of the fastest growth of the function. The gradient formula is

$$\nabla\varphi = \begin{bmatrix} \dfrac{\partial\varphi}{\partial x^1} & \dfrac{\partial\varphi}{\partial x^2} & \dfrac{\partial\varphi}{\partial x^3} \end{bmatrix}^{\mathrm{T}} \tag{4.50}$$

or so we have usually learned. Let's see how correct is that knowledge.

Let's perform a simple basis transformation. Say we grow all basis vectors by the factor 2:

$$\vec{e}_i{}' = 2\vec{e}_i \tag{4.51a}$$

Vector components would change contravariantly:

$$x'^i = \frac{x^i}{2} \tag{4.51b}$$

so that

$$\vec{v} = x'^i \vec{e}_i{}' = \frac{x^i}{2} \cdot 2\vec{e}_i = x^i \vec{e}_i$$

is unchanged. That should include the gradient vectors, since those are vectors, and, as all other vectors, should change contravariantly, right?

Let's substitute (4.51b) into (4.50):

$$\frac{\partial\varphi}{\partial x'^i} = \frac{\partial\varphi}{\partial(x^i/2)} = 2\frac{\partial\varphi}{\partial x^i}$$

It seems gradient's components have varied by the covariant law. So, gradient is not a vector? How is the direction of the fastest growth not a vector? To answer that question, let's first obtain the transformation law of the components of the gradient given by (4.50) in case of a general basis transformation.

Goal 4.18. Find the transformation law of the components of the gradient given by (4.50).

Using the chain differentiation rule yields

$$\frac{\partial\varphi}{\partial x'^i} = \frac{\partial\varphi}{\partial x^1}\frac{\partial x^1}{\partial x'^i} + \frac{\partial\varphi}{\partial x^2}\frac{\partial x^2}{\partial x'^i} + \frac{\partial\varphi}{\partial x^3}\frac{\partial x^3}{\partial x'^i} = \sum_j \frac{\partial\varphi}{\partial x^j}\frac{\partial x^j}{\partial x'^i} \tag{4.52}$$

Now by the contravariant transformation law (4.14), which in tensor index notation reads $x^j = x'^k T^j_k$, we have

$$\frac{\partial x^j}{\partial x'^i} = \frac{\partial}{\partial x'^i}(x'^k T^j_k) = \frac{\partial}{\partial x'^i}\left(\sum_k x'^k T^j_k\right) = \sum_k T^j_k \frac{\partial x'^k}{\partial x'^i} \tag{4.53}$$

where we switched to the explicit summation, since we're not sure yet how Einstein's implicit summation convention should behave with the regards to the indices in the denominators of the derivative operations. Even though the contraction is done with the index in the numerator, we chose to write the summation explicitly to avoid potential confusion. What is, however, obvious is that the derivatives $\partial x'^k/\partial x'^i$ in the right-hand side of (4.53) yield 1 if $k = i$ and 0 otherwise, so they are equal to δ_{ki}, of δ^k_i or δ^{ki}, whatever the index placement for δ is. Two considerations help to establish correct index placement. One is that the quantities $\partial x'^k/\partial x'^i$ are apparently the same in each basis, and we might remember that of the three tensor-like symbols δ_{ki}, of δ^k_i or δ^{ki} only δ^k_i behaves as a tensor. Secondly, by choosing the symbol δ^k_i we can use Einstein's summation convention, rewriting the right-hand side of (4.53) further as

$$\sum_k T^j_k \frac{\partial x'^k}{\partial x'^i} = \sum_k T^j_k \delta^k_i = T^j_k \delta^k_i = T^j_i \tag{4.54}$$

Thus we have established that

$$\frac{\partial x'^k}{\partial x'^i} = \delta^k_i \tag{4.55}$$

Substituting (4.54) into (4.53) we obtain

$$\frac{\partial x^j}{\partial x'^i} = \sum_k T^j_k \frac{\partial x'^k}{\partial x'^i} = T^j_i \tag{4.56}$$

and substituting (4.56) further into (4.52) yields

$$\frac{\partial\varphi}{\partial x'^i} = \sum_j \frac{\partial\varphi}{\partial x^j}\frac{\partial x^j}{\partial x'^i} = \sum_j \frac{\partial\varphi}{\partial x^j} T_i^j \tag{4.57}$$

This looks like the covariant transformation law, except that we had to use explicit summation. Actually, we could notice that in (4.54) treating the superscript index i in the differentiation operator's denominator $\partial x'^i$ as if it was a usual subscript index in δ_i^k yielded the possibility to use the implicit summation convention. We further reestablished this in (4.55), where the superscript i in the denominator of the left-hand side is matched against the subscript i in the right-hand side. Along the same lines, if in (4.57) we treat the index j in the differentiation operator's denominator ∂x^j as if it was a usual subscript index "in the numerator", this would also allow us to use implicit summation convention. We also have generally established by (4.57) itself that the result of differentiation by vector's components transforms by the covariant law. Therefore a superscript in the denominator (at least in the denominator of a differentiation's operator) should be treated as if it was a subscript in the numerator, and we can use the implicit summation convention rewriting (4.57) as

$$\frac{\partial\varphi}{\partial x'^i} = \frac{\partial\varphi}{\partial x^j} T_i^j \tag{4.58}$$

▷ The fact that differentiation by x^i produces a covariant index motivates the shorthand notation for the partial differentiation:

$$\partial_i \triangleq \frac{\partial}{\partial x^i} \tag{4.59}$$

e.g.

$$\partial_i\varphi = \frac{\partial\varphi}{\partial x^i}$$

Unfortunately we couldn't use (4.59) in (4.58), since we need to simultaneously express differentiation with respect to x^j and x'^i in the same formula. Although, in principle we might still write (4.58) as $\partial'_i\varphi = T_j^i\partial_j\varphi$ and even generalize it to equating the differentiation operators $\partial'_i = T_j^i\partial_j$.

Goal 4.18 completed. ■

We have thereby just established that the gradient given by (4.50) transforms covariantly, thus it is a linear form rather than a vector and we, in particular, should use the row matrix instead of column matrix in (4.50). But what about the interpretation of the gradient as the direction of the fastest growth of the function? The directions are expressed by vectors, not by linear forms.

First, we shall establish a reinterpretation of gradient as a linear form. Let $\varphi = \varphi(\vec{x})$ and let $\vec{x}$ be a function of t: $\vec{x} = \vec{x}(t)$. Combining the functions $\varphi(\vec{x})$ and $\vec{x}(t)$ we obtain the function $\varphi(t) = \varphi(\vec{x}(t))$, or, in coordinate component form

$$\varphi = \varphi\big(x^1(t), x^2(t), x^3(t)\big)$$

By the differentiation chain rule

$$\varphi'(t) = \frac{\mathrm{d}\varphi}{\mathrm{d}t} = \frac{\partial\varphi}{\partial x^1}\frac{\mathrm{d}x^1}{\mathrm{d}t} + \frac{\partial\varphi}{\partial x^2}\frac{\mathrm{d}x^2}{\mathrm{d}t} + \frac{\partial\varphi}{\partial x^3}\frac{\mathrm{d}x^3}{\mathrm{d}t} = \frac{\partial\varphi}{\partial x^i}\frac{\mathrm{d}x^i}{\mathrm{d}t} = \frac{\mathrm{d}x^i}{\mathrm{d}t}\partial_i\varphi \tag{4.60}$$

where we switched to the implicit summation convention again, contracting the superscript i in the derivative's denominator with the superscript i in the numerator.

▷ The prime notation in $\varphi'(t)$ for the differentiation in (4.60) shouldn't cause any confusion with the usage of prime to denote the second basis. Since bases are irrelevant for scalars, we wouldn't need to distinguish between different φ in different bases.

Similarly, we are about to denote the vector-valued derivative $\mathrm{d}x^i/\mathrm{d}t$ as $\vec{x}\,'$. This shouldn't be causing any confusion for the same reason: bases are irrelevant for vector notation and we don't need to distinguish between different $\vec{x}$ in different bases.

Rewriting (4.60) back in vector notation terms we write $\mathrm{d}x^i/\mathrm{d}t$ as $\vec{x}\,'$ and introduce the linear form $h(\vec{\cdot})$ with components $h_i = \partial_i\varphi = \partial\varphi/\partial x^i$ for the gradient, yielding

$$\varphi'(t) = h_i \frac{\mathrm{d}x^i}{\mathrm{d}t} = h(\vec{x}\,') \tag{4.61}$$

Equation (4.61) interprets the gradient as a linear form $h(\vec{\cdot})$, which can be used to obtain the speed of change of the function $\varphi(\vec{x})$, given the speed of change of $\vec{x}$.

If we want to reinterpret the linear form $h(\vec{\cdot})$ as a vector giving the direction of the fastest growth, we need to reinterpret the linear form application $h(\vec{x}\,')$ as a dot product of two vectors $\vec{h}\cdot\vec{x}\,'$. In order to do so, we need to raise the index of the linear form h_i, thereby obtaining its contravariant counterpart

$$h^j = h_i g^{ij} = g^{ij}\partial_i\varphi \tag{4.62}$$

which is the "vector version" of the gradient. Thus, the original formula (4.50) delivers a vector only in orthonormal bases, while a more universal formula for $\nabla\varphi$ is given by (4.62).

Now that we have established that differentiation with respect to x^i produces a covariant index, we could expect that the divergence of a vector field $\vec{v} = \vec{v}(\vec{x})$ given by

$$\nabla\cdot v = \frac{\partial v^1}{\partial x^1} + \frac{\partial v^2}{\partial x^2} + \frac{\partial v^3}{\partial x^3} = \frac{\partial v^i}{\partial x^i} = \partial_i v^i \tag{4.63}$$

(using the implicit summation again) produces a scalar due to contraction of the two indices i in the right-hand side of (4.63). Let's see if that's the case.

Goal 4.19. Verify that (4.63) gives a scalar (basis-independent) value.

Switching the basis and applying the chain differentiation rule we obtain

$$\begin{aligned}\nabla\cdot v &= \frac{\partial v'^i}{\partial x'^i} = \frac{\partial v'^i}{\partial x^j}\frac{\partial x^j}{\partial x^i} = \frac{\partial v'^i}{\partial x^j}T^j_i = \frac{\partial(T'^i_k v^k)}{\partial x^j}T^j_i = T'^i_k\frac{\partial v^k}{\partial x^j}T^j_i = \frac{\partial v^k}{\partial x^j}\delta^j_k \\ &= \frac{\partial(v^k\delta^j_k)}{\partial x^j} = \frac{\partial v^j}{\partial x^j} = \partial_j v^j\end{aligned}$$

So divergence is indeed a scalar.

Goal 4.19 completed. ■

The previous discussion of gradient and divergence suggests that the nabla operator in a non-orthonormal basis takes the form

(nabla operator in a non-orthonormal basis)

$$\nabla = \vec{e}_j \cdot g^{ij} \partial_i \tag{4.64}$$

Indeed, as we have already seen by (4.62), in a gradient operator, "∇" has to be understood as given by (4.64). We can also verify that (4.64) is compatible with the divergence formula (4.63):

$$\begin{aligned}\nabla \cdot \vec{v} &\overset{(4.64)}{=} \vec{e}_j g^{ij} \partial_i \cdot \vec{v} = \vec{e}_j g^{ij} \partial_i \cdot v^k \vec{e}_k = (\vec{e}_j \cdot \vec{e}_k) \cdot g^{ij} \partial_i \cdot v^k = g_{jk} g^{ij} \partial_i \cdot v^k \\ &= \delta^i_k \partial_i \cdot v^k = \partial_i \cdot v^i\end{aligned}$$

We won't discuss the remaining common applications of the differential operator "∇", which are the field curl $\nabla \times \vec{v}$ and the Laplacian $\nabla^2 \varphi$, in this Chapter. Some hints with respect to the covariant-form aspects of the curl may be however obtained from the material of Chapter 5. Further discussion of the "∇" operator can be found in Section 6.10.

4.15 Linear transformations of space

4.15.1 Equivalence to basis change

Throughout this chapter we have been reiterating the idea that, under a basis change, tensor components are changing by the co- and contravariant laws, which in matrix form are written as (4.10) and (4.15). Let's denote the matrix $\mathbf{T}^{-1}$ in (4.15) as $\mathbf{A}$. Then (4.15) starts to look as

$$\mathbf{v}' = \mathbf{A} \cdot \mathbf{v} \tag{4.65}$$

Now (4.65) looks like some linear transformation $\mathbf{A}$ applied to a vector $\mathbf{v}$. E.g. it could be a rotation. Suppose $\mathbf{A}$ describes a rotation (in principle the following argument holds for any non-singular linear transformation, not just rotations, but it's probably most intuitive to imagine a rotation). If we apply (4.65) to all vectors in the linear space, we thereby transform (rotate) all vectors in that space. And that transformation is described by exactly the same formula as the contravariant transformation (4.15), we only changed the formal symbol for the matrix from $\mathbf{T}^{-1}$ to $\mathbf{A}$.

This is a fundamental feature of basis changes: they are described by the same equations as linear transformations of the space, and the only distinction between those is our interpretation. We could hold the basis in place and rotate all vectors in the linear space, or we could hold all vectors and rotate the basis (but in the opposite direction), the change of the relationship between the vectors and the basis (expressed as the change of the vector component values) will be exactly the same. Fig. 4.1 illustrates. The opposite direction of rotation, depending on whether we choose to rotate the vectors or the basis, corresponds to the matrix equality $\mathbf{T}^{-1} = \mathbf{A}$: one matrix is the inverse of the other. It is also a manifestation of co- and contra-variance being the "opposites" of each other.

Now, if we transform all vectors in the space by the same linear transformation $\vec{A}(\vec{\cdot})$, we are going to thereby preserve all relationships between the

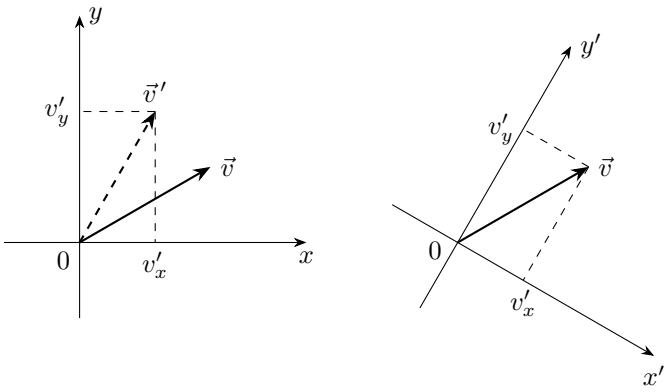


Figure 4.1: Equivalence of space rotation vs. rotation of basis in an opposite direction. The coordinates v'_x, v'_y of the rotated vector $\vec{v}\,'$ in the original coordinate system are the same as the coordinates of the original vector $\vec{v}$ in the coordinate system rotated by the same angle in an opposite direction.

vectors (or so it would seem at the first look). E.g. if we had $\vec{z} = \vec{x} + 2\vec{y}$ then $\vec{z}\,' = \vec{A}(\vec{z}) = \vec{A}(\vec{x} + 2\vec{y}) = \vec{A}(\vec{x}) + 2\vec{A}(\vec{y}) = \vec{x}\,' + 2\vec{y}\,'$. That is, the relationship $\vec{z} = \vec{x} + 2\vec{y}$ is preserved by linear transformations. This makes linear transformations look like basis changes, and, as we just found out, they are even described by exactly the same equations.

> ▷ This results in a rather common pitfall among students of linear algebra (at least in the author's experience) to pretty much ignore the idea of basis change and instead "simply" think of linear transformations of the space (4.65). The latter seem so simple and intuitive, while the idea of basis change often may look pretty artificial and academic. Hopefully this chapter served to reestablish the key importance of the idea of basis change for the reader, in case they have fallen into the same trap earlier.

Of course, more complicated relationships between vectors can be easily broken by a linear transformation. E.g. suppose we had a relationship $\vec{y} = \vec{B}(\vec{x})$, where $\vec{B}(\vec{\cdot})$ is some other linear operator, and now we apply the linear transformation $\vec{A}(\vec{\cdot})$ to the entire linear space. This gives $\vec{y}\,' = \vec{A}(\vec{y}) = \vec{A}(\vec{B}(\vec{x})) \neq \vec{B}(\vec{A}(\vec{x})) = \vec{B}(\vec{x}\,')$, because the linear transformations $\vec{A}(\vec{\cdot})$ and $\vec{B}(\vec{\cdot})$ do not have to commute (in matrix form, $\mathbf{AB} \neq \mathbf{BA}$). Similarly, expressions involving linear or bilinear forms will be broken as well.

By now, after reading through the previous material of this chapter, we should be able to realize that the situation can be easily amended by transforming not only the vectors, but also transforming linear operators, linear forms, bilinear forms, etc., to match the linear space transformation applied to vectors (e.g. we need to apply a similarity transformation to all linear operators), effectively performing the same transformations which we would do in case of a basis change. Thereby we might be able to maintain the equivalence between basis changes and linear transformations of the space (if we really want to). But still, not fully, as we shall see next.

4.15.2 Pseudotensors

Recall the cross product formula. Assuming a positively oriented orthonormal basis (for the sake of definiteness),

$$\vec{z} = \vec{x} \times \vec{y} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \begin{bmatrix} x^2y^3 - x^3y^2 \\ x^3y^1 - x^1y^3 \\ x^1y^2 - x^2y^1 \end{bmatrix} \tag{4.66}$$

We can already spot a number of "suspicious things" about the formula (4.66). First, the fact that the formula needs to assume a particular basis orientation is highly unusual. Secondly, the elements of the column matrix in the right-hand side of (4.66) look like they should be components of a $(2,0)$-type tensor, rather than the ones of a vector. A related fact is that in tensor notation is does not seem even possible to equate the expressions of the form $x^iy^j - x^jy^i$ to vector components z^k.

> ▷ Interestingly enough, the elements of the column matrix in the right-hand side of (4.66) in fact *are* components of a 2nd rank tensor, and also it *is* possible to equate $x^iy^j - x^jy^i$ to vector components z^k in tensor notation (this is discussed in Chapter 5 and, more specifically, Section 5.9). At any rate, the result of the cross product operation is not really a vector, and now we shall see, why.

Let's apply a particular linear transformation to the linear space: invert the direction of all vectors. This transformation can be expressed as $\vec{x}\,' = -\vec{x}\ \forall \vec{x}$. Now, if we apply this transformation to the cross product $\vec{x} \times \vec{y}$ we can notice that the result of the cross product is unchanged by this transformation, since $\vec{x}\,' \times \vec{y}\,' = (-\vec{x}) \times (-\vec{y}) = \vec{x} \times \vec{y}$. But, on the other hand, the vector $\vec{z}$ should have changed its direction: $\vec{z}\,' = -\vec{z}$. Thereby the relationship $\vec{z} = \vec{x} \times \vec{y}$ is broken by such transformation.

We have already been in a similar situation earlier: we noticed that vector relationships which involve linear operators, linear forms, bilinear forms, etc., are also broken by a linear transformation of the space, and we found that we can amend the situation by "adjusting the parameters" of linear operators, linear forms and bilinear forms in a way to match the linear transformation of the space. So, it seems we should apply the same idea here: we need to adjust the cross product, so that it also matches the linear transformation of the space.

Unfortunately such approach doesn't really work here. E.g. if we had some arbitrarily taken linear operator $\vec{A}(\vec{x})$, there would be no problem adjusting this operator after the transformation of the linear space. Similarly, it's not a problem to adjust some arbitrarily taken bilinear form $B(\vec{x}, \vec{y})$ after the transformation of the linear space. We had some bilinear form $B(\vec{x}, \vec{y})$ before the transformation, after the transformation we're going to use some other bilinear form $B'(\vec{x}, \vec{y})$ instead, nothing too unusual about that. The situation is very different with operations like dot or cross product: we expect these operations to be the same before and after the transformation, otherwise they even wouldn't feel like operations defined on vectors, since their definitions would be changing all the time.

Now with the dot product we don't have a problem: the dot product is linear in each of its arguments, so it changes sign upon a sign change of any of its arguments. If both arguments change sign, then the dot product is unchanged. Thereby the dot product is unchanged upon the "$\vec{x}\,' = -\vec{x}$" transformation.

The cross product, like the dot product, is also unchanged upon the "$\vec{x}\,' = -\vec{x}$" transformation, but since the result is now a vector rather than a scalar, we actually did expect the result to change its sign.

Notice that the same problem with the cross product doesn't occur under a basis change. Under a basis change we do not expect the result of the cross product to change, therefore upon inverting the directions of all basis vectors we can simply change the sign in front of the right-hand side of the cross-product formula (4.66), thereby preserving the result (in the geometric sense, as a direction with a length), which is thereby basis-independent:

$$\begin{aligned}\vec{z} = \vec{x} \times \vec{y} &= \begin{bmatrix}\vec{e}_1{}' & \vec{e}_2{}' & \vec{e}_3{}'\end{bmatrix} \cdot \begin{bmatrix} x'^2 y'^3 - x'^3 y'^2 \\ x'^3 y'^1 - x'^1 y'^3 \\ x'^1 y'^2 - x'^2 y'^1 \end{bmatrix} \\ &= (-1) \begin{bmatrix}-\vec{e}_1 & -\vec{e}_2 & -\vec{e}_3\end{bmatrix} \cdot \begin{bmatrix} (-x^2)(-y^3) - (-x^3)(-y^2) \\ (-x^3)(-y^1) - (-x^1)(-y^3) \\ (-x^1)(-y^2) - (-x^2)(-y^1) \end{bmatrix}\end{aligned}$$

It's just that the cross-product has a different expression in a negatively-oriented basis, but this is nothing special: as we have seen, the dot product also has different expressions in different bases.

So, the extra factor -1 is completely fine in (4.66) upon a basis inversion, but it cannot appear upon space inversion. The equivalence between linear space transformations and basis changes is thereby irreparably broken.

Conceptually, this is addressed by referring to the entities, obtained as cross product results, as *pseudovectors* instead of referring to them as vectors.

A similar situation (but kind of in an opposite way) occurs with the triple product of vectors. The triple product is a trilinear operation, therefore upon inverting the directions of all vectors the result changes its sign. But a scalar value is not supposed to change sign in this situation. Therefore the result of the triple product is not a scalar but a *pseudoscalar*.

The same idea can be generalized to other kinds of tensors, thereby we introduce the concept of *pseudotensors*. Apparently, whenever the rank of the result of a product-like operation (or generally of a multilinear combination of several tensors) does not match the sum of the ranks of the factors (or, more exactly, when it differs by an odd number), the same sign mismatch effect is going to appear. There is no way to fix this issue: the sign mismatch emerges directly from the multilinearity property.

▹ Some further discussion of pseudotensors will be found in Section 5.8.

4.16 Tensors and complex numbers

So far we have assumed that tensor components (as well as coefficients of linear combinations) are having real values. But sometimes linear algebra operates on the field of complex numbers. Let's take a brief look at the implications of such change on the tensor algebra.

The tensor algebra is built around the ideas of co- and contravariance, which are closely related to the idea of the dot product. Tensor index contraction is,

intuitively seen, a dot product of a row and a column matrices. We can recall from Section 3.6 that if complex numbers are involved, the dot product involves conjugation of the components of one of the factors, as specified by (3.48). The formula (4.34) should be thereby changed into something like

$$\vec{x} \cdot \vec{y} = g_{ij}\overline{x}^i y^j \tag{4.67}$$

where the bar on top denotes complex conjugation.[5] The components of the metric tensor g_{ij}, being equal to $\vec{e}_i \cdot \vec{e}_j$ would make a Hermitian matrix ($g_{ji} = \overline{g}_{ij}$), rather than a symmetric one, since, for a sesquilinear dot product, swapping the arguments conjugates the result: $\vec{e}_i \cdot \vec{e}_j = \overline{\vec{e}_j \cdot \vec{e}_i}$.

Matrix transposition is generally replaced by matrix Hermitian conjugation in complex linear algebra, therefore not only the metric tensor becomes Hermitian, but also index raising and lowering should conjugate the components, at least for rank 1 tensors, so we should be writing something like $v_i = g_{ij}\overline{v}^j$. Or should it rather be $v_i = \overline{g_{ij}v^j}$, that is shall we conjugate before or after the contraction? Actually, since g_{ij} is Hermitian, it doesn't matter, whether we conjugate before or after, however instead we should watch out for the index order, since g_{ij} is no longer equal to g_{ji}.

Let's therefore do everything carefully again:

$$g_{ij} = \vec{e}_i \cdot \vec{e}_j$$

$$\vec{x} \cdot \vec{y} = (x^i\vec{e}_i) \cdot (y^j\vec{e}_j) = \overline{x}^i y^j \cdot (\vec{e}_i \cdot \vec{e}_j) = g_{ij}\overline{x}^i y^j$$

where we used the sesquilinear property $(\alpha\vec{x}) \cdot \vec{y} = \overline{\alpha} \cdot (\vec{x} \cdot \vec{y})$ (while $\vec{x} \cdot (\beta\vec{y}) = \beta \cdot (\vec{x} \cdot \vec{y})$).

It looks like we could combine $g_{ij}\overline{x}^i$ into x_j thereby obtaining

$$\vec{x} \cdot \vec{y} = g_{ij}\overline{x}^i y^j = (g_{ij}\overline{x}^i) \cdot y^j = x_j y^j$$

thereby

$$x_j = g_{ij}\overline{x}^i = \overline{g_{ji}x^i}$$

We could continue working out further details of tensor algebra in case of a complex field, however we are going to rather stop here as nonorthonormal bases in complex linear spaces are not as common as in real case. On the other hand, under the assumption of an orthonormal (unitary) basis, the interpretation of column matrices as vectors and row matrices as linear forms is reinforced, in which case Hermitian transposition can be used as the means to "raise and lower indices".

Summary of Chapter 4

- Starting from the idea of formulas and entities invariant under basis change (such as vectors) we have developed the ideas of co- and contra-variance.
- We introduced tensor index notation to make the co- and contra-variance properties more explicitly visible.

[5] We don't use the asterisk to denote conjugation here, because it doesn't fit well with tensor index notation.

- We introduced the idea of tensors as linear entities invariant under basis changes and realized that scalars, vectors, linear operators, linear forms and bilinear forms are all examples of tensors.
- We have reviewed a number of formulas and properties established in linear algebra from the perspective of covariance and realized that some of them automatically emerge from co- and contra-variant properties and tensor index notation, including similarity and congruence matrix transformations or the invariance of matrix trace.
- We found out that matrix transposition is usually a hidden form of tensor index raising or lowering.
- We found that there are various entities which only look as if they are tensors, but actually aren't. A particular subclass of those entities is pseudotensors.

Chapter 5

Area and volume

Derivation of signed area of a parallelogram by bilinearity. Exterior product in 2D and in higher dimension counts. Relationship to dot product. Relationship to determinants. Relationship to tensors and pseudotensors.

5.1 Exterior product in 2 dimensions . . . 136
5.2 Symmetries . . . 141
5.3 Exterior vs. dot product . . . 146
5.4 Volume . . . 147
5.5 Bivectors . . . 152
5.5.1 Exterior product in 3D . . . 152
5.5.2 Bivector space in 3D . . . 157
5.5.3 Bivectors in 2D . . . 161
5.5.4 Dot product of bivectors . . . 162
5.5.5 Higher dimension counts . . . 165
5.6 k-vectors . . . 167
5.6.1 Associative wedge product in 3D . . . 167
5.6.2 3-vector space . . . 169
5.6.3 k-vector space . . . 170
5.7 Volume deformations . . . 172
5.7.1 Covariant Levi–Civita tensor . . . 172
5.7.2 Matrix determinants reviewed . . . 176
5.8 Pseudotensors revisited: Hodge star . . . 179
5.8.1 Motivation and definition . . . 179
5.8.2 Linearity . . . 182
5.8.3 1-vectors and 0-vectors . . . 184
5.8.4 Self-inverse property . . . 186
5.9 k-vectors as tensors . . . 187
5.9.1 Bivectors as tensors . . . 187
5.9.2 3D 3-vectors as tensors . . . 188
5.9.3 Hodge star on tensors . . . 191
Summary of Chapter 5 . . . 193

The topic of generalized volumes (areas, volumes and/or hypervolumes) is often only briefly touched upon by courses of linear algebra and calculus. Often, only the formulas are being presented (if at all), but the question of how these

formulas emerge is largely left out of scope. Those formulas are based on the fact that the generalized volume computation is in turn based on the determinant formula, which gives the volume of a generalized parallelepiped (parallelotope). However this fact is often given for granted, as well as the determinant formula, which is simply being presented as given "by definition".

This, naturally, leaves a huge gap in the understanding of the respective linear algebra aspects. This chapter aims to fill this gap by giving an introduction into exterior algebra, which is probably the most (if not the only) intuitive way to address these topics.

> ▷ Another manifestation of the same knowledge gap is the seeming lack of a 2D version of the vector cross product, felt at least by some students of mathematics whose math courses didn't include exterior algebra. This not uncommonly leads to students trying to introduce a "2D cross product" as
>
> $$\vec{x} \times \vec{y} = \begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix} = |\vec{x}| \cdot |\vec{y}| \cdot \sin\theta$$

Motivational starter for Chapter 5

Develop a formal linear algebraic framework to support computation of areas, volumes and hypervolumes.

5.1 Exterior product in 2 dimensions

Commonly and naturally, the idea of 2D area starts with the area of a rectangle. In the context of linear algebra or calculus, we however quickly run into severe applicability limitations with that. In particular,

- under the majority of linear transformations a rectangle turns into a parallelogram
- essentially the same mechanism results in area differentials of parametrically defined surfaces (e.g. $\vec{x} = \vec{x}(u, v)$) having parallelogram shapes.

Therefore a rectangle is not good enough as a fundamental area element and we want to be able to deal with areas of parallelograms instead.

Thus, given two vectors $\vec{x}$ and $\vec{y}$ on a 2D plane we wish to compute the area of the parallelogram spanned by these two vectors (Fig. 5.1). This area is some function $S(\vec{x}, \vec{y})$ of these vectors. For the reasons which shall become clear later, from the very beginning we will replace the function notation by the operation notation: $S(\vec{x}, \vec{y}) = \vec{x} \wedge \vec{y}$.

> ▷ Our current usage of the $\vec{x} \wedge \vec{y}$ notation is not 100% formally correct. We currently assume that $\vec{x} \wedge \vec{y}$ returns a scalar value, while actually it is an element of a separate 1-dimensional linear space, formally not compatible with scalars. This approach shall do for now, we will correct the details later.

Similarly to what we did with the dot product in Section 3.3, we shall start by demonstrating the bilinearity of the operation $\vec{x} \wedge \vec{y}$ based on its additivity

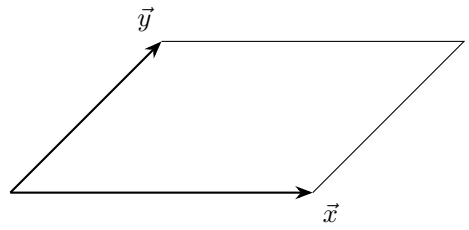


Figure 5.1: Parallelogram spanned by vectors $\vec{x}$ and $\vec{y}$.

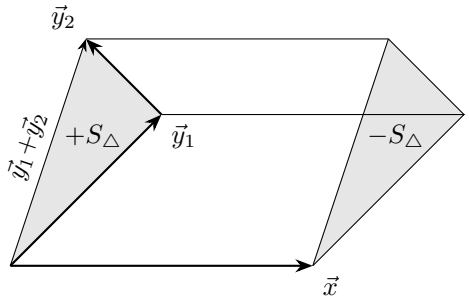


Figure 5.2: The area of the parallelogram spanned by vectors $\vec{x}$ and $\vec{y}_1 + y_2$ is the sum of the areas of parallelograms spanned by $\vec{x}$ and $\vec{y}_1$ and by $\vec{x}$ and $\vec{y}_2$. The shaded triangular areas $+S_\triangle$ and $-S_\triangle$ cancel each other.

(where we shall additionally employ continuity in order to accommodate real numbers). Fig. 5.2 demonstrates the additivity of the "$\wedge$" operation in the second argument. The idea is that there are two equal triangles, the sides of each triangle being the vectors $\vec{y}_1$, $\vec{y}_2$ and $\vec{y}_1 + \vec{y}_2$. One of these triangles is being added to the area and the other is being subtracted:

$$\vec{x} \wedge (\vec{y}_1 + \vec{y}_2) = \vec{x} \wedge \vec{y}_1 + \vec{x} \wedge \vec{y}_2 + S_\triangle - S_\triangle = \vec{x} \wedge \vec{y}_1 + \vec{x} \wedge \vec{y}_2$$

In a similar fashion, we can also demonstrate the additivity of the "$\wedge$" operation in its first argument.

> ▷ Notice that, differently from the dot product, this time we cannot use commutativity to extrapolate the additivity observations made for the second argument to the first argument, since we don't have a counterpart to the definition (3.29). As we also shall find out later, commutativity doesn't even hold in this case (although we could use anticommutativity, which does hold).

Continuity can be simply postulated: after all we expect that the area of a parallelogram spanned by vectors with rational coordinates is approximately equal to the area of a parallelogram spanned by vectors with real coordinates, which the rational coordinates of the first parallelogram approximate. This postulate would essentially state Cauchy sequence convergence for parallelogram areas. As an additional justification for that postulate, we can recall that it suffices to demonstrate continuity just at zero, and its not difficult to observe that the area of a parallelogram spanned by $\vec{x} \to 0$, $\vec{y} \to 0$ must approach zero

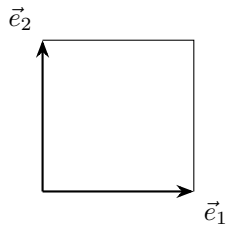


Figure 5.3: The area of the parallelogram spanned by the basis vectors $\vec{e}_1$ and $\vec{e}_2$ in an orthonormal basis is 1.

(e.g. by covering such parallelogram with a rectangle of indefinitely shrinking rational dimensions).

Now, by having additivity and continuity we also have bilinearity, and it remains to find out the values of the expressions $\vec{e}_i \wedge \vec{e}_j$. *For simplicity of the argument we shall restrict most of the discussion in the current chapter to orthonormal bases.* The expression $\vec{e}_1 \wedge \vec{e}_2$ denotes the area of a parallelogram spanned by basis vectors $\vec{e}_1$ and $\vec{e}_2$, however in an orthonormal basis this parallelogram turns into a unit square (Fig. 5.3). We will postulate the area of this unit square to be equal to 1:

$$\vec{e}_1 \wedge \vec{e}_2 = 1 \tag{5.1}$$

which essentially specifies the units to be used for measuring the area.

▷ Rather than simply postulating $\vec{e}_1 \wedge \vec{e}_2 = 1$, we could explicitly compute the area of the unit square as

$$\int \mathrm{d}S = \iint \mathrm{d}x\,\mathrm{d}y = \int_0^1 \left(\int_0^1 \mathrm{d}x \right) \mathrm{d}y = 1$$

However in doing so we would implicitly postulate pretty much the same properties:

- The assumption that the area of an infinitely small rectangle of a $\mathrm{d}x \times \mathrm{d}y$ size is equal to $\mathrm{d}x\,\mathrm{d}y$ is related to $\vec{e}_1 \wedge \vec{e}_2 = 1$ via homogeneity:

$$(\mathrm{d}x \cdot \vec{e}_1) \wedge (\mathrm{d}y \cdot \vec{e}_2) = \mathrm{d}x\,\mathrm{d}y \cdot \vec{e}_1 \wedge \vec{e}_2 = \mathrm{d}x\,\mathrm{d}y \cdot 1 = \mathrm{d}x\,\mathrm{d}y$$

- The assumption that the total area can be computed by integrating $\mathrm{d}x\,\mathrm{d}y$ de-facto postulates additivity (since the integral is essentially a sum) and continuity (since the integral is a limit of a sum).

Considering that the expression $\vec{e}_2 \wedge \vec{e}_1$ also describes a unit square spanned by the same vectors $\vec{e}_1$ and $\vec{e}_2$, just taken in a different order, we might conclude that also $\vec{e}_2 \wedge \vec{e}_1 \stackrel{?}{=} 1$, however then we are going to quickly run into contradictions. Probably the most prominent and fundamental contradiction that we would run into is violation of the rotational symmetry. In application to the "$\wedge$" operation, rotational symmetry should mean that rotating a pair of vectors $\vec{x}$, $\vec{y}$ does not change the value of $\vec{x} \wedge \vec{y}$. Because if this value did change, it would mean the existence of "preferred directions" in space. With that in mind we can notice that by applying bilinearity to (5.1) we obtain $\vec{e}_1 \wedge (-\vec{e}_2) = -(\vec{e}_1 \wedge \vec{e}_2) = -1$. On the other hand, the parallelogram spanned by $\vec{e}_1$, $-\vec{e}_2$ is related to the parallelogram spanned by $\vec{e}_2$, $\vec{e}_1$ by a 90° rotation

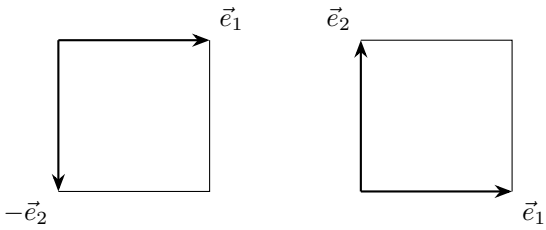


Figure 5.4: Parallelograms spanned by $\vec{e}_1$, $-\vec{e}_2$ and on $\vec{e}_2$, $\vec{e}_1$ are related by a 90° rotation.

(Fig. 5.4) and by rotational symmetry their areas must be equal and therefore

$$\vec{e}_2 \wedge \vec{e}_1 = -1 \tag{5.2}$$

But isn't the parallelogram spanned by $\vec{e}_2$, $\vec{e}_1$ the same as the one spanned by $\vec{e}_1$, $\vec{e}_2$? After all, the order in which we specify vectors shouldn't matter, right? It turns out, our previous discussion implies exactly the opposite: the order should matter, if we do not want to run into contradictions. We thus have

$$\vec{e}_2 \wedge \vec{e}_1 = -(\vec{e}_1 \wedge \vec{e}_2) \tag{5.3}$$

This means that the operation "$\wedge$" must yield signed rather than simply non-negative area.

▷ The idea of signed (or *oriented*) area doesn't really deprecate the idea of unsigned (or *unoriented*) area. Rather, both ideas coexist. Depending on the context we might be talking of signed or unsigned area, although in this chapter we will almost exclusively use signed areas.

The relationship between a signed and an unsigned area is quite similar to the relationship between a vector and its length. The details of this should become more clear throughout the course of this chapter.

We still need to obtain the values of $\vec{e}_1 \wedge \vec{e}_1$ and $\vec{e}_2 \wedge \vec{e}_2$, where we can immediately notice that the respective parallelograms are singular and have zero areas:

$$\vec{e}_1 \wedge \vec{e}_1 = \vec{e}_2 \wedge \vec{e}_2 = 0 \tag{5.4}$$

▷ It's easiest to infer (5.4) using the concept of unsigned area. The singular parallelograms in (5.4) can be covered by rectangles of arbitrarily small area, therefore their (unsigned) areas must be less than or equal to an arbitrarily small positive number, and hence must be simply zero. Since unsigned area is equal to the absolute value of the respective signed area, the signed areas of the singular parallelograms are also zero.

The formula for the unsigned area of a rectangle (as a product of lengths of the rectangle's sides), in turn, can be inferred using the same approach that we're taking in this chapter to infer the formula for the signed area of a parallelogram. The readers might take this as an additional exercise.

Having established the bilinearity of "$\wedge$" as well as the equalities

(exterior products of 2D basis vectors in a positively-oriented orthonormal basis)

$$\vec{e}_1 \wedge \vec{e}_2 = 1 \qquad \vec{e}_2 \wedge \vec{e}_1 = -1 \qquad \vec{e}_1 \wedge \vec{e}_1 = \vec{e}_2 \wedge \vec{e}_2 = 0$$

we can obtain the expression for $\vec{x} \wedge \vec{y}$ in terms of vector components.

Goal 5.1. Express $\vec{x} \wedge \vec{y}$ (in 2D) in terms of components of $\vec{x}$ and $\vec{y}$. *Hint:* follow the approach taken in the derivations of (3.31) and (4.34).

Using the bilinearity of "$\wedge$", we obtain

$$\begin{aligned}
\vec{x} \wedge \vec{y} &= (x_1\vec{e}_1 + x_2\vec{e}_2) \wedge (y_1\vec{e}_1 + y_2\vec{e}_2) \\
&= \left(\sum_{i=1}^{2} x_i\vec{e}_i\right) \wedge \left(\sum_{j=1}^{2} y_j\vec{e}_j\right) = \sum_{i=1}^{2}\sum_{j=1}^{2}(x_i\vec{e}_i \wedge y_j\vec{e}_j) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2}\Big(x_i y_j \cdot (\vec{e}_i \wedge \vec{e}_j)\Big) \\
&= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{bmatrix} \vec{e}_1 \wedge \vec{e}_1 & \vec{e}_1 \wedge \vec{e}_2 \\ \vec{e}_2 \wedge \vec{e}_1 & \vec{e}_2 \wedge \vec{e}_2 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \\
&= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \\
&= \mathbf{x}^{\mathrm{T}} \cdot \mathbf{\Gamma} \cdot \mathbf{y}
\end{aligned} \tag{5.5}$$

or

(2D exterior product in an orthonormal basis in tensor index notation)

$$\vec{x} \wedge \vec{y} = x^i\vec{e}_i \wedge y^j\vec{e}_j = (\vec{e}_i \wedge \vec{e}_j) \cdot x^i y^i = \varepsilon_{ij} x^i y^j \tag{5.6}$$

That is, we obtain the result similar to (4.34), just with ε_{ij} in place of the metric tensor g_{ij}. Also the matrix $\mathbf{\Gamma}$ in (5.5) is not a matrix of dot products anymore and hence cannot be referred to as Gram matrix. For the lack of a better name, we will refer to it here as a Gram-like matrix. Respectively, in tensor notation we don't get the metric tensor g_{ij}, but some other tensor-like entity ε_{ij}, defined by $\varepsilon_{11} = \varepsilon_{22} = 0$, $\varepsilon_{12} = 1$, $\varepsilon_{21} = -1$. This entity is not a tensor, but, not unlike δ_{ij}, a tensor-like symbol, referred to as the (2-dimensional) *Levi–Civita symbol*. As we shall see in Section 5.7, the Levi–Civita symbol doesn't transform as a $(0, 2)$-type tensor (but a similar entity does).

Expanding the matrix product in (5.5) we obtain

(2D exterior product in component terms in a positively-oriented orthonormal basis)

$$\vec{x} \wedge \vec{y} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = x_1 y_2 - x_2 y_1 = \det \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix} \tag{5.7}$$

The operation $\vec{x} \wedge \vec{y}$, which we just introduced is referred to as the *exterior product* (alternatively called *wedge product*[1]) of vectors $\vec{x}$ and $\vec{y}$, and the formula (5.7) gives an expression for the computation of the exterior product in an orthonormal basis in 2 dimensions.

▷ The readers are reminded about the earlier made reservation that formally the exterior product is not a scalar value but an element of a special linear space. From that perspective, the formula (5.7) just gives an expression for that space element's component.

▷ Notice that we have obtained the formula (5.7), which gives an expression for the area of a parallelogram, without even giving a "real definition" of what the area is. Instead we just postulated a number of properties of the area inferred from geometric observations and common sense.

Goal 5.1 completed. ■

▷ The 3D generalization of the area of a parallelogram is obviously the volume of a parallelepiped, which we shall consider a bit later. What might be less obvious, is that the 1D version of that is the length of a closed interval. As an exercise, the readers might try to construct a formal approach to the length of a closed interval (in a 1D space) along the same lines how we proposed to approach the parallelogram area in a 2D space in this text.

5.2 Symmetries

Anticommutativity

The property (5.3) generalizes to arbitrary vectors

(anticommutativity of 2D exterior product)

$$\vec{y} \wedge \vec{x} = -(\vec{x} \wedge \vec{y}) \tag{5.8}$$

which means that exterior product is anticommutative.

Goal 5.2. Demonstrate the anticommutativity (5.8) of the exterior product. *Hint:* use the already demonstrated anticommutativity of the exterior product of basis vectors.

Noticing that

$$\vec{e}_j \wedge \vec{e}_i = -(\vec{e}_i \wedge \vec{e}_j) \quad \forall i, j \tag{5.9}$$

(where for $i \neq j$ (5.9) directly states (5.3) while for $i = j$ (5.9) follows from (5.4)), we have

$$\begin{aligned}
\vec{y} \wedge \vec{x} &= (y_1\vec{e}_1 + y_2\vec{e}_2) \wedge (x_1\vec{e}_1 + x_2\vec{e}_2) \\
&= y_1x_1 \cdot \vec{e}_1 \wedge \vec{e}_1 + y_2x_1 \cdot \vec{e}_2 \wedge \vec{e}_1 + y_1x_2 \cdot \vec{e}_1 \wedge \vec{e}_2 + y_2x_2 \cdot \vec{e}_2 \wedge \vec{e}_2 \\
&= -(y_1x_1 \cdot \vec{e}_1 \wedge \vec{e}_1 + y_2x_1 \cdot \vec{e}_1 \wedge \vec{e}_2 + y_1x_2 \cdot \vec{e}_2 \wedge \vec{e}_1 + y_2x_2 \cdot \vec{e}_2 \wedge \vec{e}_2)
\end{aligned}$$

[1] In this text we are going to use the term "exterior product" to stress the area-based interpretation and aspects, while the term "wedge product" will be preferred when focusing on "∧" as a more abstract operation.

$$= -\Big((x_1\vec{e}_1 + x_2\vec{e}_2) \wedge (y_1\vec{e}_1 + y_2\vec{e}_2)\Big) = -(\vec{x} \wedge \vec{y})$$

or, in tensor notation

$$\vec{y} \wedge \vec{x} = (y^j\vec{e}_j) \wedge (x^i\vec{e}_i) = y^j x^i \cdot (\vec{e}_j \wedge \vec{e}_i) = -\Big(x^i y^j \cdot (\vec{e}_i \wedge \vec{e}_j)\Big) = -(\vec{x} \wedge \vec{y})$$

Effectively, the anticommutativity of basis vectors (5.9) is "promoted" through the bilinearity property to the anticommutativity of arbitrary vectors.

Goal 5.2 completed. ■

It follows that

(2D exterior product of a vector with itself is zero)

$$\vec{x} \wedge \vec{x} = 0 \quad \forall \vec{x} \tag{5.10}$$

which is a generalization of (5.4) to arbitrary vectors.

Goal 5.3. Demonstrate (5.10).

Letting $\vec{y} = \vec{x}$ in (5.8) we obtain $\vec{x} \wedge \vec{x} = -(\vec{x} \wedge \vec{x})$, and (5.10) follows.

▷ We can also obtain (5.10) directly from (5.4) in an intuitive way. Indeed, by rotational invariance (5.4) generalizes to arbitrary unit vectors, and then by homogeneity to vectors of arbitrary length and direction.

Goal 5.3 completed. ■

Notably, the property (5.10) automatically implies (5.8) via bilinearity:

$$\begin{aligned} 0 &\overset{(5.10)}{=} (\vec{x} + \vec{y}) \wedge (\vec{x} + \vec{y}) = \vec{x} \wedge \vec{x} + \vec{y} \wedge \vec{x} + \vec{x} \wedge \vec{y} + \vec{y} \wedge \vec{y} \\ &\overset{(5.10)}{=} 0 + \vec{y} \wedge \vec{x} + \vec{x} \wedge \vec{y} + 0 = \vec{y} \wedge \vec{x} + \vec{x} \wedge \vec{y} \end{aligned} \tag{5.11}$$

Shear transformations

The property (5.10) is also closely related to the "shearing property" of the exterior product. A shear transformation of a parallelogram adds a (fractional) multiple of one vector to the other, e.g. instead of $\vec{x} \wedge \vec{y}$ we consider $\vec{x} \wedge (\vec{y} + \lambda\vec{x})$. Fig. 5.5 illustrates, where we can see that this effectively adds and subtracts one and the same triangle (spanned by vectors $\vec{y}$, $\lambda\vec{x}$ and $\vec{y} + \lambda\vec{x}$) to/from the area of the parallelogram, thereby leaving the area unchanged:

$$\vec{x} \wedge (\vec{y} + \lambda\vec{x}) = \vec{x} \wedge \vec{y} + S_\triangle - S_\triangle = \vec{x} \wedge \vec{y} \tag{5.12a}$$

The same result could have been obtained from bilinearity and (5.10):

$$\vec{x} \wedge (\vec{y} + \lambda\vec{x}) = \vec{x} \wedge \vec{y} + \lambda \cdot \vec{x} \wedge \vec{x} \overset{(5.10)}{=} \vec{x} \wedge \vec{y}. \tag{5.12b}$$

Conversely, we could infer (5.10) from (5.12a) and bilinearity:

$$\vec{x} \wedge \vec{y} \overset{(5.12a)}{=} \vec{x} \wedge (\vec{y} + \lambda\vec{x}) = \vec{x} \wedge \vec{y} + \lambda \cdot \vec{x} \wedge \vec{x}$$

and it follows that $\vec{x} \wedge \vec{x} = 0$.

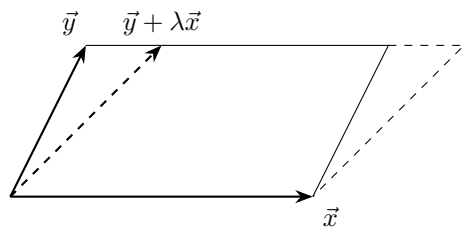


Figure 5.5: Shear transformation of a parallelogram doesn't change its area.

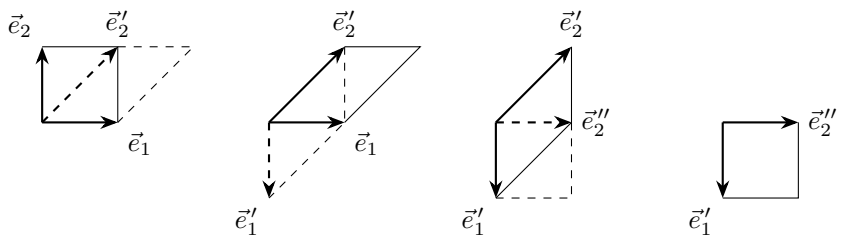


Figure 5.6: Series of shear transformations of $\vec{e}_1$, $\vec{e}_2$ resulting in a rotation by 90°. The first shear transformation rotates $\vec{e}_2$ by 45°, the second one rotates $\vec{e}_1$ by 90°, and the third one rotates $\vec{e}_2'$ by another 45°.

▷ It follows that areas in general do not change under shear transformations. Indeed, if we compute an area of a geometric figure as a double integral of $\mathrm{d}x\,\mathrm{d}y$, the areas of the $\mathrm{d}x \times \mathrm{d}y$ rectangles will be invariant under shear transformation and so will be their sum.

Notably, the "shearing property" (5.12) implies that the product $\vec{e}_1 \wedge \vec{e}_2$ is unchanged by 90° clockwise rotation (which was the basis for inferring (5.3)). Specifically, we can use a series of 3 shear transformations to convert $\vec{e}_1 \wedge \vec{e}_2$ into $(-\vec{e}_2) \wedge \vec{e}_1$, as shown in Fig. 5.6. Therefore (5.12) implies (5.3), pretty much like (5.10) (to which (5.12) is closely related) does.

Rotations

In the derivation of the exterior product we have been postulating rotation invariance. However in doing so we have only been applying rotation invariance in certain particular cases, which still leaves an opportunity for the rotation invariance not holding in general. Let's therefore verify whether the exterior product is rotationally invariant.

It would suffice to show that $\vec{e}_i \wedge \vec{e}_j$ is rotationally invariant $\forall i, j$, the same property for arbitrary vectors will follow by bilinearity. In other words, we need to show that the Gram-like matrix $\mathbf{\Gamma}$ defined in (5.5) doesn't change upon rotations of basis vectors. Notice that here (differently from our main approach in Chapter 4) we are not looking for ways to somehow compensate for the change of basis. On the contrary, $\mathbf{\Gamma}$ is explicitly expressed via the exterior products of basis vectors, and we want to know what happens to these products if basis vectors are changed.

Goal 5.4. Show that the components of $\mathbf{\Gamma}$ in (5.5) are unchanged by basis rotations.

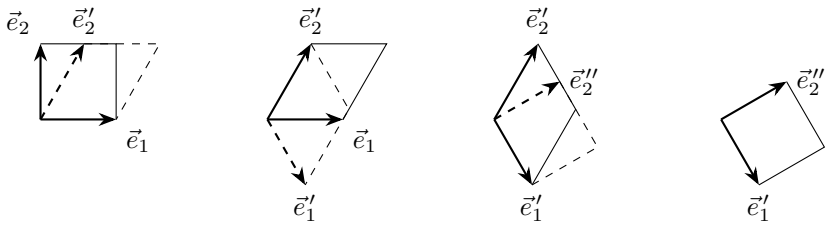


Figure 5.7: Series of shear transformations of $\vec{e}_1$, $\vec{e}_2$ resulting in a rotation by an arbitrary given angle α. The first shear transformation rotates $\vec{e}_2$ by $\alpha/2$, the second one rotates $\vec{e}_1$ by α, and the third one rotates $\vec{e}_2'$ by another $\alpha/2$.

Rotating $\vec{e}_1$ and $\vec{e}_2$ by α we get

$$\begin{aligned} \vec{e}_1' &= \vec{e}_1 \cos\alpha + \vec{e}_2 \sin\alpha \\ \vec{e}_2' &= -\vec{e}_1 \sin\alpha + \vec{e}_2 \cos\alpha \end{aligned} \tag{5.13}$$

Respectively

$$\begin{aligned} \vec{e}_1' \wedge \vec{e}_2' &= -\vec{e}_1 \wedge \vec{e}_1 \cdot \cos\alpha \sin\alpha + \vec{e}_1 \wedge \vec{e}_2 \cdot \cos^2\alpha \\ &\quad - \vec{e}_2 \wedge \vec{e}_1 \cdot \sin^2\alpha + \vec{e}_2 \wedge \vec{e}_2 \cdot \sin\alpha\cos\alpha \\ &\overset{(5.10)}{=} 0 + \vec{e}_1 \wedge \vec{e}_2 \cdot (\cos^2\alpha + \sin^2\alpha) + 0 = \vec{e}_1 \wedge \vec{e}_2 \end{aligned} \tag{5.14a}$$

and by (5.8)

$$\vec{e}_2' \wedge \vec{e}_1' = -(\vec{e}_1' \wedge \vec{e}_2') = -(\vec{e}_1 \wedge \vec{e}_2) = \vec{e}_2 \wedge \vec{e}_1 \tag{5.14b}$$

while the equalities $\vec{e}_i' \wedge \vec{e}_i' = \vec{e}_i \wedge \vec{e}_i$ follow directly from (5.10). Thereby the components $\vec{e}_i \wedge \vec{e}_j$ of the matrix $\mathbf{\Gamma}$ are rotationally invariant $\forall i, j$.

Notice that we again used (5.10), this time to derive rotational invariance. We can also do the same in a graphical way, using the shearing property (5.12a) as a "graphical expression" of (5.10). Specifically, we can generalize the idea of Fig. 5.6 to arbitrary angles, as shown in Fig. 5.7.

Goal 5.4 completed. ■

Since the products $\vec{e}_i \wedge \vec{e}_j$ are rotationally invariant, it follows by bilinearity that so are the exterior products of arbitrary vectors $\vec{x} \wedge \vec{y}$.

Mirror reflections

As we are sticking to orthonormal bases for now, the "allowed" basis transformations are orthogonal transformations. We can recall from our discussion in Subsection 3.4.4 that, besides rotations, there is one more type of orthogonal transformations of linear space: mirror reflections.

A mirror reflection in arbitrary direction can be represented as a combination of two rotations and a mirror reflection in some fixed direction, e.g. in the direction of $\vec{e}_1$ (where the first rotation would match the desired rotation direction with $\vec{e}_1$ and the second rotation would undo the first rotation). We already know that exterior product is invariant under rotations, so it suffices

to study how the exterior product behaves under space reflections along $\vec{e}_1$. As with rotations, we can simply look at how the matrix $\mathbf{\Gamma}$ in (5.5) changes upon flipping the sign of $\vec{e}_1$.

Since the elements of $\mathbf{\Gamma}$ are equal to $\vec{e}_i \wedge \vec{e}_j$, replacing $\vec{e}_1$ by $-\vec{e}_1$ inverts the auxiliary diagonal elements of $\mathbf{\Gamma}$, while the main diagonal elements stay equal to zero. So if originally we had $\mathbf{\Gamma} = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$, upon reflection we obtain $\mathbf{\Gamma} = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$. Thereby $\mathbf{\Gamma}$ is effectively replaced by $-\mathbf{\Gamma}$, and therefore the exterior product $\vec{x} \wedge \vec{y} = \mathbf{x}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{y}$ changes sign under mirror reflections of the linear space.

> ▷ We mentioned earlier that exterior product is an element of a special 1-dimensional linear space, formally not compatible with scalars. By now we have seen that, if taken as a numeric value, exterior product behaves as a pseudoscalar, and pseudoscalars are not formally compatible with scalars (one generally would not equate or add a scalar and a pseudoscalar).

Basis changes

So, as long as we stick to orthonormal bases, any basis transformation is either a rotation or a mirror reflection or a combination of those. We also may recall from Section 4.15 that basis transformations are mathematically equivalent to linear space transformations (at least as long as only vectors are involved). In particular, basis rotations are equivalent to rotations of the linear space by the same angle but in the opposite direction, and mirror reflections of the basis are equivalent to mirror reflections of the entire space along the same direction.

We have just seen that rotations of the linear space do not change exterior products. Therefore, replacing a basis by a rotated one (and adjusting the vector coordinates contravariantly) does not change the value of the expression (5.5). Therefore the formula (5.5) is valid in all bases obtained from the original basis by rotations.

We also have seen that reflections of the linear space change the sign of the exterior product. To compensate for that, after a mirror reflection of the basis, the exterior product formula also has to be replaced by one with an opposite sign:

$$\vec{x} \wedge \vec{y} = -(x_1 y_2 - x_2 y_1) = x_2 y_1 - x_1 y_2 \qquad \text{(negatively oriented basis)} \tag{5.15}$$

Looking from another perspective, we can notice that the Gram-like matrix $\mathbf{\Gamma}$ in (5.5) is the same in all bases related to each other by rotation, since, as we have found out, the elements $\vec{e}_i \wedge \vec{e}_j$ of the matrix $\mathbf{\Gamma}$ are unchanged by rotations. However, performing a reflection of the basis turns the matrix $\mathbf{\Gamma}$ into $-\mathbf{\Gamma}$. This splits all orthonormal bases in 2D into two groups. In one group we have $\vec{e}_1 \wedge \vec{e}_2 = 1$ and respectively $\mathbf{\Gamma} = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$. In the other group we have $\vec{e}_1 \wedge \vec{e}_2 = -1$ and respectively $\mathbf{\Gamma} = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$. Thus, there are two different formulas for the exterior product: the (5.7) and the (5.15). How do we know which one is applicable to the basis we're currently in? This question can be rephrased: how do we know whether $\vec{e}_1 \wedge \vec{e}_2 = 1$ or $\vec{e}_1 \wedge \vec{e}_2 = -1$? The answer actually is: there is no way to know! Remember, that we simply postulated that $\vec{e}_1 \wedge \vec{e}_2 = 1$. Instead, we could have postulated $\vec{e}_1 \wedge \vec{e}_2 = -1$. Exterior product is a "signed" area quantity, and which unit areas have positive sign and which do have negative sign is merely a matter of convention. It is common to choose the positive sign in case the 90° rotation from $\vec{e}_1$ to $\vec{e}_2$ is going counterclockwise

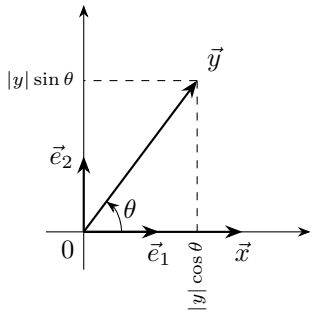


Figure 5.8: Derivation of the geometric interpretation of the exterior product.

(as in Fig. 5.3) and choose the negative sign otherwise. The "counterclockwise" basis is referred to as *positively oriented* and the "clockwise" basis is referred to as *negatively oriented*.[2]

5.3 Exterior vs. dot product

Now that we have demonstrated the invariance of the exterior product under rotations, it's time to obtain a geometric interpretation-based formula for it, similar to the formula (3.29) for the dot product.

Goal 5.5. Construct a geometric interpretation of the exterior product $\vec{x} \wedge \vec{y}$ in 2D, by obtaining an exterior product formula in a form similar to (3.29). *Hint:* since exterior product is rotation-invariant, we can try to choose a convenient basis for our computations.

Given an exterior product $\vec{x} \wedge \vec{y}$, let's choose the basis vector $\vec{e}_1$ going in the same direction as $\vec{x}$ (so that $\vec{x} = x_1 \vec{e}_1$, where $x_1 \geq 0$) and let's choose $\vec{e}_2$ so that $\vec{e}_1 \wedge \vec{e}_2 = 1$ (Fig. 5.8). Assuming the orthonormality of the basis, this uniquely specifies the basis choice. In this basis the components of $\vec{x}$ are

$$\begin{aligned} x_1 &= |\vec{x}| \\ x_2 &= 0 \end{aligned} \tag{5.16a}$$

and the components of $\vec{y}$ are

$$\begin{aligned} y_1 &= |\vec{y}| \cdot \cos\theta \\ y_2 &= |\vec{y}| \cdot \sin\theta \end{aligned} \tag{5.16b}$$

where θ is the angle from $\vec{x}$ to $\vec{y}$, where counterclockwise rotations are considered positive.

Substituting (5.16a) and (5.16b) into the positively oriented-basis formula (5.7) we obtain

$$\vec{x} \wedge \vec{y} = |\vec{x}| \cdot |\vec{y}| \cdot \sin\theta \tag{5.17}$$

[2]The author is unsure whether there is only one convention universally employed across the world, especially if we go into higher dimension counts.

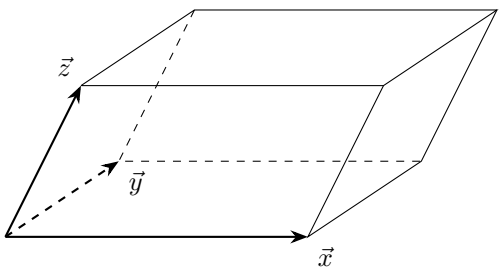


Figure 5.9: Parallelepiped spanned by vectors $\vec{x}$, $\vec{y}$ and $\vec{z}$.

▷ We can verify whether the value given by (5.17) coincides with the area computed by integration:

$$S = \int \mathrm{d}S = \iint \mathrm{d}x\,\mathrm{d}y = \int_0^{|\vec{y}|\sin\theta} \left(\int_{\frac{y}{\tan\theta}}^{|\vec{x}|+\frac{y}{\tan\theta}} \mathrm{d}x \right) \mathrm{d}y$$
$$= \int_0^{|\vec{y}|\sin\theta} |\vec{x}| \cdot \mathrm{d}y = |\vec{x}| \cdot |\vec{y}| \cdot \sin\theta$$

which is the same expression as (5.17).

Goal 5.5 completed. ■

Comparing (5.17) to (3.29) we can notice a certain duality between the two formulas, it's just one is containing $\cos\theta$ and the other one containing $\sin\theta$. Thereby $\vec{x}\cdot\vec{y}$ and $\vec{x}\wedge\vec{y}$ together uniquely specify the angle between the vectors (up to a $2\pi k$ term). It also follows that

(relationship between exterior and dot products in 2D)

$$(\vec{x}\cdot\vec{y})^2 + (\vec{x}\wedge\vec{y})^2 = |\vec{x}|^2 \cdot |\vec{y}|^2 \tag{5.18}$$

▷ Formula (5.18) is a particular case of the more general formula (5.51).

5.4 Volume

Let's try to generalize the previous ideas from 2 to 3 dimensions. Instead of the area function $S(\vec{x},\vec{y})$ introduced in the beginning of Section 5.1 (which we then immediately replaced with the $\vec{x}\wedge\vec{y}$ notation) we shall introduce the volume function $V(\vec{x},\vec{y},\vec{z})$, denoting the volume of a parallelepiped spanned by the 3 vectors $\vec{x}$, $\vec{y}$, $\vec{z}$ (Fig. 5.9).

Similarly to how $S(\vec{x},\vec{y})$ was found to be bilinear, the function $V(\vec{x},\vec{y},\vec{z})$ is trilinear (linear in each of its 3 arguments). E.g.

$$V(\vec{x}_1+\vec{x}_2,\vec{y},\vec{z}) = V(\vec{x}_1,\vec{y},\vec{z}) + V(\vec{x}_2,\vec{y},\vec{z})$$

etc. The argumentation for the trilinearity is pretty similar to how the bilinearity of $S(\vec{x},\vec{y})$ (that is, of $\vec{x}\wedge\vec{y}$) was argued, and we shall not repeat it here,

the readers may work it out by themselves, if they wish so (and this could be a good exercise).

With $S(\vec{x}, \vec{y})$ we have been defining the units of reference by postulating $S(\vec{e}_1, \vec{e}_2) = 1$. We can do a similar thing here and postulate

(volume of the unit cube)

$$V(\vec{e}_1, \vec{e}_2, \vec{e}_3) = 1 \tag{5.19}$$

▷ Or, we could compute the volume of a unit cube by integration

$$\int \mathrm{d}V = \iiint \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \int_0^1 \left(\int_0^1 \left(\int_0^1 \mathrm{d}x \right) \mathrm{d}y \right) \mathrm{d}z = 1$$

with the usual reservations about the definitions implicitly applied by using this formula.

Now we shall generalize some of the properties previously established for $S(\vec{x}, \vec{y})$ (in the form of the properties of $\vec{x} \wedge \vec{y}$) to the respective properties of $V(\vec{x}, \vec{y}, \vec{z})$.

Goal 5.6. Generalize the property $S(\vec{x}, \vec{x}) = 0$ to the case of $V(\vec{x}, \vec{y}, \vec{z})$.

As the property $S(\vec{e}_i, \vec{e}_i) = \vec{e}_i \wedge \vec{e}_i = 0$ goes, and, more generally, the property $S(\vec{x}, \vec{x}) = \vec{x} \wedge \vec{x} = 0$, they generalize to the 3D as follows: if at least 2 of the 3 vectors coincide, the parallelepiped becomes "flat" and its volume turns to zero:

(volume of a parallelepiped with at least two coinciding edges is zero)

$$V(\vec{x}, \vec{x}, \vec{y}) = V(\vec{x}, \vec{y}, \vec{x}) = V(\vec{x}, \vec{y}, \vec{y}) = 0 \quad \forall \vec{x}, \vec{y} \tag{5.20}$$

It also follows (by letting $\vec{y} = \vec{x}$) that $V(\vec{x}, \vec{x}, \vec{x}) = 0\ \forall \vec{x}$. Thus, if any argument of $V(\cdot, \cdot, \cdot)$ is repeated, the volume is zero.

Goal 5.6 completed. ■

Goal 5.7. Generalize the property $S(\vec{y}, \vec{x}) = -S(\vec{x}, \vec{y})$ to the case of $V(\vec{x}, \vec{y}, \vec{z})$.

Recall that the expression $S(\vec{e}_1, \vec{e}_2)$ was anticommutative:

$$S(\vec{e}_2, \vec{e}_1) = \vec{e}_2 \wedge \vec{e}_1 = -(\vec{e}_1 \wedge \vec{e}_2) = -S(\vec{e}_1, \vec{e}_2)$$

The expression $V(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ has a similar property, but now that we have 3 rather than 2 arguments, this property is generalized to anticommutativity in each pair of the arguments. E.g.

$$V(\vec{e}_2, \vec{e}_1, \vec{e}_3) = -V(\vec{e}_1, \vec{e}_2, \vec{e}_3)$$

The argumentation is similar to $S(\vec{e}_1, \vec{e}_2)$ but more complicated.

One option is to deduce the anticommutativity from (5.20) by trilinearity. E.g.

$$0 = V(\vec{x} + \vec{y}, \vec{x} + \vec{y}, \vec{z}) = V(\vec{x}, \vec{x}, \vec{z}) + V(\vec{x}, \vec{y}, \vec{z}) + V(\vec{y}, \vec{x}, \vec{z}) + V(\vec{y}, \vec{y}, \vec{z})$$

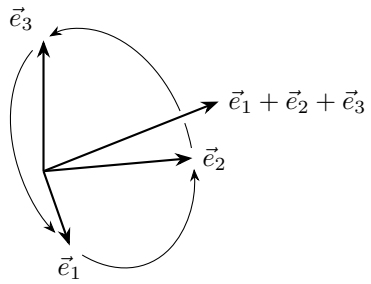


Figure 5.10: Positively oriented 3D basis (the vector $\vec{e}_1$ is pointing "out of the page").

$$= 0 + V(\vec{x}, \vec{y}, \vec{z}) + V(\vec{y}, \vec{x}, \vec{z}) + 0 = V(\vec{x}, \vec{y}, \vec{z}) + V(\vec{y}, \vec{x}, \vec{z})$$

The other one is by space rotation. Let's first imagine the 3 orthonormal basis vectors $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$ being looked at "from the tip" of their sum $\vec{e}_1+\vec{e}_2+\vec{e}_3$ (that is we draw all 3 vectors $\vec{e}_i$ from the origin and also draw the vector $\vec{e}_1+\vec{e}_2+\vec{e}_3$ from the origin, and we look at the vectors $\vec{e}_i$ from the tip of the vector $\vec{e}_1+\vec{e}_2+\vec{e}_3$). The tips of the vectors $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$, when looked at from this point, will be going either clock- or counterclockwise (Fig. 5.10)

Let's, for the sake of definiteness, assume that the vectors are going counterclockwise (this is referred to as a "positively oriented 3D basis"). Now let's say that we change the order of the arguments in the expression $V(\vec{e}_1, \vec{e}_2, \vec{e}_3)$. We want to treat the vector $\vec{e}_2$ as if it was the first vector in the triplet, $\vec{e}_3$ as the second, and $\vec{e}_1$ as the third one, so that our volume expression becomes $V(\vec{e}_2, \vec{e}_3, \vec{e}_1)$. Apparently such transformation can be achieved just by rotating the vectors around the $\vec{e}_1+\vec{e}_2+\vec{e}_3$ axis by 120° and thereby doesn't change the volume. The same holds for $V(\vec{e}_3, \vec{e}_1, \vec{e}_2)$, which can be achieved by a further 120° rotation. Thus

(volumes of positively-oriented basis triplets)

$$V(\vec{e}_1, \vec{e}_2, \vec{e}_3) = V(\vec{e}_2, \vec{e}_3, \vec{e}_1) = V(\vec{e}_3, \vec{e}_1, \vec{e}_2) = 1 \tag{5.21}$$

Now, say, instead we replace $\vec{e}_1$ with $-\vec{e}_1$, obtaining the triplet $-\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$. By trilinearity we have

$$V(-\vec{e}_1, \vec{e}_2, \vec{e}_3) = -V(\vec{e}_1, \vec{e}_2, \vec{e}_3)$$

However the triplet $-\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$ can be converted into the triplet $\vec{e}_2$, $\vec{e}_1$, $\vec{e}_3$ (and back) by a 90° rotation around the axis $\vec{e}_3$ (pretty much how in 2D we could convert $-\vec{e}_1$, $\vec{e}_2$ into $\vec{e}_2$, $\vec{e}_1$ by a 90° rotation). Thereby their volumes must be equal and

$$V(\vec{e}_2, \vec{e}_1, \vec{e}_3) = V(-\vec{e}_1, \vec{e}_2, \vec{e}_3) = -V(\vec{e}_1, \vec{e}_2, \vec{e}_3) = -1$$

By rotating the argument order in the triplet $\vec{e}_2$, $\vec{e}_1$, $\vec{e}_3$ (which is the same as 120° rotations of the space) we obtain

(volumes of negatively-oriented basis triplets)

$$V(\vec{e}_2, \vec{e}_1, \vec{e}_3) = V(\vec{e}_1, \vec{e}_3, \vec{e}_2) = V(\vec{e}_3, \vec{e}_2, \vec{e}_1) = -1 \tag{5.22}$$

Comparing (5.21) and (5.22) we notice that swapping any two arguments changes the expression sign, thereby these expressions are anticommutative in each pair of the arguments. It doesn't take much to derive the anticommutativity of the function $V(\vec{x}, \vec{y}, \vec{z})$ for arbitrary vectors as arguments. E.g.

$$\begin{aligned}
V(\vec{x}, \vec{y}, \vec{z}) &= V\left(\left(\sum_i x_i \vec{e}_i\right), \left(\sum_j y_j \vec{e}_j\right), \vec{z}\right) = \sum_{i,j} x_i y_i V(\vec{e}_i, \vec{e}_j, \vec{z}) \\
V(\vec{y}, \vec{x}, \vec{z}) &= V\left(\left(\sum_j y_j \vec{e}_j\right), \left(\sum_i x_i \vec{e}_i\right), \vec{z}\right) = \sum_{i,j} x_i y_j V(\vec{e}_j, \vec{e}_i, \vec{z}) \\
&= -\sum_{i,j} x_i y_j V(\vec{e}_i, \vec{e}_j, \vec{z}) = -V(\vec{x}, \vec{y}, \vec{z})
\end{aligned}$$

and therefore

(antisymmetry of volume in a pair of arguments)

$$V(\vec{x}, \vec{y}, \vec{z}) = -V(\vec{y}, \vec{x}, \vec{z})$$

The same holds for any other pair of arguments.

Goal 5.7 completed. ■

At this point we shall (finally) introduce the "$\wedge$" notation for the volume $V(\vec{x}, \vec{y}, \vec{z})$. Since now we have 3 arguments instead of 2, we shall write a "triple exterior product" (or "triple wedge product")

(3D volume as a triple wedge product)

$$\vec{x} \wedge \vec{y} \wedge \vec{z} = V(\vec{x}, \vec{y}, \vec{z}) \tag{5.23}$$

This seemingly somewhat awkward notation has a good justification: actually, pairwise wedge products, like e.g. $\vec{x} \wedge \vec{y}$, are also possible, we just shall discuss them a bit later. For now we shall treat each expression of the form $\vec{x} \wedge \vec{y} \wedge \vec{z}$ as a single 3-argument operation defined by (5.23).

Rewriting (5.21) and (5.22) in the wedge product notation, we obtain

(nonzero triple wedge products of basis vectors in 3D)

$$\begin{aligned}
\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3 = \vec{e}_2 \wedge \vec{e}_3 \wedge \vec{e}_1 = \vec{e}_3 \wedge \vec{e}_1 \wedge \vec{e}_2 = 1 \\
\vec{e}_2 \wedge \vec{e}_1 \wedge \vec{e}_3 = \vec{e}_1 \wedge \vec{e}_3 \wedge \vec{e}_2 = \vec{e}_3 \wedge \vec{e}_2 \wedge \vec{e}_1 = -1
\end{aligned} \tag{5.24}$$

noticing the same anticommutativity property as in (5.21) and (5.22). Now (5.24) contains all six possible ordering variations of $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$. All other triplets built from $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$ would necessarily have one of the vectors repeated, e.g.

$\vec{e}_1$, $\vec{e}_1$, $\vec{e}_2$. But we already know that the volumes spanned by such triplets with vector repetitions are zero. Therefore (5.24) lists all nonzero triple wedge products of the basis vectors, all other products are zero.

Goal 5.8. Obtain a component-based expression for $\vec{x} \wedge \vec{y} \wedge \vec{z}$, which is the counterpart of (5.6).

The expressions (5.24) along with the knowledge that all other triple wedge products of basis vectors are zero allow us to compute a triple wedge product of any 3 vectors $\vec{x}$, $\vec{y}$, $\vec{z}$, by expanding these vectors in the basis and then applying the trilinearity:

$$\vec{x} \wedge \vec{y} \wedge \vec{z} = \left(\sum_i x_i \vec{e}_i\right) \wedge \left(\sum_j y_j \vec{e}_j\right) \wedge \left(\sum_k z_k \vec{e}_k\right) = \sum_{i,j,k} x_i y_j z_k \varepsilon_{ijk} \quad (5.25a)$$

or

(triple wedge product in an orthonormal 3D basis in tensor index notation)

$$\vec{x} \wedge \vec{y} \wedge \vec{z} = x^i \vec{e}_i \wedge y^j \vec{e}_j \wedge z^k \vec{e}_k = x^i y^j z^k \cdot (\vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k) = x^i y^j z^k \varepsilon_{ijk} \quad (5.25b)$$

where

$$\varepsilon_{ijk} = \vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k \quad (5.25c)$$

is the 3-dimensional Levi–Civita symbol. The values ε_{ijk} can be described by two simple rules:

1. $\varepsilon_{123} = \vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3 = 1$
2. Antisymmetry in each pair of indices: swapping the values of any two of the three indices in ε_{ijk} changes the sign. E.g.

$$\varepsilon_{jik} = \vec{e}_j \wedge \vec{e}_i \wedge \vec{e}_k = -(\vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k) = -\varepsilon_{ijk}$$

In particular, from the second rule we obtain:

- $\varepsilon_{213} = -\varepsilon_{123} = -1$
- $\varepsilon_{231} = -\varepsilon_{213} = 1$ (by continuing from the previous result)
- $\varepsilon_{112} = -\varepsilon_{112}$ (where the two swapped indices are both equal to 1), therefore $\varepsilon_{112} = 0$
- etc.

▷ A simpler way to describe the values of the components of the Levi–Civita symbol is to say that it is equal to the sign of the permutation of the indices if all indices are different and zero if any of the indices are repeated.

Goal 5.8 completed. ■

The readers are encouraged to convince themselves that the formula (5.25) coincides with the matrix determinant formula for

$$\det \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix}$$

We thereby have obtained that the volume of a parallelepiped spanned by the three vectors $\vec{x}$, $\vec{y}$, $\vec{z}$ is

(determinant of a matrix gives the volume of a parallelepiped spanned by the vectors represented by matrix columns)

$$V(\vec{x}, \vec{y}, \vec{z}) = \vec{x} \wedge \vec{y} \wedge \vec{z} = \det \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix} \tag{5.26}$$

▷ Comparing (5.26) to (5.7), where we had a determinant of a 2×2 matrix, it doesn't take too much to realize how (5.26) generalizes to higher dimension counts: a determinant gives the hypervolume of a "higher-dimensional parallelepiped" in a positively-oriented orthonormal basis.

5.5 Bivectors

We have introduced the "triple exterior product" $\vec{x} \wedge \vec{y} \wedge \vec{z}$, however we still do not know the meaning of the exterior product $\vec{x} \wedge \vec{y}$ of two vectors in 3D. Let's try to come up with some reasonable idea.

5.5.1 Exterior product in 3D

The first idea could be to simply use the formula (5.17). However here immediately comes a problem. In 2D there were only 2 possible orientations between the 2 vectors, and we could have a simple rule, which angle is positive, which one is negative. However in 3D there are infinitely many such orientations and it's not clear at all, which of those should count as positive and which as negative.

Yet another way to arrive at the same problem is by postulating something like $\vec{e}_1 \wedge \vec{e}_2 = \vec{e}_2 \wedge \vec{e}_3 = \vec{e}_3 \wedge \vec{e}_1 \stackrel{?}{=} 1$ and trying to derive the remaining values $\vec{e}_i \wedge \vec{e}_j$ by anticommutativity. This immediately conflicts with bilinearity. E.g. consider the product $(\vec{e}_1 + \vec{e}_2) \wedge \vec{e}_3$. By bilinearity we should obtain

$$(\vec{e}_1 + \vec{e}_2) \wedge \vec{e}_3 = \vec{e}_1 \wedge \vec{e}_3 + \vec{e}_2 \wedge \vec{e}_3 \stackrel{?}{=} -1 + 1 = 0 \tag{5.27a}$$

But having zero as a result here is weird. We rather would expect $\pm\sqrt{2}$, since $|\vec{e}_1 + \vec{e}_2| = \sqrt{2}$ and the vectors $\vec{e}_1 + \vec{e}_2$ and $\vec{e}_3$ are orthogonal. Also, if we rotate the vectors around the $\vec{e}_3$ axis, so that the former $\vec{e}_1 + \vec{e}_2$ coincides in direction with one of the basis vectors $\vec{e}_1$ or $\vec{e}_2$, our exterior product becomes $\sqrt{2}\vec{e}_1 \wedge \vec{e}_3$ or $\sqrt{2}\vec{e}_2 \wedge \vec{e}_3$ respectively and *will* deliver $\pm\sqrt{2}$ as its result value.

On the other hand, we are also going to get

$$(\vec{e}_2 - \vec{e}_1) \wedge \vec{e}_3 = \vec{e}_2 \wedge \vec{e}_3 - \vec{e}_1 \wedge \vec{e}_3 \stackrel{?}{=} 1 - (-1) = 2 \tag{5.27b}$$

whereas we also rather expected $\pm\sqrt{2}$. So, our definition conflicts with bilinearity and is not invariant under rotations either. Not very useful.

Equations (5.27) however give us a hint: different exterior products $\vec{e}_i \wedge \vec{e}_j$ (such as the ones found in the intermediate expressions of (5.27)) should add as if they were vectors, same as the vectors in the parentheses in the left-hand sides of (5.27). In other words, the result of an exterior product of two 3D vectors shouldn't be a one-dimensional scalar but an element of some multidimensional linear space. We still (so far) expect this exterior product to be bilinear:

$$\begin{aligned}
\vec{x} \wedge \vec{y} &= (x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3) \wedge (y_1\vec{e}_1 + y_2\vec{e}_2 + y_3\vec{e}_3) \\
&= \left(\sum_{i=1}^{3} x_i\vec{e}_i\right) \wedge \left(\sum_{j=1}^{3} y_j\vec{e}_j\right) = \sum_{i=1}^{3}\sum_{j=1}^{3}(x_i\vec{e}_i \wedge y_j\vec{e}_j) \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3}\Big(x_iy_j \cdot (\vec{e}_i \wedge \vec{e}_j)\Big) = \sum_{i=1}^{3}\sum_{j=1}^{3}(x_iy_j \cdot e_{ij}) \\
&= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \cdot \begin{bmatrix} \vec{e}_1 \wedge \vec{e}_1 & \vec{e}_1 \wedge \vec{e}_2 & \vec{e}_1 \wedge \vec{e}_3 \\ \vec{e}_2 \wedge \vec{e}_1 & \vec{e}_2 \wedge \vec{e}_2 & \vec{e}_2 \wedge \vec{e}_3 \\ \vec{e}_3 \wedge \vec{e}_1 & \vec{e}_3 \wedge \vec{e}_2 & \vec{e}_3 \wedge \vec{e}_3 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\
&= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \cdot \begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \\ e_{31} & e_{32} & e_{33} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}
\end{aligned} \tag{5.28a}$$

where we introduced the notation $e_{ij} = \vec{e}_i \wedge \vec{e}_j$. This exterior product probably should also still be anticommutative, like in 2D, so $e_{ij} = \vec{e}_i \wedge \vec{e}_j = -(\vec{e}_j \wedge \vec{e}_i) = -e_{ji}$, and we can rewrite the latter result as

$$\vec{x} \wedge \vec{y} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \cdot \begin{bmatrix} 0 & e_{12} & -e_{31} \\ -e_{12} & 0 & e_{23} \\ e_{31} & -e_{23} & 0 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \tag{5.28b}$$

where, like in (5.5), the bilinear form matrix is antisymmetric, but, differently from (5.5), the elements $e_{ij} = \vec{e}_i \wedge \vec{e}_j$ of that matrix (and respectively the values yielded by the bilinear form) are not numbers but elements of the linear space that we are constructing. Thus $\vec{x} \wedge \vec{y}$ given by (5.28) is still bilinear and anticommutative.

Rewriting (5.28b) in tensor index notation and then in the component form, we have

(basis expansion of an exterior product of two 3D vectors)

$$\begin{aligned}
\vec{x} \wedge \vec{y} &= x^i\vec{e}_i \wedge y^j\vec{e}_j = x^iy^j \cdot (\vec{e}_i \wedge \vec{e}_j) = x^iy^j \cdot e_{ij} \\
&= (x^1y^2 - x^2y^1) \cdot e_{12} + (x^2y^3 - x^3y^2) \cdot e_{23} + (x^3y^1 - x^1y^3) \cdot e_{31}
\end{aligned} \tag{5.29}$$

It looks like we have constructed another 3D linear space (of a yet unknown meaning/interpretation) with the basis elements $e_{12} = \vec{e}_1 \wedge \vec{e}_2$, $e_{23} = \vec{e}_2 \wedge \vec{e}_3$, $e_{31} = \vec{e}_3 \wedge \vec{e}_1$. The elements of this space are referred to as *bivectors* (or *2-vectors*) and the space itself as *bivector space* (or *2-vector space*).

▷ The resulting expression in (5.29) looks very much like a cross product of vectors, and it is indeed closely related to the cross product, but *is not* a cross product. In particular $\vec{e}_1 \wedge \vec{e}_2 = e_{12} \neq \vec{e}_3$, $\vec{e}_2 \wedge \vec{e}_3 = e_{23} \neq \vec{e}_1$ and $\vec{e}_3 \wedge \vec{e}_1 = e_{31} \neq \vec{e}_2$, it is a different (although still three-dimensional) linear space (the bivector space) and a different basis. This will be further discussed in Section 5.8.

So suppose $e_{13} = \vec{e}_1 \wedge \vec{e}_3$ and $e_{23} = \vec{e}_2 \wedge \vec{e}_3$ are two different elements of the bivector space. In order to satisfy the first equalities of (5.27a) and (5.27b), we would like both of e_{13} and e_{23} to have unit length:

$$|e_{13}| = |e_{23}| = 1$$

and to be mutually orthogonal, so that

$$|e_{13} + e_{23}|^2 = |e_{13}|^2 + |e_{23}|^2 = 2$$

and

$$|e_{13} + e_{23}| = \sqrt{2}$$

which then would match the expectation

$$\sqrt{2} = |(\vec{e}_1 + \vec{e}_2) \wedge \vec{e}_3| \overset{(5.27a)}{=} |\vec{e}_1 \wedge \vec{e}_3 + \vec{e}_2 \wedge \vec{e}_3| = \sqrt{2}$$

We seem to be getting somewhere. So we generally let different products $\vec{e}_i \wedge \vec{e}_j$ yield unit-length elements e_{ij} of the bivector space:

$$|\vec{e}_i \wedge \vec{e}_j| = |e_{ij}| = \begin{cases} 1 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{5.30}$$

where different nonzero products $e_{ij} = \vec{e}_i \wedge \vec{e}_j$ $(i \neq j)$ are mutually orthogonal:

$$e_{ij} \perp e_{kl} \quad (\text{where } i \neq j, k \neq l) \tag{5.31a}$$

with the exception of

$$e_{ij} = -e_{ji} \quad \forall i, j \tag{5.31b}$$

in which case we obtain mutually opposite rather than orthogonal elements.

▷ More precisely, (5.31a) should have been written under the additional condition $\{i, j\} \neq \{k, l\}$, that is that the (unordered) set of indices of the first bivector e_{ij} is different from the set of indices of the second bivector e_{kl}. We just specified the same in words, by saying that e_{ij} and e_{kl} are different products of the basis vectors and that there is an exception given by (5.31b).

We thus have come up with the proposal to use (5.28) and (5.29) as the exterior product of two 3D vectors, which we constructed by trying to ensure that the absolute value of such exterior product still delivers the unsigned area of the parallelogram spanned by $\vec{x}$ and $\vec{y}$. Let's see if we managed to achieve that.

Goal 5.9. Demonstrate that $|\vec{x} \wedge \vec{y}|$ (where $\vec{x} \wedge \vec{y}$ is by (5.29)) delivers the unsigned area of the parallelogram spanned by $\vec{x}$ and $\vec{y}$. *Hint:* demonstrate rotational invariance of $|\vec{x} \wedge \vec{y}|$.

Let

$$\zeta = \vec{x} \wedge \vec{y} = \zeta_{12} \cdot e_{12} + \zeta_{23} \cdot e_{23} + \zeta_{31} \cdot e_{31} \tag{5.32}$$

where $\vec{x} \wedge \vec{y}$ is by (5.29). Since bivectors e_{ij} are mutually orthogonal,

(absolute magnitude of a bivector in an orthonormal basis)

$$\begin{aligned}|\zeta|^2 &= |\zeta_{12}\cdot e_{12}|^2 + |\zeta_{23}\cdot e_{23}|^2 + |\zeta_{31}\cdot e_{31}|^2 \\ &= |\zeta_{12}|^2\cdot 1 + |\zeta_{23}|^2\cdot 1 + |\zeta_{31}|^2\cdot 1 = (\zeta_{12})^2 + (\zeta_{23})^2 + (\zeta_{31})^2 \end{aligned} \tag{5.33}$$

If $|\zeta|$ given by (5.33) is equal to the unsigned area of the parallelogram spanned by $\vec{x}$ and $\vec{y}$ indeed, it must be invariant under space rotations. In that case we might be able to rotate the vectors $\vec{x}$ and $\vec{y}$ into a particularly convenient position for demonstrating that $|\zeta|$ is equal to the parallelogram's area.

So our plan is the following:

1. demonstrate that $|\zeta|$ is invariant under rotations,
2. rotate $\vec{x}$ and $\vec{y}$ into one particularly convenient position and demonstrate that $|\zeta|$ is equal to the parallelogram's area in that position.

As the first point goes, by the discussion in Subsection 3.4.4, any rotational transformation can be decomposed into rotations done within coordinate planes. It would therefore suffice to demonstrate the invariance of $|\zeta|$ under rotations within coordinate planes. We can just demonstrate the invariance of $|\zeta|$ under rotations in the $\vec{e}_1$, $\vec{e}_2$ plane. The other coordinate planes are not essentially different and the respective invariance can follow by the same argument.

Goal 5.9a. Demonstrate the invariance of $|\zeta|$ under rotations in the $\vec{e}_1$, $\vec{e}_2$ plane.

Apparently, x_3 and y_3 are unchanged by a rotation in the $\vec{e}_1$, $\vec{e}_2$ plane, while the remaining coordinates are changed, turning into some new values x_1', x_2', y_1', y_2'. Looking at (5.29), we notice that $z_{12} = x_1y_2 - x_2y_1$ is unchanged due to the rotational invariance of the 2D exterior product. The remaining components of the rotated bivector ζ' yield

$$\begin{aligned}(\zeta_{23}')^2 + (\zeta_{31}')^2 &= (x_2'y_3 - x_3y_2')^2 + (x_3y_1' - x_1'y_3)^2 \\ &= (x_2')^2(y_3)^2 - 2x_2'y_2'x_3y_3 + (y_2')^2(x_3)^2 \\ &\qquad + (y_1')^2(x_3)^2 - 2x_1'y_1'x_3y_3 + (x_1')^2(y_3)^2 \\ &= (y_3)^2\big((x_1')^2 + (x_2')^2\big) + (x_3)^2\big((y_1')^2 + (y_2')^2\big) - 2x_3y_3(x_1'y_1' + x_2'y_2')\end{aligned}$$

where we can notice that

$$\begin{aligned}(x_1')^2 + (x_2')^2 &= (x_1)^2 + (x_2')^2 \\ (y_1')^2 + (y_2')^2 &= (y_1)^2 + (y_2')^2 \\ x_1'y_1' + x_2'y_2' &= x_1y_1 + x_2y_2\end{aligned}$$

since the first two lines represent the lengths of the projections of $\vec{x}$ and $\vec{y}$ onto the $\vec{e}_1$, $\vec{e}_2$ plane, and the third line represents the dot product of these projections. All 3 values are invariant under rotations in the $\vec{e}_1$, $\vec{e}_2$ plane, thereby so is the value of $(\zeta_{23})^2 + (\zeta_{31})^2$, and respectively so is $|\zeta|$.

Goal 5.9a completed. ■

Goal 5.9b. Complete Goal 5.9.

Now as the "particularly convenient position" for vectors $\vec{x}$ and $\vec{y}$ we shall choose the one where both $\vec{x}$ and $\vec{y}$ are lying in the $\vec{e}_1$, $\vec{e}_2$ plane, which apparently

is reachable by rotational transformations of the space for any initial $\vec{x}$, $\vec{y}$. In this case $x_3 = y_3 = 0$, therefore $\zeta_{23} = \zeta_{31} = 0$ and (5.29) effectively collapses into the 2D exterior product:

$$|\zeta| = |\vec{x} \wedge \vec{y}| = |(x_1 y_2 - x_2 y_1) \cdot e_{12}| = |(x_1 y_2 - x_2 y_1)| \cdot 1 = |(x_1 y_2 - x_2 y_1)|$$

thereby yielding the unsigned area of the parallelogram spanned by $\vec{x}$, $\vec{y}$.

Goal 5.9b completed. ■

Goal 5.9 completed. ■

▷ It is not difficult to realize that $|\zeta|$ is invariant under any orthogonal transformations of the space. By the discussion in Subsection 3.4.4, any orthogonal transformation is decomposable into rotations and a mirroring reflection along a chosen coordinate axis. Let's choose $\vec{e}_3$ as the mirroring direction.

The sole effect of inverting the signs of x_3 and y_3 in (5.29) is the inversion of the signs of ζ_{23} and ζ_{31}. Thereby (5.33) is unchanged. It follows that $|\zeta|$ is invariant under orthogonal transformations of the space.

The previous discussion suggests that any exterior product $\vec{x} \wedge \vec{y}$ of 3D vectors $\vec{x}$ and $\vec{y}$ can be interpreted as an *oriented area* of the parallelogram spanned by the vectors $\vec{x}$, $\vec{y}$. Indeed, we can always apply rotational transformations of the 3D vector space to bring $\vec{x}$ and $\vec{y}$ into the $\vec{e}_1$, $\vec{e}_2$ coordinate plane, where the exterior product $\vec{x} \wedge \vec{y}$ of 3D vectors $\vec{x}$ and $\vec{y}$ essentially turns into a 2D exterior product, which in turn has this parallelogram-based interpretation.

Except that in 2D the area orientation was specified in the form of the area's sign, which in turn was related to the sign of the angle θ in (5.17). However we cannot do the same in 3D: depending on from which side we are looking at the plane spanned by $\vec{x}$ and $\vec{y}$, one and the same angle θ can be positive or negative. So, in 3D we have to explicitly refer to the rotation direction associated with this angle. Thus,

3D exterior product $\vec{x} \wedge \vec{y}$ defines an oriented planar area in the 3D space, determined by the following:

- the absolute magnitude $|\vec{x} \wedge \vec{y}|$ is the unsigned area of the parallelogram spanned by $\vec{x}$ and $\vec{y}$
- the area's orientation in 3D is given by
  - the plane spanned by $\vec{x}$ and $\vec{y}$
  - the area's orientation within that plane, which can be expressed by the rotation direction going in the shortest way from $\vec{x}$ to $\vec{y}$.

Fig. 5.11 illustrates.

▷ Notice that expressing the area orientation via rotation direction is nothing more than a visualization convention: there is no actual rotation happening, it's just a way for us to visually think of the difference between $\vec{x} \wedge \vec{y}$ and $\vec{y} \wedge \vec{x}$.

It is also common to visualize oriented areas by plane normals, the length of the normal determining the absolute magnitude of the area, while the direction of the normal simultaneously determining the plane as well as the area orientation, opposite normal directions corresponding to two opposite orientations of the area.

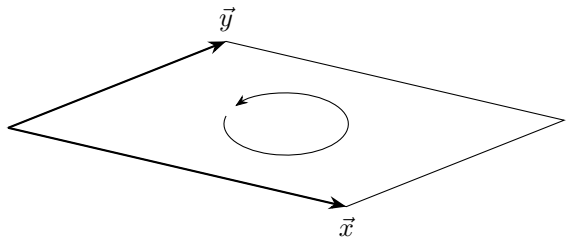


Figure 5.11: Geometric interpretation of the exterior product $\vec{x} \wedge \vec{y}$ in 3D space.

This idea is however associated with the cross product of vectors and works only in 3D, while defining the area orientation by rotation is universal.

5.5.2 Bivector space in 3D

We have developed a geometric interpretation of the exterior product $\vec{x} \wedge \vec{y}$ in 3D. However, by our previous construction, the exterior products $\vec{x} \wedge \vec{y}$ of 3D vectors $\vec{x}$ and $\vec{y}$ are elements of the bivector space, which is a linear space spanned by the basis elements $e_{12} = \vec{e}_1 \wedge \vec{e}_2$, $e_{23} = \vec{e}_2 \wedge \vec{e}_3$, $e_{31} = \vec{e}_3 \wedge \vec{e}_1$. We would like to obtain further interpretations, refining our picture of the entire bivector space, covering the elements of the space as well as the linear operations on those elements.

Since the bivector space is a linear space, we can also obtain further elements of that space by simply constructing their linear combinations, e.g. $\vec{x} \wedge \vec{y} + \vec{u} \wedge \vec{v}$. Therefore it doesn't seem to be immediately guaranteed that a random bivector ζ (which is a randomly picked element of the bivector space) is representable as an exterior product of two 3D vectors $\zeta = \vec{x} \wedge \vec{y}$. We either need to demonstrate the latter, or otherwise suggest an interpretation for the remaining bivectors, which would not be representable as exterior products.

Goal 5.10. Find whether any bivector ζ is representable as an exterior product of two 3D vectors.

Thinking about this question, we can realize that the way how we earlier demonstrated the invariance of the absolute magnitude of a bivector $\zeta = \vec{x} \wedge \vec{y}$ under rotations also suggests that any element of the bivector space can be interpreted as a parallelogram area. The idea is that we can use rotations to make the ζ_{23} and ζ_{31} components vanish, effectively leaving us just with a 2D area. Although, there are some fine details with that: we don't know how to rotate an arbitrary bivector ζ, which is not a priori given as an exterior product of two 3D vectors. Or, more precisely speaking, we don't know how this arbitrary bivector ζ transforms under rotations of the 3D vector space, or, equivalently, under rotations of the 3D vector space basis.

This is what we shall now find out. Similarly to the previous argument about the rotational invariance of $|\vec{x} \wedge \vec{y}|$, it suffices to consider rotations of an orthonormal basis in one of the coordinate planes.

Goal 5.10a. Find how the components of a generic bivector ζ (not given a priori in the $\zeta = \vec{x} \wedge \vec{y}$ form) transform under rotations of an orthonormal 3D vector basis in the $\vec{e}_1$, $\vec{e}_2$ plane.

Let the basis transformation be given by (5.13), while $\vec{e}_3$ is unchanged by the rotation. The transformations (5.14) are perfectly applicable in 3D, since they only rely on bilinearity and anticommutativity of the exterior product and do not care about the number of dimensions or about whether the exterior product's values are real numbers or linear space elements. It follows that ζ_{12} is unchanged by this rotation.

Substituting (5.13) into $\vec{e}_2{}' \wedge \vec{e}_3$ and $\vec{e}_3 \wedge \vec{e}_1{}'$ yields

$$\begin{aligned}\vec{e}_2{}' \wedge \vec{e}_3 &= (-\vec{e}_1 \sin\alpha + \vec{e}_2 \cos\alpha) \wedge \vec{e}_3 = \vec{e}_2 \wedge \vec{e}_3 \cdot \cos\alpha + \vec{e}_3 \wedge \vec{e}_1 \cdot \sin\alpha \\ \vec{e}_3 \wedge \vec{e}_1{}' &= \vec{e}_3 \wedge (\vec{e}_1 \cos\alpha + \vec{e}_2 \sin\alpha) = -\vec{e}_2 \wedge \vec{e}_3 \cdot \sin\alpha + \vec{e}_3 \wedge \vec{e}_1 \cdot \cos\alpha\end{aligned}$$

So $\vec{e}_2 \wedge \vec{e}_3$ and $\vec{e}_3 \wedge \vec{e}_1$ transform like ordinary basis vectors (in fact, we could have already realized this simply by bilinearity of the exterior product), and respectively ζ_{23} and ζ_{31} transform in a contravariant manner:

$$\begin{aligned}\zeta_{23} &= \zeta_{23} \cos\alpha - \zeta_{31} \sin\alpha \\ \zeta_{31} &= -\zeta_{23} \sin\alpha + \zeta_{31} \cos\alpha\end{aligned}$$

Goal 5.10a completed. ■

It follows that by rotations in the $\vec{e}_1$, $\vec{e}_2$ plane we can zero one of the components ζ_{23} and ζ_{31}, while keeping ζ_{12} intact. Say, we zeroed ζ_{31}, while $\zeta_{12} \neq 0$, $\zeta_{23} \neq 0$. Now let's rotate in the $\vec{e}_3 \wedge \vec{e}_1$ plane. This keeps $\zeta_{31} = 0$ untouched, while we can additionally bring ζ_{23} to zero. We are thus left with a single nonzero component ζ_{12} and effectively with a 2D exterior product in the $\vec{e}_1$, $\vec{e}_2$ plane. This 2D product can be expressed as an exterior product of two vectors $\vec{x}$, $\vec{y}$ lying in the $\vec{e}_1$, $\vec{e}_2$ plane in a trivial manner, e.g. by (5.17).

It follows that any 3D bivector is representable as a product of two 3D vectors $\vec{x} \wedge \vec{y}$ and respectively can be associated with an oriented parallelogram area in the $\vec{x}$, $\vec{y}$ plane.

Goal 5.10 completed. ■

Furthermore, we can realize that in 2D we have been speaking just of signed areas in the 2D plane, without the need to refer to the specific form of the 2D shape. This shape can be a parallelogram, or it could be something else. Similarly, we can associate any 3D bivector ζ with an oriented planar shape of unspecified form. In that case $|\zeta|$ gives the unsigned area of that shape, while the plane, in which the shape is located, is given by the distribution of $|\zeta|$ among $|\zeta_{12}|$, $|\zeta_{23}|$, $|\zeta_{31}|$. We might conjecture that $|\zeta_{12}|$, $|\zeta_{23}|$, $|\zeta_{31}|$ give the unsigned areas of the projections of the shape on the respective coordinate planes.

Goal 5.11. Develop geometric interpretation of the absolute magnitudes of the components ζ_{12}, ζ_{23}, ζ_{31}. *Hint:* assume $\zeta = \vec{x} \wedge \vec{y}$ and notice that ζ_{12} is independent of x_3 and y_3.

Let $\vec{x}$, $\vec{y}$ be two 3D vectors such that $\zeta = \vec{x} \wedge \vec{y}$. Looking at $\zeta_{12} = x_1 y_2 - x_2 y_1$ we notice that this formula is identical to (5.7). That is z_{12} is the 2D signed area formed by the projections of $\vec{x}$ and $\vec{y}$ onto the $\vec{e}_1$, $\vec{e}_2$ plane (Fig. 5.12). The same can be said about ζ_{23} and ζ_{31}. Therefore $|\zeta_{12}|$, $|\zeta_{23}|$, $|\zeta_{31}|$ are the projections of the area $\vec{x} \wedge \vec{y}$ onto the respective coordinate planes.

Now, given whatever planar shape representing an arbitrary bivector ζ, by the additivity of the area, we can split this shape into smaller parallelograms up

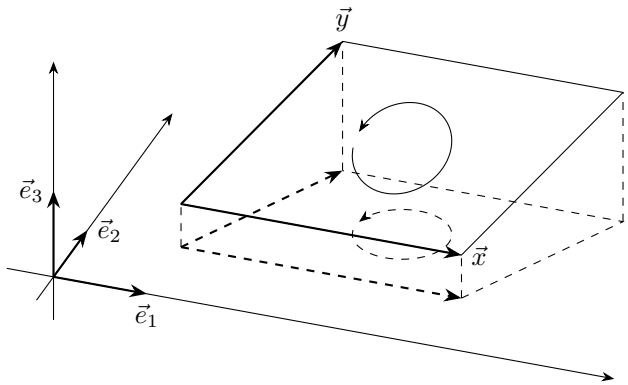


Figure 5.12: An oriented parallelogram implied by $\vec{x} \wedge \vec{y}$ and its projection onto the $\vec{e}_1$, $\vec{e}_2$ coordinate plane.

to any arbitrary given precision.[3] Each of this parallelograms can be projected separately onto a given coordinate plane, justifying that the absolute magnitude of the respective component of ζ is the unsigned area of the projection of the entire shape onto that plane.

Goal 5.11 completed. ■

It remains to develop geometric interpretation of the signs of the components of ζ.

Goal 5.12. Develop geometric interpretation of the signs of the components ζ_{12}, ζ_{23}, ζ_{31}.

Looking at Fig. 5.12 we can notice that the sign of ζ_{12} is determined by the relative orientation of the projections of $\vec{x}$, $\vec{y}$ onto the $\vec{e}_1$, $\vec{e}_2$ plane. Effectively, this is simply the projection of the rotation direction specified in the $\vec{x}$, $\vec{y}$ plane onto $\vec{e}_1$, $\vec{e}_2$ plane. Thus, the signs of ζ_{12}, ζ_{23}, ζ_{31} are effectively specifying the rotation direction's projections.

Goal 5.12 completed. ■

We thus arrive at the following shape-abstract interpretation of an arbitrarily given bivector ζ in the 3D space.

The oriented area implied by a 3D bivector ζ consists of the following:

- the absolute magnitude $|\zeta|$ is the unsigned area of a planar shape of an unspecified form (any shape having the area of the necessary magnitude is good enough)
- the area's orientation in 3D is given by
  - the plane in which the planar shape is lying (this plane is always the same, independently of the specific form of the shape)

[3] Effectively, we are constructing Riemann sums approximating the area of the shape. More details on that can be found in Section 6.8.

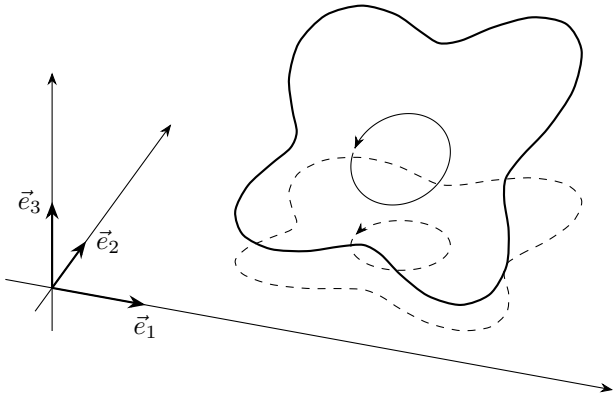


Figure 5.13: A free-form oriented shape representing a bivector ζ (solid) and its projection $\zeta_{12} \cdot (\vec{e}_1 \wedge \vec{e}_2)$ onto the $\vec{e}_1$, $\vec{e}_2$ coordinate plane (dashed).

- the area's orientation within that plane, which can be expressed as one of the two possible rotation directions in that plane.

Fig. 5.13 illustrates.

Having constructed geometric interpretation of 3D bivectors, we now need to develop geometric interpretations of linear operations (addition and multiplication by a scalar) on bivectors. Multiplication of a bivector ζ by a scalar λ is trivially interpreted as simply scaling the area $|\zeta|$ by the factor $|\lambda|$, additionally inverting the area orientation if $\lambda < 0$. It remains to construct the interpretation of bivector addition.

Goal 5.13. Construct a geometric interpretation of bivector addition. *Hint:* look at (5.27a).

So, what does it mean to add two oriented areas in 3D? If it was 2D, we simply would have added the signed real numbers representing those areas. In 3D we could do effectively the same if both areas are lying in one and the same plane, but we cannot do this if they are lying in different planes: this was exactly our problem with (5.27), and this is exactly we promoted the areas represented by real numbers to the ones represented by bivectors.

The formulas (5.27) however simultaneously give us a hint. E.g. looking at the first equality in the formula (5.27a) we notice that it effectively reexpresses the sum of bivectors $\vec{e}_1 \wedge \vec{e}_3$ and $\vec{e}_2 \wedge \vec{e}_3$ as a single exterior product $(\vec{e}_1 + \vec{e}_2) \wedge \vec{e}_3$:

$$\vec{e}_1 \wedge \vec{e}_3 + \vec{e}_2 \wedge \vec{e}_3 = (\vec{e}_1 + \vec{e}_2) \wedge \vec{e}_3$$

This suggests an interpretation of bivector addition for the cases when we can factor a common vector out of the sum: by factoring out this common vector we have an expression, which is directly interpretable as a parallelogram area, and this is how we can interpret such sums.

The question is only, if we there is always such common vector factor in the general case of adding up two arbitrary bivectors $\zeta + \eta$. By the previous argument we can represent each of the bivectors as an exterior product: $\zeta +$

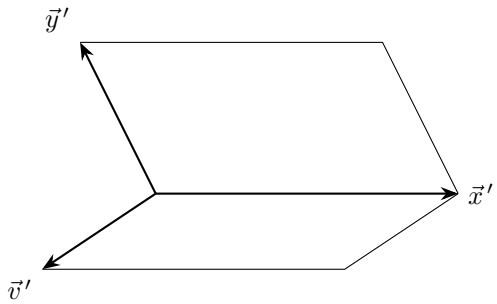


Figure 5.14: A bivector sum $\zeta + \eta = \vec{x} \wedge \vec{y} + \vec{u} \wedge \vec{v}$ transformed to a sum with a common factor $\vec{x}\,' \wedge \vec{y}\,' + \vec{x}\,' \wedge \vec{v}\,'$.

$\eta = \vec{x} \wedge \vec{y} + \vec{u} \wedge \vec{v}$, but what then? It doesn't seem like there is a common vector factor. We should however recall that there are infinitely many different exterior products, all delivering one and the same bivector equal to ζ, and there are infinitely many different exterior products, all delivering one and the same bivector equal to η. Maybe some of those do have common factors?

Thinking a bit further, we can realize that such common factor is nothing more than a vector lying in the intersection line of the planes of the bivectors $\zeta = \vec{x} \wedge \vec{y}$ and $\eta = \vec{u} \wedge \vec{v}$. We could therefore e.g. simply rotate $\vec{x}$ and $\vec{y}$ until $\vec{x}$ lies in this intersection line, and we could rotate $\vec{u}$ and $\vec{v}$ until $\vec{v}$ lies in this intersection line. This transformation doesn't change the bivectors $\zeta = \vec{x} \wedge \vec{y}$ and $\eta = \vec{u} \wedge \vec{v}$. Now it remains to rescale $\vec{x}$ making it coincide with $\vec{u}$, and rescale $\vec{y}$ by the reciprocal factor to make sure $\vec{x} \wedge \vec{y}$ is unchanged. We thus transformed the 4 vectors $\vec{x}$, $\vec{y}$, $\vec{u}$, $\vec{v}$ to some other 4 vectors $\vec{x}\,'$, $\vec{y}\,'$, $\vec{u}\,'$, $\vec{v}\,'$, such that $\vec{x}\,' = \vec{u}\,'$, and we have

$$\begin{aligned}\zeta + \eta &= \vec{x} \wedge \vec{y} + \vec{u} \wedge \vec{v} = \vec{x}\,' \wedge \vec{y}\,' + \vec{u}\,' \wedge \vec{v}\,' = \vec{x}\,' \wedge \vec{y}\,' + \vec{x}\,' \wedge \vec{v}\,' \\ &= \vec{x}\,' \wedge (\vec{y}\,' + \vec{v}\,')\end{aligned}$$

(Fig. 5.14). We thereby represented the sum $\zeta + \eta$ as an exterior product of two vectors, which thereby defines an oriented area in 3D.

▷ By (5.29), it follows that any element of the linear space implied by (5.29) is representable as an exterior product of two vectors, and thus can be interpreted as an oriented area in 3D.

As mentioned earlier and as we shall see later, this doesn't hold for higher dimension counts.

Goal 5.13 completed. ■

5.5.3 Bivectors in 2D

So the primary difference between the exterior product in 3D from 2D is that instead of scalar values it produces elements from another linear space, which is the bivector space spanned by the basis bivectors $e_{12} = \vec{e}_1 \wedge \vec{e}_2$, $e_{23} = \vec{e}_2 \wedge \vec{e}_3$ and $e_{31} = \vec{e}_3 \wedge \vec{e}_1$. We can, however, apply the same thinking in 2D, treating $e_{12} = \vec{e}_1 \wedge \vec{e}_2$ as an element of some linear space rather than as a real number.

In that case we cannot do the final transformations in (5.5) and have to stop at

$$\vec{x} \wedge \vec{y} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{bmatrix} \vec{e}_1 \wedge \vec{e}_1 & \vec{e}_1 \wedge \vec{e}_2 \\ \vec{e}_2 \wedge \vec{e}_1 & \vec{e}_2 \wedge \vec{e}_2 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{bmatrix} 0 & e_{12} \\ -e_{12} & 0 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

with $e_{12} = \vec{e}_1 \wedge \vec{e}_2$. Respectively, (5.7) becomes

(exterior product of two vectors is actually a bivector, not a number)

$$\vec{x} \wedge \vec{y} = (x_1 y_2 - x_2 y_1) \cdot e_{12} = (x_1 y_2 - x_2 y_1) \cdot (\vec{e}_1 \wedge \vec{e}_2) \tag{5.34}$$

the value $x_1 y_2 - x_2 y_1$ thereby turning from directly being the value of $\vec{x} \wedge \vec{y}$ to being the coordinate of the bivector $\vec{x} \wedge \vec{y}$ in a 1D space of 2D bivectors.

Thereby 2D bivectors become a true subset of 3D bivectors. In particular, suppose $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$ is a 3D vector basis. Simultaneously we can consider the 2D subspace of that 3D space, spanned by the vectors $\vec{e}_1$, $\vec{e}_2$. Then $\vec{e}_1 \wedge \vec{e}_2$ denotes one and the same bivector, regardless whether we construct that bivector in the context of the 3D space or its 2D subspace. Similarly, an exterior product $\vec{x} \wedge \vec{y}$ of vectors $\vec{x}$, $\vec{y}$ lying in the $\vec{e}_1$, $\vec{e}_2$ plane denotes one and the same bivector regardless of whether its treated in 2D or 3D context.

It follows that (5.34) is applicable in 3D as well (and in higher dimension counts). In fact, (5.34) emerges simply by bilinearity and anticommutativity of the wedge product, thereby it also holds in nonorthonormal bases.

5.5.4 Dot product of bivectors

The length formula (5.33) uniquely suggests the corresponding dot (inner) product

$$\begin{aligned} \zeta \cdot \eta &= (\zeta_{12} e_{12} + \zeta_{23} e_{23} + \zeta_{31} e_{31}) \cdot (\eta_{12} e_{12} + \eta_{23} e_{23} + \eta_{31} e_{31}) \\ &= \zeta_{12} \eta_{12} + \zeta_{23} \eta_{23} + \zeta_{31} \eta_{31} \end{aligned} \tag{5.35}$$

where $e_{ij} \cdot e_{ij} = 1$ and the remaining dot products of basis bivectors are zero, thereby expressing (5.30) and (5.31) respectively.

▷ We have earlier shown the invariance of $|\zeta|$ under orthogonal transformations of the vector space. It follows that bivector dot products are also invariant under orthogonal transformations, since dot product can be expressed via length as the "defect of the Pythagorean theorem":

$$\zeta \cdot \eta = |\zeta + \eta|^2 - |\zeta|^2 - |\eta|^2$$

▷ A dot product implicitly defines the formal angle θ between the respective linear space elements via

$$\zeta \cdot \eta = |\zeta| \cdot |\eta| \cdot \cos\theta = \sqrt{\zeta \cdot \zeta} \cdot \sqrt{\eta \cdot \eta} \cdot \cos\theta$$

On the other hand, the components of ζ in an orthonormal basis are given by the dot product $\zeta_{ij} = \zeta \cdot e_{ij} = |\zeta| \cdot \cos\theta_{ij}$. Thereby θ_{ij} are the geometric angles between ζ and e_{ij}, and it follows by rotational invariance that θ is the geometric angle between the planes of bivectors ζ and η.

We have seen that the length of a bivector $\zeta = \vec{x} \wedge \vec{y}$ can be related to the dot product of vectors $\vec{x}$ and $\vec{y}$. Indeed, in 2D we had (5.18), which, now that we promoted the exterior product's result from a number to a bivector, should be rewritten as

(relationship between exterior and dot products)

$$(\vec{x} \cdot \vec{y})^2 + |\vec{x} \wedge \vec{y}|^2 = |\vec{x}|^2 \cdot |\vec{y}|^2 \tag{5.36}$$

Apparently, given two 3D vectors $\vec{x}$ and $\vec{y}$, we can rotate them (or the basis) so that they are both lying in the $\vec{e}_1$, $\vec{e}_2$ plane. This wouldn't change any of the scalar values (the dot product and the vector and bivector lengths) involved in (5.36). Upon such rotation we are going to have an essentially 2D setup, where (5.36) is known to hold. It follows that (5.36) holds in 3D as well.

▷ By the same argument, (5.36) holds in any higher dimension count.

But we can also write (5.36) as

$$(\vec{x} \cdot \vec{y})^2 + (\vec{x} \wedge \vec{y})^2 = \vec{x}^2 \cdot \vec{y}^2$$

where

$$\begin{aligned} \vec{x}^2 &= \vec{x} \cdot \vec{x} \\ \vec{y}^2 &= \vec{y} \cdot \vec{y} \\ (\vec{x} \wedge \vec{y})^2 &= (\vec{x} \wedge \vec{y}) \cdot (\vec{x} \wedge \vec{y}) \end{aligned}$$

Thus $|\vec{x} \wedge \vec{y}|^2$ in (5.36) can be seen as a dot square $\zeta^2 = \zeta \cdot \zeta$ of the bivector $\zeta = \vec{x} \wedge \vec{y}$. Thereby (5.36) can be used to express the dot square of an exterior product of two vectors via the dot product and the dot squares of the vectors:

$$\zeta^2 = (\vec{x} \wedge \vec{y})^2 = \vec{x}^2 \cdot \vec{y}^2 - (\vec{x} \cdot \vec{y})^2$$

We would like to know if a similar, more general, relationship exists for a dot product of two different bivectors $\zeta \cdot \eta = (\vec{x} \wedge \vec{y}) \cdot (\vec{u} \cdot \vec{v})$.

Goal 5.14. Express the dot product $\zeta \cdot \eta$ of two bivectors $\zeta = \vec{x} \wedge \vec{y}$ and $\eta = \vec{u} \wedge \vec{v}$ via the dot products and squares of $\vec{x}$, $\vec{y}$, $\vec{u}$ and $\vec{v}$. *Hint:* consider (5.7) in terms of projections, which in turn can be interpreted as dot products.

Now that we agreed that an exterior product of two vectors yields a bivector, even in 2D, we can interpret (5.7) as a formula for the ζ_{12} component of the bivector $\zeta = \vec{x} \wedge \vec{y}$:

$$\zeta_{12} = x_1 y_2 - x_2 y_1 = \det \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix} \tag{5.37}$$

Notice that by (5.29), the formula (5.37) equally holds in 3D (and would hold in any higher dimension count).

Assuming an orthonormal basis, we can interpret ζ_{12} in the left-hand side of (5.37) as projection of the bivector ζ onto the $\vec{e}_1$, $\vec{e}_2$ plane, which is simply given by

$$\zeta_{12} = \zeta \cdot e_{12} = (\vec{x} \wedge \vec{y}) \cdot (\vec{e}_1 \wedge \vec{e}_2)$$

Similarly, we can interpret x_1, x_2, y_1 and y_2 in the right-hand side of (5.37) as projections of the vectors $\vec{x}$ and $\vec{y}$ onto the vectors $\vec{e}_1$ and $\vec{e}_2$:

$$x_1 = \vec{x} \cdot \vec{e}_1 \quad x_2 = \vec{x} \cdot \vec{e}_2 \quad y_1 = \vec{y} \cdot \vec{e}_1 \quad y_2 = \vec{y} \cdot \vec{e}_2$$

Notice that this interpretation holds both in 2D and in 3D (or in higher dimension counts). With that in mind, we can rewrite (5.37) as

$$(\vec{x} \wedge \vec{y}) \cdot (\vec{e}_1 \wedge \vec{e}_2) = \det \begin{bmatrix} \vec{x} \cdot \vec{e}_1 & \vec{y} \cdot \vec{e}_1 \\ \vec{x} \cdot \vec{e}_2 & \vec{y} \cdot \vec{e}_2 \end{bmatrix} \tag{5.38}$$

So, we have expressed a dot product of two bivectors $\vec{x} \wedge \vec{y}$ and $\vec{e}_1 \wedge \vec{e}_2$ via the underlying vectors $\vec{x}$, $\vec{y}$, $\vec{e}_1$ and $\vec{e}_2$. We'd like to generalize our previous result to arbitrary vectors in place of $\vec{e}_1$ and $\vec{e}_2$.

We can recall that rotations (or, more generally, orthogonal transformations) in the vectors space correspond to rotations (or orthogonal transformation) in the bivector space. None of those change the lengths or dot products. We can therefore apply (5.38) to any pair of mutually orthogonal unit-length 3D vectors $\vec{u} \perp \vec{v}$ in place of $\vec{e}_1$ and $\vec{e}_2$:

$$(\vec{x} \wedge \vec{y}) \cdot (\vec{u} \wedge \vec{v}) = \det \begin{bmatrix} \vec{x} \cdot \vec{u} & \vec{y} \cdot \vec{u} \\ \vec{x} \cdot \vec{v} & \vec{y} \cdot \vec{v} \end{bmatrix} \qquad \text{where } \vec{u} \perp \vec{v} \tag{5.39}$$

It is also obvious that multiplying $\vec{u}$ and/or $\vec{v}$ a constant still doesn't break (5.39), thus we don't need to require $\vec{u}$ and $\vec{v}$ to have unit length, but so far the vectors $\vec{u}$ and $\vec{v}$ must still be mutually orthogonal. Intuitively, we should be able to apply the linearity of (5.39) with respect to each of the involved vectors to break the orthogonality requirement. E.g. we could replace $\vec{v}$ by $\vec{v} + \lambda\vec{u}$ (where $\lambda \in \mathbb{R}$), effectively performing a shear transformation of $\vec{u} \wedge \vec{v}$. In order to be able to do that, however, we first need to demonstrate that replacing $\vec{v}$ by $\vec{u}$ in (5.39) doesn't break the equation:

$$(\vec{x} \wedge \vec{y}) \cdot (\vec{u} \wedge \vec{u}) = \det \begin{bmatrix} \vec{x} \cdot \vec{u} & \vec{y} \cdot \vec{u} \\ \vec{x} \cdot \vec{u} & \vec{y} \cdot \vec{u} \end{bmatrix} \tag{5.40}$$

But it indeed doesn't break the equation, since both sides of the equation simply turn to zero in this case!

At this point we may realize that we even don't need to involve the shear transformation and can argue the whole thing in a simpler manner by noticing that, by (5.40), the equation (5.38) holds not just for any pair of mutually orthogonal basis vectors $\vec{e}_i$, $\vec{e}_j$ where $i \neq j$, but also in case $i = j$:

$$(\vec{x} \wedge \vec{y}) \cdot (\vec{e}_i \wedge \vec{e}_j) = \det \begin{bmatrix} \vec{x} \cdot \vec{e}_i & \vec{y} \cdot \vec{e}_i \\ \vec{x} \cdot \vec{e}_j & \vec{y} \cdot \vec{e}_j \end{bmatrix} \quad \forall i, j \tag{5.41}$$

Applying the bilinearity of (5.41) with respect to the vector pair $\vec{e}_i$, $\vec{e}_j$ (that is, taking an arbitrary vector $\vec{u}$ and expressing it as a linear combination of basis vectors, and doing the same with vector $\vec{v}$) yields

(bivector dot product in terms of the dot products of underlying vectors)

$$\zeta \cdot \eta = (\vec{x} \wedge \vec{y}) \cdot (\vec{u} \wedge \vec{v}) = \det \begin{bmatrix} \vec{x} \cdot \vec{u} & \vec{y} \cdot \vec{u} \\ \vec{x} \cdot \vec{v} & \vec{y} \cdot \vec{v} \end{bmatrix} \tag{5.42}$$

that is (5.39) holds for any vectors.

Goal 5.14 completed. ■

▷ Letting $\eta = \zeta$ in (5.42), that is $\vec{u} = \vec{x}$, $\vec{v} = \vec{y}$, yields

$$\begin{aligned} |\vec{x} \wedge \vec{y}|^2 &= (\vec{x} \wedge \vec{y}) \cdot (\vec{x} \wedge \vec{y}) = \det \begin{bmatrix} \vec{x} \cdot \vec{x} & \vec{y} \cdot \vec{x} \\ \vec{x} \cdot \vec{y} & \vec{y} \cdot \vec{y} \end{bmatrix} = \det \mathbf{G} \\ &= |\vec{x}|^2 \cdot |\vec{y}|^2 - (\vec{x} \cdot \vec{y})^2 \end{aligned} \tag{5.43}$$

(where $\mathbf{G}$ is the Gram matrix of vectors $\vec{x}$, $\vec{y}$), which is exactly the statement of (5.36). Thus (5.36) is essentially the particular case of (5.42) at $\eta = \zeta$.

▷ Now that we introduced (5.34) and the bivector interpretation of exterior products of two 2D vectors, we may lack a formal way to distinguish between two different bivector orientations in 2D. Previously this orientation was given by the sign of (5.7), but now (5.7) is interpreted as (5.37), that is simply as a bivector component is some specific basis, and it's a bit awkward to relate the bivector orientation to the sign of a bivector component.

A more elegant formal way to determine the bivector orientation sign in 2D is now offered by (5.42). By substituting a normal basis bivector for η into (5.42), we can obtain the projection of ζ onto η. More generally, the sign of $\zeta \cdot \eta$ allows to compare the orientations of ζ and η. If we know that η is positively oriented (which ultimately needs to be postulated for at least one bivector, in order to agree on the orientation sign definition), then so are all ζ such that $\zeta \cdot \eta > 0$. So, if we let $\eta = e_{12} = \vec{e}_1 \wedge \vec{e}_2$ and we know that e_{12} is positively oriented, we can simply check the positive orientation of ζ by checking the sign of the dot product $\zeta \cdot \eta$. Notice that by (5.42) this gives the same result as by using (5.7) in a positively-oriented orthonormal basis (which is what (5.7) assumed).

5.5.5 Higher dimension counts

In order to generalize the bivector idea to higher dimension counts, we can just reuse the approach from 3D, effectively defining $\vec{x} \wedge \vec{y}$ as an antisymmetric bilinear form, yielding elements of some linear space (the bivector space), this space having a specific associated length $|\vec{x} \wedge \vec{y}|$ giving the unsigned area of the parallelogram spanned by the respective $\vec{x}$, $\vec{y}$. The exterior product formula thereby simply becomes

(basis expansion of an exterior product of two vectors in an arbitrary dimension count)

$$\vec{x} \wedge \vec{y} = x^i y^j \cdot (\vec{e}_i \wedge \vec{e}_j) \tag{5.44}$$

with the additional restriction that the bilinear form (5.44) is antisymmetric: $\vec{x} \wedge \vec{y} = -\vec{y} \wedge \vec{x}$.

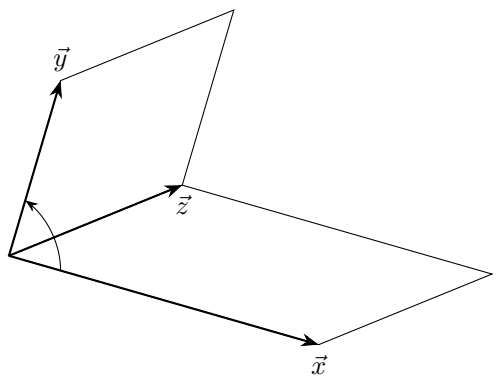


Figure 5.15: The angle between the planes $\vec{x} \wedge \vec{z}$ and $\vec{y} \wedge \vec{z}$, where $\vec{x} \perp \vec{z}$ and $\vec{y} \perp \vec{z}$, is equal to the angle between the vectors $\vec{x}$ and $\vec{y}$.

▷ E.g. in 4D we obtain in place of (5.28b)

$$\vec{x} \wedge \vec{y} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \cdot \begin{bmatrix} 0 & e_{12} & e_{13} & e_{14} \\ -e_{12} & 0 & e_{23} & e_{24} \\ -e_{13} & -e_{23} & 0 & e_{34} \\ -e_{14} & -e_{24} & -e_{34} & 0 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

where $e_{ij} = \vec{e}_i \wedge \vec{e}_j$. We thus have 6 different basis bivectors, the dimension of the 4D bivector space is thereby 6.

Respectively, the bivector length emerges by postulating the orthonormality of the basis bivectors $e_{ij} = \vec{e}_i \wedge \vec{e}_j$ obtained from orthonormal basis vectors e_i as in (5.30) and (5.31). This orthonormality then implies the "Pythagorean theorem formula" (5.33) generalized to the necessary dimension count. By the same orthonormality we also obtain the dot product formula (5.35) generalized to the necessary dimension count. The orthonormality also effectively implies (5.37) and thereby (5.42) (both formulas being directly applicable in any dimension count).

▷ The general approach of postulating the orthonormality of basis bivectors (in an orthonormal vector basis) can be traced back to two aspects:

- the unit length of basis bivectors effectively just specifies the units of area
- the orthogonality of basis bivectors is a particular case of the desired condition that the angle between bivectors $\vec{x} \wedge \vec{z}$ and $\vec{y} \wedge \vec{z}$, where $\vec{x}$ and $\vec{y}$ are orthogonal to $\vec{z}$, must be equal to the angle between vectors $\vec{x}$ and $\vec{y}$ (which is what we effectively wanted in (5.27a) and (5.27b)). Fig. 5.15 illustrates.

2-blades

The 4D and higher dimension-count cases have an important distinction from 3D (or 2D): we cannot guarantee that an arbitrary bivector ζ is representable as an exterior product of two vectors. The simplest argument is brought up by considering the sum $\vec{e}_1 \wedge \vec{e}_2 + \vec{e}_3 \wedge \vec{e}_4$. The plane spanned by $\vec{e}_1$, $\vec{e}_2$ and the plane spanned by $\vec{e}_3$, $\vec{e}_4$ intersect at a single point, which is the space's origin. Thereby

they do not have an intersection line and respectively there is no common vector factor between the two terms of the sum. The sum $\vec{e}_1 \wedge \vec{e}_2 + \vec{e}_3 \wedge \vec{e}_4$ thus cannot be reduced to a single exterior product of two vectors.

Thereby the bivector space elements fall into 2 different classes: those which can be represented as an exterior product of two vectors, and those which can not. The former kind of bivectors is more specifically referred to as *2-blades* (or, alternatively, *decomposable bivectors* or *simple bivectors*).

▷ This is raising an interesting question of what the length of the non-decomposable bivector $\vec{e}_1 \wedge \vec{e}_2 + \vec{e}_3 \wedge \vec{e}_4$ should be. There is also an equivalent question of what the dot product of $e_{12} = \vec{e}_1 \wedge \vec{e}_2$ and $e_{34} = \vec{e}_3 \wedge \vec{e}_4$ (for an orthonormal vector basis) should be. We mentioned that the mutual orthogonality of e.g. the basis vectors e_{12} and e_{23} is motivated by the desire to have the angle between the e_{12} and e_{23} planes (which share a common vector $\vec{e}_2$) equal to the angle between the vectors $\vec{e}_1$ and $\vec{e}_3$. However the planes e_{12} and e_{34} do not have a common vector. What should be the angle between e_{12} and e_{34}?

An obvious answer would be to always let all basis bivectors be orthonormal, given an orthonormal vector basis. This however would result in

$$|\vec{e}_1 \wedge \vec{e}_2 + \vec{e}_3 \wedge \vec{e}_4| = \sqrt{|\vec{e}_1 \wedge \vec{e}_2|^2 + |\vec{e}_3 \wedge \vec{e}_4|^2} = \sqrt{2}$$

while intuitively we might have expected that the area associated with the bivector $\vec{e}_1 \wedge \vec{e}_2 + \vec{e}_3 \wedge \vec{e}_4$ is equal to 2, since it "falls into two independent, unconnected unit areas", suggesting that maybe we should have postulated $e_{12} \cdot e_{34} = 1$. The latter idea, however, doesn't convincingly work, since in that case we'd also have e.g. $|\vec{e}_1 \wedge \vec{e}_2 - \vec{e}_3 \wedge \vec{e}_4| = 0$, while intuitively we would have expected to obtain the value of 2 again. It seems therefore consistent to simply let all basis bivectors be orthonormal, so, in particular, $e_{12} \cdot e_{34} = 0$.

Another argument in favor of $e_{12} \cdot e_{34} = 0$ would be to notice that, while there is no shared vector between e_{12} and e_{34}, all vectors in these two planes are mutually orthogonal.

It follows that the formula (5.42) cannot be used to obtain dot products of non-decomposable bivectors. But we can still expand these non-decomposable bivectors into a sum of decomposable bivectors (basis expansion being one trivial way to do that) and apply (5.42) to the terms of that expansion. And we always can simply use (5.35) generalized to the necessary dimension count.

5.6 k-vectors

5.6.1 Associative wedge product in 3D

We are now going to further generalize the ideas introduced in Section 5.5. As the first step we shall allow the elements of the 3D bivector space to be further wedge-multiplied by vectors of the original 3D space.

The base of a unit cube is a unit square. This suggests that we could define the wedge product $(\vec{e}_1 \wedge \vec{e}_2) \wedge \vec{e}_3$ to be equal to $V(\vec{e}_1, \vec{e}_2, \vec{e}_3) = 1$:

$$(\vec{e}_1 \wedge \vec{e}_2) \wedge \vec{e}_3 = V(\vec{e}_1, \vec{e}_2, \vec{e}_3) = 1$$

where $\vec{e}_1 \wedge \vec{e}_2$ is a bivector corresponding to the base of the cube and $\vec{e}_3$ is a vector corresponding to the height of the cube. Obviously, we could apply the

same idea to other combinations of basis vectors, yielding:

$$(\vec{e}_i \wedge \vec{e}_j) \wedge \vec{e}_k \triangleq V(\vec{e}_i, \vec{e}_j, \vec{e}_k)$$

where we get ± 1 if i, j, k are all different and 0 otherwise. Or, dropping the $V(\vec{\cdot}, \vec{\cdot}, \vec{\cdot})$ notation in favor of pure wedge products,

$$(\vec{e}_i \wedge \vec{e}_j) \wedge \vec{e}_k = \vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k \tag{5.45a}$$

We can also allow wedge-multiplication of a vector by a bivector, where choosing $\vec{e}_j \wedge \vec{e}_k$ as the base of the cube and $\vec{e}_i$ as the height, we have

$$\vec{e}_i \wedge (\vec{e}_j \wedge \vec{e}_k) \triangleq V(\vec{e}_i, \vec{e}_j, \vec{e}_k)$$

or, in pure wedge product form,

$$\vec{e}_i \wedge (\vec{e}_j \wedge \vec{e}_k) = \vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k \tag{5.45b}$$

The idea is to combine (5.45a) and (5.45b) yielding an associative wedge operation:

(wedge product of a basis bivector and a basis vector)

$$(\vec{e}_i \wedge \vec{e}_j) \wedge \vec{e}_k = \vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k = \vec{e}_i \wedge (\vec{e}_j \wedge \vec{e}_k) \tag{5.46}$$

where the associativity can be intuitively understood as "permission to drop the parentheses".

By the bilinearity of wedge product of two vectors and by the trilinearity of triple wedge product, the definition (5.46) naturally generalizes to arbitrary vectors as

(wedge product of a bivector and a vector, associativity of wedge product of 3 vectors)

$$(\vec{x} \wedge \vec{y}) \wedge \vec{z} = \vec{x} \wedge \vec{y} \wedge \vec{z} = \vec{x} \wedge (\vec{y} \wedge \vec{z}) \tag{5.47}$$

It follows that the newly introduced wedge product between a vector and a bivector must be bilinear:

(bilinearity of wedge product of a bivector and a vector)

$$(\zeta + \eta) \wedge \vec{v} = \zeta \wedge \vec{v} + \eta \wedge \vec{v}$$
$$(\vec{x} + \vec{y}) \wedge \zeta = \vec{x} \wedge \zeta + \vec{y} \wedge \zeta$$

etc.

Contrary to some possible expectations, the wedge product of a bivector and a vector is not anticommutative, but commutative.

Goal 5.15. Show that $\zeta \wedge \vec{v} = \vec{v} \wedge \zeta$, where ζ is a bivector.

For decomposable $\zeta = \vec{x} \wedge \vec{y}$,

$$(\vec{x} \wedge \vec{y}) \wedge \vec{v} = \vec{x} \wedge \vec{y} \wedge \vec{v} = -(\vec{x} \wedge \vec{v} \wedge \vec{y}) = \vec{v} \wedge \vec{x} \wedge \vec{y} = \vec{v} \wedge (\vec{x} \wedge \vec{y})$$

For non-decomposable ζ the previous result can be generalized by bilinearity.

Goal 5.15 completed. ■

Earlier we have found component-form representations of the expressions $\vec{x} \wedge \vec{y}$ and $\vec{x} \wedge \vec{y} \wedge \vec{z}$. Now we shall find a component form representation of the wedge product of a bivector and a vector.

Goal 5.16. Find a component-form representation of $\zeta \wedge \vec{v}$, where ζ is a bivector. *Hint:* use bilinearity of $\zeta \wedge \vec{v}$.

Expanding ζ and $\vec{v}$ in basis terms,

$$\begin{aligned}
\zeta \wedge \vec{v} &= \Big(\zeta_{12} \cdot (\vec{e}_1 \wedge \vec{e}_2) + \zeta_{23} \cdot (\vec{e}_2 \wedge \vec{e}_3) + \zeta_{31} \cdot (\vec{e}_3 \wedge \vec{e}_1)\Big) \cdot (v_1\vec{e}_1 + v_2\vec{e}_2 + v_3\vec{e}_3) \\
&= \zeta_{12}v_1 \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_1) + \zeta_{23}v_1 \cdot (\vec{e}_2 \wedge \vec{e}_3 \wedge \vec{e}_1) + \zeta_{31}v_1 \cdot (\vec{e}_3 \wedge \vec{e}_1 \wedge \vec{e}_1) \\
&\quad + \zeta_{12}v_2 \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_2) + \zeta_{23}v_2 \cdot (\vec{e}_2 \wedge \vec{e}_3 \wedge \vec{e}_2) + \zeta_{31}v_2 \cdot (\vec{e}_3 \wedge \vec{e}_1 \wedge \vec{e}_2) \\
&\quad + \zeta_{12}v_3 \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) + \zeta_{23}v_3 \cdot (\vec{e}_2 \wedge \vec{e}_3 \wedge \vec{e}_3) + \zeta_{31}v_3 \cdot (\vec{e}_3 \wedge \vec{e}_1 \wedge \vec{e}_3) \\
&= (\zeta_{12}v_1 \cdot 0 + \zeta_{23}v_1 \cdot 1 + \zeta_{31}v_1 \cdot 0) + (\zeta_{12}v_2 \cdot 0 + \zeta_{23}v_2 \cdot 0 + \zeta_{31}v_2 \cdot 1) \\
&\quad + (\zeta_{12}v_3 \cdot 1 + \zeta_{23}v_3 \cdot 0 + \zeta_{31}v_3 \cdot 0) \\
&= \zeta_{23}v_1 + \zeta_{31}v_2 + \zeta_{12}v_3
\end{aligned}$$

and the product $\vec{v} \wedge \zeta$ can be obtained in a similar fashion, yielding the same expression (and reconfirming the commutativity of the product).

Goal 5.16 completed. ■

5.6.2 3-vector space

Similarly to how earlier we proposed to stop treating the wedge products of 2D vectors $\vec{x} \wedge \vec{y}$ as real numbers and treat them as elements of a 1-dimensional bivector space (with the default basis $\vec{e}_1 \wedge \vec{e}_2$) instead, we could start treating triple wedge products of 3D vectors $\vec{x} \wedge \vec{y} \wedge \vec{z}$ as elements of a 1-dimensional *3-vector* space with the default basis 3-vector $\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3$. Given an orthonormal 3D vector basis $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$, the respective basis 3 vector has unit length:

$$|\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3| = 1$$

Respectively, (5.25) turns into

(triple wedge product of vectors is actually a 3-vector, not a number)

$$\vec{x} \wedge \vec{y} \wedge \vec{z} = x^i y^j z^j \varepsilon_{ijk} \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) = \det \begin{bmatrix} x^1 & y^1 & z^1 \\ x^2 & y^2 & z^2 \\ x^3 & y^3 & z^3 \end{bmatrix} \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) \quad (5.48)$$

Differently from (5.25), the formula (5.48) holds in any basis, since it emerges simply by multilinearity and anticommutativity of the wedge product.

Using (5.48), we can also derive the counterpart of (5.42) for 3-blades:

(dot product of 3-blades in terms of the dot products of underlying vectors)

$$\zeta \cdot \eta = (\vec{x} \wedge \vec{y} \wedge \vec{z}) \cdot (\vec{u} \wedge \vec{v} \wedge \vec{w}) = \det \begin{bmatrix} \vec{x} \cdot \vec{u} & \vec{y} \cdot \vec{u} & \vec{z} \cdot \vec{u} \\ \vec{x} \cdot \vec{v} & \vec{y} \cdot \vec{v} & \vec{z} \cdot \vec{v} \\ \vec{x} \cdot \vec{w} & \vec{y} \cdot \vec{w} & \vec{z} \cdot \vec{w} \end{bmatrix} \tag{5.49}$$

▷ As with (5.42) in 2D, we can use (5.49) to check for 3-vector orientations in 3D. A positive dot product value $\zeta \cdot \eta > 0$ means that the orientations of 3-vectors ζ and η coincide.

5.6.3 k-vector space

We can already guess how this generalizes to higher dimension counts and longer chains of wedge products:

- Wedge product of k vectors is a function, linear in each of the k arguments and anticommutative in each pair of arguments[4], producing an element of a special linear space:

  $$\zeta = \vec{v}_1 \wedge \vec{v}_2 \wedge \ldots \wedge \vec{v}_k$$

  This element ζ is referred to as *k-blade* (or as decomposable or simple k-vector). The general name for the elements of this linear space (including those which cannot be decomposed into wedge products of k vectors) is *k-vectors*. The linear space itself is referred to as *k-vector space*.

- Wedge product of a k-vector with an l-vector is defined by postulating the associativity of wedge product, which effectively allows to "drop the parentheses", thereby yielding a $(k+l)$-vector:

  $$\begin{aligned}&(\vec{u}_1 \wedge \vec{u}_2 \wedge \ldots \wedge \vec{u}_k) \wedge (\vec{v}_1 \wedge \vec{v}_2 \wedge \ldots \wedge \vec{v}_l) \\ &\quad = \vec{u}_1 \wedge \vec{u}_2 \wedge \ldots \wedge \vec{u}_k \wedge \vec{v}_1 \wedge \vec{v}_2 \wedge \ldots \wedge \vec{v}_l\end{aligned}$$

- The default k-vector basis is obtained from wedge products of k different basis vectors, where each combination of basis vectors participates only once (since one and the same subset of basis vectors, wedge-multiplied in different orders, yields "cohyperplanar" k-vectors). If the vector basis is orthonormal, so is the obtained k-vector basis.

Goal 5.17. Assuming the vector space has N dimensions, find the dimension count of the k-vector space. *Hint:* find the number of linearly independent k-vectors obtained as wedge products of basis vectors.

Given N basis vectors, we can construct C_N^k different subsets of those (where C_N^k is the binomial coefficient). Therefore the dimension of the k-vector space of an N-dimensional vector space is C_N^k. If $k = N$, this dimension is thereby 1.

[4] It would suffice to require anticommutativity in each pair of neighboring arguments.

▷ If $k > N$, the k-vector space contains only the zero element. E.g. in 2D any wedge product of 3 vectors is zero, since 3 vectors are always linearly dependent in 2D. The simplest example is provided by

$$\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_1 = -(\vec{e}_1 \wedge \vec{e}_1 \wedge \vec{e}_2) = -((\vec{e}_1 \wedge \vec{e}_1) \wedge \vec{e}_2) = -(0 \wedge \vec{e}_2) = 0$$

Goal 5.17 completed. ■

Goal 5.18. Generalize (5.48) to k-vectors in k-dimensional space.

By an argument, similar to 2- and 3-vectors,

(k-blade in terms of components of the underlying vectors in a k-dimensional space)

$$\vec{x}_k \wedge \ldots \wedge \vec{x}_k = x_1^{i_1} \ldots x_k^{i_k} \varepsilon_{i_1 \ldots i_k} \cdot (\vec{e}_1 \wedge \ldots \wedge \vec{e}_3) = \det \begin{bmatrix} x_1^1 & \cdots & x_k^1 \\ \vdots & \ddots & \vdots \\ x_1^k & \cdots & x_k^k \end{bmatrix} \cdot (\vec{e}_1 \wedge \ldots \wedge \vec{e}_k) \tag{5.50}$$

Goal 5.18 completed. ■

Goal 5.19. Generalize (5.42) and (5.49) to a dot product of two k-blades.

By an argument, similar to the derivation of (5.42) and (5.49),

(dot product of k-blades in terms of the dot products of underlying vectors)

$$\zeta \cdot \eta = (\vec{x}_1 \wedge \ldots \wedge \vec{x}_k) \cdot (\vec{y}_1 \wedge \ldots \wedge \vec{y}_k) = \det \begin{bmatrix} \vec{x}_1 \cdot \vec{y}_1 & \cdots & \vec{x}_k \cdot \vec{y}_1 \\ \vdots & \ddots & \vdots \\ \vec{x}_1 \cdot \vec{y}_k & \cdots & \vec{x}_k \cdot \vec{y}_k \end{bmatrix} \tag{5.51}$$

▷ As with (5.42) in 2D and (5.49) in 3D, we can use (5.51) to check for k-vector orientations in a k-dimensional vector space. A positive dot product value $\zeta \cdot \eta > 0$ means that the orientations of k-vectors ζ and η coincide.

Goal 5.19 completed. ■

5.7 Volume deformations

▷ For the sake of simpler readability in the following discussion we are going to speak of 3D volumes and 3-vectors, however the material generally assumes k-vectors in k-dimensional spaces and the associated k-dimensional hypervolumes. The readers are expected to apply the respective generalizations themselves.

In the following discussion we'll need to make a clear distinction between 3-vectors emerging from a 3D vector space and the signed volumes implied by these 3-vectors. Since we need signed volumes, using the absolute value operation on a 3-vector to obtain the respective volume won't do, and we need something else. We will introduce the "$\star$" operation, which converts from a 3-vector in a 3D space to the associated volume and back (or, more generally from a k-vector in a k-dimensional space to the associated hypervolume and back). So, given the 3-vector $\zeta = \vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3$, where $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$ is a positively oriented orthonormal basis, we have:

$$\begin{aligned} \star\zeta &= 1 && \text{(the signed volume associated with } \zeta \text{ is } +1) \\ \star 1 &= \zeta && \text{(the 3-vector representing the unit volume is } \zeta) \end{aligned}$$

▷ We just introduced the "$\star$" operation in a preliminary, limited fashion, just for the needs of this section. A proper introduction will be done in Section 5.8.

5.7.1 Covariant Levi–Civita tensor

In the derivation of (5.25) we have equated the triple wedge products $\vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k$ to scalar values ε_{ijk}, which we by now agreed not to do any longer and consider these expressions to yield 3-vectors rather than scalars. We can however now use the "$\star$" operation to convert 3D 3-vectors into scalars. We could therefore have written

$$\varepsilon_{ijk} \stackrel{?}{=} \star(\vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k)$$

except that, by convention, the components of the Levi–Civita symbol ε_{ijk} are supposed to take one of the $+1$, 0 and -1 values, therefore such equality would hold only in orthonormal bases, or actually only in positively oriented orthonormal bases (since the components of ε_{ijk} are supposed to take specific values, particularly ε_{123} is supposed to be equal to 1).

To avoid breaking this convention and simultaneously allow using any kind of basis, we introduce a new entity E_{ijk}:

(covariant Levi–Civita tensor in 3D)

$$E_{ijk} = \star(\vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k) = V(\vec{e}_i, \vec{e}_j, \vec{e}_k) \tag{5.52}$$

We thereby can rewrite the trilinear expansion of the triple wedge product in (5.25b) as

(3D volume of a parallelepiped spanned by vectors $\vec{x}$, $\vec{y}$, $\vec{z}$)

$$V(\vec{x}, \vec{y}, \vec{z}) = \star(\vec{x} \wedge \vec{y} \wedge \vec{z}) = \star(x^i \vec{e}_i \wedge y^j \vec{e}_j \wedge z^k \vec{e}_k)$$

$$= x^i y^j z^k \cdot \star(\vec{e}_i \wedge \vec{e}_j \wedge \vec{e}_k) \overset{(5.52)}{=} x^i y^j z^k E_{ijk} \tag{5.53}$$

where we used the obvious fact $\star(\lambda\zeta) = \lambda \cdot \star\zeta$ for any scalar λ and any 3-vector ζ, simply meaning that multiplication of a 3-vector ζ by a scalar λ increases the associated volume λ times.

Differently from (5.25), the equation (5.53) holds in any, not just positively oriented orthonormal, basis. Thereby (5.53) is invariant under basis changes, that is (5.53) is in a covariant form. Thereby E_{ijk} must be a true $(0,3)$-type tensor. It is referred to as (the 3D version of) *covariant Levi–Civita tensor*.

Goal 5.20. Find the relationship between covariant Levi–Civita tensor and Levi–Civita symbol.

By (5.48)

$$\begin{aligned} V(\vec{x}, \vec{y}, \vec{z}) &\overset{(5.48)}{=} \star(\vec{x} \wedge \vec{y} \wedge \vec{z}) = \star\big(x^i y^j z^j \varepsilon_{ijk} \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3)\big) \\ &= x^i y^j z^j \varepsilon_{ijk} \cdot \star(\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) \end{aligned}$$

Comparing to (5.53) yields

(relationship between covariant Levi–Civita tensor and Levi–Civita symbol)

$$E_{ijk} = \varepsilon_{ijk} \cdot \star(\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) = \varepsilon_{ijk} \cdot V(\vec{e}_1, \vec{e}_2, \vec{e}_3) \tag{5.54}$$

▷ It follows that ε_{ijk} is not a tensor, since the components of ε_{ijk} are fixed by convention, while, according to (5.54), the components of a tensor entity of this kind would be scaled proportionally to the basis volume $V(\vec{e}_1, \vec{e}_2, \vec{e}_3)$.

Goal 5.20 completed. ■

Goal 5.21. Find the relationship connecting the covariant Levi–Civita tensor to the metric tensor.

The components of the metric tensor are given by the Gram matrix, consisting of the dot products of basis vectors. This looks like the right-hand side of (5.49) (and generally (5.51)) if we let

$$\zeta = \eta = \vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3$$

in which case (5.49) yields

$$\zeta \cdot \zeta = (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) = \det \begin{bmatrix} \vec{e}_1 \cdot \vec{e}_1 & \vec{e}_2 \cdot \vec{e}_1 & \vec{e}_3 \cdot \vec{e}_1 \\ \vec{e}_1 \cdot \vec{e}_2 & \vec{e}_2 \cdot \vec{e}_2 & \vec{e}_3 \cdot \vec{e}_2 \\ \vec{e}_1 \cdot \vec{e}_3 & \vec{e}_2 \cdot \vec{e}_3 & \vec{e}_3 \cdot \vec{e}_3 \end{bmatrix} = \det \mathbf{G}$$

The components of the Gram matrix $\mathbf{G}$ are the same as the components of the metric tensor g_{ij} which motivates the alternative notations $\det g$ or $|g|$ instead of $\det \mathbf{G}$. We thereby obtain $|\zeta|^2 = |g|$ and $|\zeta| = \sqrt{|g|}$.

It follows that

$$V(\vec{e}_1, \vec{e}_2, \vec{e}_3) = \star(\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) = \star\zeta = \pm|\zeta| = \pm\sqrt{|g|}$$

or, dropping the helper entity ζ,

(the unsigned basis volume or hypervolume is equal to the squared root of the determinant of the metric tensor)

$$V(\vec{e}_1, \vec{e}_2, \vec{e}_3) = \star(\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) = \pm\sqrt{|g|} \tag{5.55}$$

where, apparently the plus sign is to be taken for a positively-oriented basis and the minus sign for a negatively-oriented basis. The dot products of basis vectors are the same in both orientations of the basis, thereby they cannot contain information about basis orientation, and this information needs to be supplied additionally in the form of the "±" sign in (5.55).

> ▷ This is similar to the formula $\sin\alpha = \pm\sqrt{1-\cos^2\alpha}$, where the value of the cosine is sufficient to determine the absolute value of the sine, but the information about the sign of the sine is lost.

Subsituting (5.55) into (5.54), we obtain

(covariant Levi–Civita tensor expressed via the determinant of the metric tensor, 3D version)

$$E_{ijk} = \pm\sqrt{|g|} \cdot \varepsilon_{ijk} \tag{5.56}$$

Goal 5.21 completed. ■

Substituting (5.54) into (5.53) we obtain

$$V(\vec{x}, \vec{y}, \vec{z}) = x^i y^j z^j \varepsilon_{ijk} \cdot V(\vec{e}_1, \vec{e}_2, \vec{e}_3) = \det \begin{bmatrix} x^1 & y^1 & z^1 \\ x^2 & y^2 & z^2 \\ x^3 & y^3 & z^3 \end{bmatrix} \cdot V(\vec{e}_1, \vec{e}_2, \vec{e}_3) \tag{5.57}$$

(which is effectively simply a scalar version of (5.48) and alternatively could have been obtained from the latter simply by applying the "⋆" operation to both sides). It follows that the determinant of the matrix

$$\mathbf{A} = \begin{bmatrix} x^1 & y^1 & z^1 \\ x^2 & y^2 & z^2 \\ x^3 & y^3 & z^3 \end{bmatrix}$$

is equal to the ratio between the volumes $V(\vec{x}, \vec{y}, \vec{z})$ and $V(\vec{e}_1, \vec{e}_2, \vec{e}_3)$.

We can consider the matrix $\mathbf{A}$ as the transformation matrix from the basis $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$ to the set of vectors $\vec{x}$, $\vec{y}$, $\vec{z}$:

$$\begin{bmatrix} \vec{x} & \vec{y} & \vec{z} \end{bmatrix} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \begin{bmatrix} x^1 & y^1 & z^1 \\ x^2 & y^2 & z^2 \\ x^3 & y^3 & z^3 \end{bmatrix} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \mathbf{A}$$

In fact, we can consider $\vec{x}$, $\vec{y}$, $\vec{z}$ as a new basis and respectively $\mathbf{A}$ as the basis transformation matrix. To notationally reflect that, we will rename $\vec{x}$, $\vec{y}$, $\vec{z}$ to $\vec{e}_1{}'$, $\vec{e}_2{}'$, $\vec{e}_3{}'$:

$$\begin{bmatrix} \vec{e}_1{}' & \vec{e}_2{}' & \vec{e}_3{}' \end{bmatrix} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \mathbf{A} \tag{5.58}$$

thus (5.57) turns into

$$V(\vec{e}_1{}', \vec{e}_2{}', \vec{e}_3{}') = \det \mathbf{A} \cdot V(\vec{e}_1, \vec{e}_2, \vec{e}_3) \tag{5.59}$$

The value of the matrix determinant $\det \mathbf{A}$ is thereby equal to the *deformation factor* of the basis volume $V(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ by the basis transformation (5.58).

Deforming the basis volume by the factor $\det \mathbf{A}$ will deform all other volumes in the space by the same factor, since by (5.57) any other volume in the space is proportional to the basis volume $V(\vec{e}_1, \vec{e}_2, \vec{e}_3)$. Thereby $\det \mathbf{A}$ describes the deformation of all volumes in the space occuring under the basis transformation (5.58).

> ▷ The basis transformation formula (5.58) is formally indistinguishable from the basis change formula, however in the current context is has a very different semantics. During a basis change we only change the basis, while we aim at preserving all vectors and other covariant entities in the linear space, thereby we have to adjust the components of those entities to compensate for the basis change. In our current context however, we do not change the values of those components at all, since we do want all entities to transform along with the basis.

Intuitively we would expect that the basis transformation (5.58) (where the components of all vectors in the space are unchanged) can be reexpressed in terms of a linear space transformation, where we do not change the basis but change the vectors (and thereby the vector components). If we are successful with that, we might be able to associate matrix determinants with volume deformations caused by linear transformations of the space.

Goal 5.22. Try to equivalently reexpress the basis transformation (5.58) as a linear transformation of the vector space.

Consider a random vector in our vector space:

$$\vec{v} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \mathbf{v}$$

where

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

After the basis transformation by (5.58) this vector will turn into

$$\vec{v}\,' = \left(\begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \mathbf{A} \right) \cdot \mathbf{v} = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot (\mathbf{A} \cdot \mathbf{v}) = \begin{bmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \end{bmatrix} \cdot \mathbf{v}'$$

where

$$\mathbf{v}' = \mathbf{A} \cdot \mathbf{v} \tag{5.60}$$

Thus, we have equivalently reexpressed the basis transformation (5.58) as a vector space transformation (5.60), where we transform the vector components by the same matrix $\mathbf{A}$. It follows that

(the determinant of a matrix is equal to the volume deformation factor of the linear transformation described by that matrix)

$$V\left(\vec{A}(\vec{v}_1), \vec{A}(\vec{v}_2), \vec{A}(\vec{v}_3)\right) = \det \mathbf{A} \cdot V(\vec{v}_1, \vec{v}_2, \vec{v}_3) \tag{5.61a}$$

or, in terms of wedge products,

$$\star\left(\vec{A}(\vec{v}_1)\wedge\vec{A}(\vec{v}_2)\wedge\vec{A}(\vec{v}_3)\right)=\det\mathbf{A}\cdot\star(\vec{v}_1\wedge\vec{v}_2\wedge\vec{v}_3) \tag{5.61b}$$

where $\mathbf{A}$ is the matrix of the components of the linear transformation operator $\vec{A}(\cdot)$.

▷ Formulas (5.61) immediately suggest that $\det\mathbf{A}$ is invariant under basis changes, since the rest of the formulas (5.61) *is* invariant. We will address this question once again in (5.66).

Goal 5.22 completed. ■

5.7.2 Matrix determinants reviewed

Goal 5.23. Review the definition and the main properties of matrix determinants from the perspective of the material discussed in this chapter and, in particular, in this section.

Until now we have been implicitly assuming that the determinant of a matrix is defined as a particular alternating sum of matrix components, e.g. for a 3×3 matrix

(algebraic definition of matrix determinant)

$$\begin{aligned}\det\mathbf{A}&=\det\begin{bmatrix}a_1^1 & a_2^1 & a_3^1\\ a_1^2 & a_2^2 & a_3^2\\ a_1^3 & a_2^3 & a_3^3\end{bmatrix}\\ &\triangleq a_1^1a_2^2a_3^3-a_1^1a_3^2a_2^3-a_2^1a_1^2a_3^3+a_2^1a_3^2a_1^3+a_3^1a_1^2a_2^3-a_3^1a_2^2a_1^3\end{aligned} \tag{5.62a}$$

where, instead of shuffling the lower indices, we can shuffle the upper indices, obtaining an equal to the previous one result:

$$\det\mathbf{A}\triangleq a_1^1a_2^2a_3^3-a_1^1a_2^3a_2^3-a_1^2a_2^1a_3^3+a_1^2a_2^3a_3^1+a_1^3a_2^1a_3^2-a_1^3a_2^2a_3^1 \tag{5.62b}$$

From the perspective of the material of this chapter, however, we can create a determinant definition based on volume interpretation.

Goal 5.23a. Offer an alternative definition of matrix determinant using the volume-based interpretation.

We can use the formula (5.57) in a positively oriented orthonormal basis, where $V(\vec{e}_1,\vec{e}_2,\vec{e}_3)=1$ and therefore (5.57) turns into (5.26). Thus, given a matrix $\mathbf{A}$, let's interpret the columns of the matrix as coordinates of some vectors $\vec{a}_i$ in a positively oriented orthonormal basis. Then we can define $\det\mathbf{A}$ as the signed value of the volume (or hypervolume) spanned by those vectors:

(definition of matrix determinant as volume)

$$\det\mathbf{A}=\det\begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3\end{bmatrix}\triangleq V(\vec{a}_1,\vec{a}_2,\vec{a}_3)=\star(\vec{a}_1\wedge\vec{a}_2\wedge\vec{a}_3) \tag{5.63}$$

where

$$\mathbf{a}_i = \begin{bmatrix} a_i^1 \\ a_i^2 \\ a_i^3 \end{bmatrix} = \begin{bmatrix} \vec{a}_i \cdot \vec{e}_1 \\ \vec{a}_i \cdot \vec{e}_2 \\ \vec{a}_i \cdot \vec{e}_3 \end{bmatrix}$$

is a matrix column of coordinates of vector $\vec{a}_i$ in the implied orthonormal basis $\{\vec{e}_j\}$.

Goal 5.23a completed. ■

Goal 5.23b. Obtain the formula for the determinant of a matrix product. *Hint:* consider two consecutive linear transformations of the vector space.

Let $\mathbf{B}$ be a linear transformation of the vector space, its associated volume deformation factor being $\det \mathbf{B}$. Let's apply another linear transformation $\mathbf{A}$ on top of $\mathbf{B}$. This will perform another volume deformation of factor $\det \mathbf{A}$. The total volume deformation factor is thereby $\det \mathbf{A} \cdot \det \mathbf{B}$. On the other hand, we can consider a sequence of two linear transformations $\mathbf{B}$, $\mathbf{A}$ as a single linear transformation by the matrix $\mathbf{AB}$, its respective volume deformation factor being $\det(\mathbf{AB})$. The transformation result is the same in either case and so must be the volume deformation factors:

(determinant of a matrix product is equal to the product of determinants)

$$\det(\mathbf{AB}) = \det \mathbf{A} \cdot \det \mathbf{B} \tag{5.64}$$

Goal 5.23b completed. ■

Goal 5.23c. Obtain the formula for the determinant of a matrix inverse.

By the definition of the matrix inverse $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}$. Now $\mathbf{I}$ defines an identity linear transformation of the vector space, its volume deformation factor being equal to 1. It follows by (5.64) that

$$1 = \det \mathbf{I} = \det(\mathbf{A} \cdot \mathbf{A}^{-1}) = \det \mathbf{A} \cdot \det(\mathbf{A}^{-1})$$

and therefore

(determinant of a matrix inverse is reciprocal to the determinant of the original matrix)

$$\det(\mathbf{A}^{-1}) = (\det \mathbf{A})^{-1} \tag{5.65}$$

Goal 5.23c completed. ■

Goal 5.23d. Find how a matrix determinant transforms under a similarity transformation.

By (5.65)

$$\det(\mathbf{T}^{-1} \cdot \mathbf{A} \cdot \mathbf{T}) = \det(\mathbf{T}^{-1}) \cdot \det \mathbf{A} \cdot \det \mathbf{T} = \big(\det(\mathbf{T}^{-1}) \cdot \det \mathbf{T}\big) \cdot \det \mathbf{A}$$
$$\overset{(5.65)}{=} \det \mathbf{A}$$

thus

(determinant of a matrix is invariant under similarity transformations)

$$\det(\mathbf{T}^{-1} \cdot \mathbf{A} \cdot \mathbf{T}) = \det \mathbf{A} \tag{5.66}$$

Goal 5.23d completed. ■

Goal 5.23e. Obtain the formula for the determinant of a matrix transpose in two different ways: using the algebraic definition of the determinant and using the volume-based interpretation. *Hint:* for the volume-based approach use (5.49).

The difference between the two versions (5.62a) and (5.62b) of the algebraic definition of matrix determinant is swapped indices of matrix components. But this is the same as the difference between the elements of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ and it follows that $\det \mathbf{A} = \det \mathbf{A}^{\mathrm{T}}$.

Alternatively we can notice that the matrix in the right-hand side of (5.49) (which is emerging from (5.48), which is effectively the same as the volume-based definition of matrix determinant (5.63)) can represent a product of a matrix and its transpose. Indeed, let

$$\mathbf{A} = \begin{bmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ a_1^3 & a_2^3 & a_3^3 \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \end{bmatrix}$$

where

$$\mathbf{a}_i = \begin{bmatrix} a_i^1 \\ a_i^2 \\ a_i^3 \end{bmatrix}$$

Therefore

$$\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \\ \mathbf{a}_3^{\mathrm{T}} \end{bmatrix}$$

and

$$\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A} = \begin{bmatrix} \mathbf{a}_1^{\mathrm{T}} \cdot \mathbf{a}_1 & \mathbf{a}_1^{\mathrm{T}} \cdot \mathbf{a}_2 & \mathbf{a}_1^{\mathrm{T}} \cdot \mathbf{a}_3 \\ \mathbf{a}_2^{\mathrm{T}} \cdot \mathbf{a}_1 & \mathbf{a}_2^{\mathrm{T}} \cdot \mathbf{a}_2 & \mathbf{a}_2^{\mathrm{T}} \cdot \mathbf{a}_3 \\ \mathbf{a}_3^{\mathrm{T}} \cdot \mathbf{a}_1 & \mathbf{a}_3^{\mathrm{T}} \cdot \mathbf{a}_2 & \mathbf{a}_3^{\mathrm{T}} \cdot \mathbf{a}_3 \end{bmatrix}$$

In an orthonormal basis the matrix $\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A}$ is simply a matrix of dot products $\vec{a}_i \cdot \vec{a}_j$, where $\vec{a}_i$ are vectors with coordinates given by the components of the column matrices $\mathbf{a}_i$.

Therefore letting

$$\zeta = \eta = \vec{a}_1 \wedge \vec{a}_2 \wedge \vec{a}_3$$

we obtain exactly the matrix $\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A}$ in the right-hand side of (5.49), while in the left-hand side we obtain the squared volume

$$\zeta \cdot \zeta = |\zeta|^2 = V^2(\vec{a}_1, \vec{a}_2, \vec{a}_3) = (\det \mathbf{A})^2$$

Therefore (5.49) reads

$$(\det \mathbf{A})^2 = \det(\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A}) = \det(\mathbf{A}^{\mathrm{T}}) \cdot \det \mathbf{A} \tag{5.67}$$

and it follows that

(determinant of a matrix transpose is equal to the determinant of the original matrix)

$$\det(\mathbf{A}^{\mathrm{T}}) = \det \mathbf{A} \tag{5.68}$$

▷ Strictly speaking, in case of $\det \mathbf{A} = 0$ we cannot immediately infer $\det(\mathbf{A}^{\mathrm{T}}) = 0$ from (5.67), since (5.67) would also be satisfied by any other value of $\det(\mathbf{A}^{\mathrm{T}})$. However with a bit of further thinking we can realize that a nonzero value $\det(\mathbf{A}^{\mathrm{T}})$ won't do in this case.

Indeed, suppose $\det(\mathbf{A}^{\mathrm{T}}) \neq 0$. Noticing that $(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}$, we can apply the previous argument to $\mathbf{A}^{\mathrm{T}}$ instead of $\mathbf{A}$, arriving at the equation

$$\Big(\det(\mathbf{A}^{\mathrm{T}})\Big)^2 = \det(\mathbf{A} \cdot \mathbf{A}^{\mathrm{T}}) = \det \mathbf{A} \cdot \det(\mathbf{A}^{\mathrm{T}})$$

instead of (5.67), and it follows that $\det \mathbf{A} = \det(\mathbf{A}^{\mathrm{T}}) \neq 0$, thereby excluding the case of simultnaneously having $\det \mathbf{A} = 0$, $\det(\mathbf{A}^{\mathrm{T}}) \neq 0$.

Goal 5.23e completed. ■

Goal 5.23f. Show that a matrix determinant is invariant under "shear transformations" of a matrix, that is if we take a matrix row (or a column), multiply it by some scalar value and add it to another row (or respectively column). *Hint:* use (5.63).

Recalling that wedge products are invariant under shear transformations, we are going to use (5.63), which interprets a matrix determinant in terms of wedge product. For the sake of definiteness let's take the first column of a matrix $\mathbf{A}$, multiply it by scalar λ and add to the second column, obtaining a modified matrix $\mathbf{A}'$. By (5.63),

$$\begin{aligned}
\det \mathbf{A}' &= \star\big(\vec{a}_1 \wedge (\vec{a}_2 + \lambda \vec{a}_1) \wedge \vec{a}_3\big) \\
&= \star(\vec{a}_1 \wedge \vec{a}_2 \wedge \vec{a}_3 + \lambda \cdot \vec{a}_1 \wedge \vec{a}_1 \wedge \vec{a}_3) \\
&= \star(\vec{a}_1 \wedge \vec{a}_2 \wedge \vec{a}_3 + 0) \\
&= \star(\vec{a}_1 \wedge \vec{a}_2 \wedge \vec{a}_3) = \det \mathbf{A}
\end{aligned}$$

The same property for matrix rows can be inferred by additionally using (5.68).

Goal 5.23f completed. ■

Goal 5.23 completed. ■

5.8 Pseudotensors revisited: Hodge star

We now shall drop the previously introduced operation "$\star$", which we used to convert k-vectors to (hyper)volume scalar values and back, and reintroduce it again properly, in particular making it applicable to any k-vector associated with a vector space of any dimension.

5.8.1 Motivation and definition

In the discussion of 3D bivector space, we mentioned that the formula (5.29) looks very similar to the cross product formula. This suggests that, in a positively oriented orthonormal vector basis, the basis bivectors $\vec{e}_1 \wedge \vec{e}_2$, $\vec{e}_2 \wedge \vec{e}_3$,

$\vec{e}_3 \wedge \vec{e}_1$ of the 3D bivector space might coincide with the ordinary 3D basis as

$$\begin{aligned} \vec{e}_1 \wedge \vec{e}_2 \quad &\longleftrightarrow \quad \vec{e}_3 \\ \vec{e}_2 \wedge \vec{e}_3 \quad &\longleftrightarrow \quad \vec{e}_2 \\ \vec{e}_3 \wedge \vec{e}_1 \quad &\longleftrightarrow \quad \vec{e}_1 \end{aligned} \tag{5.69}$$

We also suggested that these two spaces and these two bases shouldn't be treated as identical ones. A 1:1 mapping between those should be possible nevertheless, or so it would seem, at least looking at (5.69). On each side we have a 3-dimensional space and we should be able to map them to each other. In this mapping linear combinations of basis bivectors from the left of (5.69) would be mapped to linear combinations of basis vectors from the right of (5.69) or vice versa, the linear combination coefficients being identical:

$$\zeta^{12} \cdot (\vec{e}_1 \wedge \vec{e}_2) + \zeta^{23} \cdot (\vec{e}_2 \wedge \vec{e}_3) + \zeta^{31} \cdot (\vec{e}_3 \wedge \vec{e}_1) \quad \longleftrightarrow \quad \zeta^{12}\vec{e}_3 + \zeta^{23}\vec{e}_1 + \zeta^{31}\vec{e}_2 \tag{5.70}$$

We can also notice that (5.69) maps basis bivectors to basis vectors that are normal to the bivector's plane. This property is preserved in their linear combinations in (5.70), which maps the bivector

$$\zeta = \zeta^{12} \cdot (\vec{e}_1 \wedge \vec{e}_2) + \zeta^{23} \cdot (\vec{e}_2 \wedge \vec{e}_3) + \zeta^{31} \cdot (\vec{e}_3 \wedge \vec{e}_1) \tag{5.71a}$$

to the vector

$$\vec{z} = \zeta^{12}\vec{e}_3 + \zeta^{23}\vec{e}_1 + \zeta^{31}\vec{e}_2 \tag{5.71b}$$

Indeed, the mutual orthogonality of ζ and $\vec{z}$ can be seen from

$$\begin{aligned} \zeta \wedge \vec{z} &= \big(\zeta^{12} \cdot (\vec{e}_1 \wedge \vec{e}_2) + \zeta^{23} \cdot (\vec{e}_2 \wedge \vec{e}_3) + \zeta^{31} \cdot (\vec{e}_3 \wedge \vec{e}_1)\big) \\ &\qquad \wedge \big(\zeta^{12}\vec{e}_3 + \zeta^{23}\vec{e}_1 + \zeta^{31}\vec{e}_2\big) \\ &= (\zeta^{12})^2 \cdot (\vec{e}_1 \wedge \vec{e}_2) \wedge \vec{e}_3 + (\zeta^{23})^2 \cdot (\vec{e}_2 \wedge \vec{e}_3) \wedge \vec{e}_1 + (\zeta^{31})^2 \cdot (\vec{e}_3 \wedge \vec{e}_1) \wedge \vec{e}_2 \\ &= \big((\zeta^{12})^2 + (\zeta^{23})^2 + (\zeta^{31})^2\big) \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) \\ &= |\zeta| \cdot |\vec{z}| \cdot (\vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3) \end{aligned}$$

(if ζ and $\vec{z}$ weren't mutually orthogonal, the absolute magnitude of their wedge product $\zeta \wedge \vec{z}$ would have been smaller than $|\zeta| \cdot |\vec{z}|$). The mapping also preserves the absolute magnitude: $|\zeta| = |\vec{z}|$, as it's easy to see from (5.71).

Thus, we want a mapping between a bivector ζ and a vector $\vec{z}$ normal to ζ, such that $|\zeta| = |\vec{z}|$. We can rephrase the vector normality here in more general terms by saying that each time we are mapping a subspace of the 3D vector space to its orthogonal complement. Either we are mapping a plane spanned by ζ to a line spanned by $\vec{z}$, or we are mapping this plane back to this line, the plane and the line being two mutually orthogonal subspaces of the 3D vector space.

We can see how this generalizes to an operation "$\star$" defined for arbitrary k-vectors: given a k-vector ζ in an N-dimensional vector space, we want to map it to a k'-vector $\star\zeta$, such that the vector subspaces spanned by ζ and $\star\zeta$ are mutually orthogonal and $|\star\zeta| = |\zeta|$. Except that, at least for now, maybe we should talk of k-blades rather than k-vectors, as it's not exactly clear what a vector subspace spanned by a non-decomposable k-vector would be.

It's not difficult to realize that $k+k' = N$ therefore $k' = N-k$. This is further reinforced (and made possible in the first place) by the equal dimensionalities of the spaces of k-vectors and $(N-k)$-vectors, given by $\mathrm{C}_N^k = \mathrm{C}_N^{N-k}$.

We should also realize that requiring $\star\zeta$ to be defining an orthogonal complement subspace with respect to ζ together with requiring $|\star\zeta| = |\zeta|$ still doesn't uniquely specify $\star\zeta$, as this leaves two possible orientations of $\star\zeta$. Indeed, $|-(\star\zeta)| = |\star\zeta|$, while the subspaces spanned by $-(\star\zeta)$ and $\star\zeta$ are identical. We need some extra condition to uniquely specify the orientation of ζ. This condition can be specified by requiring that the N-dimensional hypervolume formed by $\zeta \wedge \star\zeta$ is positively (actually, nonnegatively, to include the case $\zeta = 0$) oriented. Notice that this latter requirement is immediately compatible with all of (5.69). E.g. for the 2-vector $\zeta = \vec{e}_1 \wedge \vec{e}_2$ and for the 1-vector $\star\zeta = \vec{e}_3$ we obtain a positively oriented unit 3D volume

$$\zeta \wedge \star\zeta = (\vec{e}_1 \wedge \vec{e}_2) \wedge \vec{e}_3 = \vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3$$

Conversely, for $\zeta = \vec{e}_3$ and $\star\zeta = \vec{e}_1 \wedge \vec{e}_2$ we obtain a positively oriented unit 3D volume again:

$$\zeta \wedge \star\zeta = \vec{e}_3 \wedge (\vec{e}_1 \wedge \vec{e}_2) = \vec{e}_3 \wedge \vec{e}_1 \wedge \vec{e}_2 = \vec{e}_1 \wedge \vec{e}_2 \wedge \vec{e}_3$$

etc.

We are thus arriving at the following definition of the "$\star$" operation:

(geometric definition of Hodge star for k-blades)

Given a k-blade ζ, the $(N-k)$-blade $\star\zeta$ is defined by the following:

- the subspace spanned by $\star\zeta$ is the orthogonal complement of the subspace spanned by ζ (5.72a)
- $|\star\zeta| = |\zeta|$ (5.72b)
- $\zeta \wedge \star\zeta$ is a nonnegatively oriented hypervolume (5.72c)

This operation is referred to as *Hodge star*.[5]

We thereby have

$$\star(\vec{e}_1 \wedge \vec{e}_2) = \vec{e}_3 = \vec{e}_1 \times \vec{e}_2 \tag{5.73a}$$
$$\star(\vec{e}_2 \wedge \vec{e}_3) = \vec{e}_1 = \vec{e}_2 \times \vec{e}_3 \tag{5.73b}$$
$$\star(\vec{e}_3 \wedge \vec{e}_1) = \vec{e}_2 = \vec{e}_3 \times \vec{e}_1 \tag{5.73c}$$

We would therefore generally expect:

(cross product in terms of wedge product)

$$\star(\vec{x} \wedge \vec{y}) = \vec{x} \times \vec{y} \tag{5.74}$$

but, in order to demonstrate that, we'd need to demonstrate the linearity of the just introduced Hodge star operation first.

[5] The Hodge star shouldn't be confused with other usages of the "$\star$" symbol, e.g. to denote cross-correlation.

5.8.2 Linearity

Mutual orthogonality of vector subspaces is a linear property. Also the mapping (5.70) looks like a linear one. Therefore we should expect Hodge star to be a linear operation. Let's verify that.

Goal 5.24. Show that Hodge star is a linear operation. *Hint:* replace the condition (5.72a) with some other condition, expressible algebraically, then expand ζ is a positively oriented orthonormal basis.

It might be easier to argue the linearity of Hodge star if we expressed the subspace orthogonality (5.72a) of ζ and $\star\zeta$ in algebraic terms. We could do it by writing

$$|\zeta \wedge \star\zeta| = |\zeta|^2$$

(since, if the subspaces spanned by ζ and $\star\zeta$ were not orthogonal, the absolute magnitude of their wedge product would be smaller than $|\zeta| \cdot |\star\zeta|$, which, since $|\star\zeta| = |\zeta|$, can be rewritten as $|\zeta|^2$).

Thus, our definition (5.72) of Hodge star turns into the three following conditions

(algebraic definition of Hodge star for k-blades)

$$|\zeta \wedge \star\zeta| = |\zeta|^2 \qquad \text{(ensures the orthogonality of subspaces)} \tag{5.75a}$$

$$|\star\zeta| = |\zeta| \qquad \text{(equal absolute magnitudes)} \tag{5.75b}$$

$$\zeta \wedge \star\zeta \text{ is a nonnegatively oriented hypervolume} \tag{5.75c}$$

With that in mind, let's expand the k-blade ζ in terms of a positively-oriented orthonormal basis:

$$\zeta = \sum_{i_1 \ldots i_k} \left(\zeta^{i_1 \ldots i_k} \cdot (\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})\right) \tag{5.76a}$$

If "$\star$" is a linear operation, then we should have

(expansion of $\star\zeta$ in terms of the components of ζ, can be also used for non-decomposable k-vectors ζ)

$$\star\zeta = \sum_{i_1 \ldots i_k} \left(\zeta^{i_1 \ldots i_k} \cdot \star(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})\right) \tag{5.76b}$$

Let's verify if (5.76b) indeed gives the expression for $\star\zeta$. Computing $\zeta \wedge \star\zeta$ (where ζ and $\star\zeta$ are by (5.76)), we obtain wedge products between different terms of (5.76a) and (5.76b):

$$\begin{aligned}&\left(\zeta^{i_1 \ldots i_k} \cdot (\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})\right) \wedge \left(\zeta^{j_1 \ldots j_k} \cdot \star(\vec{e}_{j_1} \wedge \ldots \wedge \vec{e}_{j_k})\right)\\ &\quad = \zeta^{i_1 \ldots i_k} \zeta^{j_1 \ldots j_k} \cdot (\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}) \wedge \star(\vec{e}_{j_1} \wedge \ldots \wedge \vec{e}_{j_k})\end{aligned} \tag{5.77}$$

Now we argue that, whenever the set of indices $i_1, \ldots, i_k$ differs from $j_1, \ldots, j_k$, we obtain zero from the wedge product (5.77).

First, let's consider the case when these two sets coincide. Then, by the very definition (5.75), we obtain from (5.77):

$$\left(\zeta^{i_1 \ldots i_k} \cdot (\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})\right) \wedge \left(\zeta^{i_1 \ldots i_k} \cdot \star(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})\right)$$

$$
\begin{aligned}
&= (\zeta^{i_1\ldots i_k})^2 \cdot (\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}) \wedge \star(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}) \\
&\overset{(5.75)}{=} (\zeta^{i_1\ldots i_k})^2 \cdot (\vec{e}_1 \wedge \ldots \wedge \vec{e}_N)
\end{aligned}
$$

Indeed, the product $(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}) \wedge \star(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})$, by the very definition (5.75), should yield a positively oriented unit hypervolume, since $\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}$ is a unit-magnitude k-blade. This positively oriented unit hypervolume can be written as $\vec{e}_1 \wedge \ldots \wedge \vec{e}_N$. Notice that the N-blade $\vec{e}_1 \wedge \ldots \wedge \vec{e}_N$ is one and the same positively oriented unit hypervolume, regardless of the choice of the basis vectors $\vec{e}_i$, thus this result is basis-invariant.

Obviously, $\star(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})$ is an $(N-k)$-blade consisting of the remaining basis vectors, which are not included into the set $\vec{e}_{i_1}, \ldots, \vec{e}_{i_k}$. The order in which these remaining vectors are wedge-multiplied together is restricted by the condition (5.75c). E.g.

$$
\star(\vec{e}_1 \wedge \ldots \wedge \vec{e}_k) = \vec{e}_{k+1} \wedge \ldots \wedge \vec{e}_N \tag{5.78}
$$

which is easily verified by

$$
\begin{aligned}
(\vec{e}_1 \wedge \ldots \wedge \vec{e}_k) \wedge \star(\vec{e}_1 \wedge \ldots \wedge \vec{e}_k) &= (\vec{e}_1 \wedge \ldots \wedge \vec{e}_k) \wedge (\vec{e}_{k+1} \wedge \ldots \wedge \vec{e}_N) \\
&= \vec{e}_1 \wedge \ldots \wedge \vec{e}_N
\end{aligned}
$$

On the other hand, all other products $(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}) \wedge \star(\vec{e}_{j_1} \wedge \ldots \wedge \vec{e}_{j_k})$, where the basis vector sets $\vec{e}_{i_1}, \ldots, \vec{e}_{i_k}$ and $\vec{e}_{j_1}, \ldots, \vec{e}_{j_k}$ are different, will have at least one shared basis vector between $\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}$ and $\star(\vec{e}_{j_1} \wedge \ldots \wedge \vec{e}_{j_k})$. Thereby all these products turn to zero and do not participate in the sum emerging from wedge-multiplication of (5.76a) and (5.76b), which thus yields

$$
\begin{aligned}
\zeta \wedge \star\zeta &= \sum_{i_1\ldots i_k} \Big(\big(\zeta^{i_1\ldots i_k} \cdot (\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})\big) \wedge \big(\zeta^{i_1\ldots i_k} \cdot \star(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k})\big) \Big) \\
&= \sum_{i_1\ldots i_k} \big((\zeta^{i_1\ldots i_k})^2 \cdot (\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}) \wedge \star(\vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}) \big) \\
&\overset{(5.75)}{=} \sum_{i_1\ldots i_k} \big((\zeta^{i_1\ldots i_k})^2 \cdot (\vec{e}_1 \wedge \ldots \wedge \vec{e}_N) \big) \\
&= \left(\sum_{i_1\ldots i_k} (\zeta^{i_1\ldots i_k})^2 \right) \cdot (\vec{e}_1 \wedge \ldots \wedge \vec{e}_N) = |\zeta|^2 \cdot (\vec{e}_1 \wedge \ldots \wedge \vec{e}_N)
\end{aligned} \tag{5.79}
$$

Now, by (5.79), the conditions (5.75a) and (5.75c) are immediately satisfied for ζ and $\star\zeta$, while the condition (5.75b) is directly satisfied by (5.76). It follows that (5.76b) is indeed the expression for $\star\zeta$. The linearity of the operation $\star\zeta$ then directly follows from the expression (5.76b).

Goal 5.24 completed. ■

The equality (5.74) expressing the relationship between wedge and cross products follows from (5.73) by the linearity of Hodge star.

(Hodge star for non-decomposable k-vectors)

The linearity of the Hodge star operation allows to naturally extend this operation from k-blades to other k-vectors, where, in order to apply Hodge star to a non-decomposable k-vector we would simply apply it separately to each of the k-blade components of this k-vector.

5.8.3 1-vectors and 0-vectors

We have implicitly introduced the idea that 1-vectors are just the ordinary vectors, thereby the Hodge star operation establishes a bidirectional mapping between $(N-1)$-vectors and ordinary vectors in an N-dimensional vector space. However what about a mapping between N-vectors and 0-vectors? What is actually a 0-vector? Intuition might suggest to us that a 0-vector should be simply a scalar. Let's look a bit deeper into this suggestion, to see if it consistently works out.

Consider an orthonormal basis $\{\vec{e}_i\}$ in an N-dimensional vector space, and consider the application of Hodge star to various subsets of that basis, where we can use the previously discussed formula (5.78). At $k = N-2$, (5.78) yields a 2-vector $\vec{e}_{N-1} \wedge \vec{e}_N$:

$$\star(\vec{e}_1 \wedge \ldots \wedge \vec{e}_{N-2}) = \vec{e}_{N-1} \wedge \vec{e}_N$$

which is verified by

$$(\vec{e}_1 \wedge \ldots \wedge \vec{e}_{N-2}) \wedge (\vec{e}_{N-1} \wedge \vec{e}_N = \vec{e}_1 \wedge \ldots \wedge \vec{e}_{N-1} \wedge \vec{e}_N$$

At $k = N-1$, (5.78) yields a 1-vector $\vec{e}_N$:

$$\star(\vec{e}_1 \wedge \ldots \wedge \vec{e}_{N-1}) = \vec{e}_N$$

which is verified by

$$(\vec{e}_1 \wedge \ldots \wedge \vec{e}_{N-1}) \wedge \vec{e}_N = \vec{e}_1 \wedge \ldots \wedge \vec{e}_{N-1} \wedge \vec{e}_N$$

At $k = N$ all basis vectors are gone from the result of the Hodge star application in (5.78) and we are simply left with a factor of 1:

$$\star(\vec{e}_1 \wedge \ldots \wedge \vec{e}_N) = 1$$

The respective verification expression reads

$$(\vec{e}_1 \wedge \ldots \wedge \vec{e}_N) \wedge 1 = \vec{e}_1 \wedge \ldots \wedge \vec{e}_N$$

and we need to agree that we can wedge-multiply a k-vector by a scalar, which is the same as an ordinary multiplication of a k-vector by a scalar:

(scalars as 0-vectors being used as an operand in a wedge product)

$$\zeta \wedge \lambda = \lambda \wedge \zeta \triangleq \lambda \cdot \zeta \tag{5.80}$$

for any k-vector ζ and any scalar λ.

Conversely, at $k = 0$ all basis vectors are gone from the argument of Hodge star in (5.78) and we are left with

$$\star 1 = \vec{e}_1 \wedge \ldots \wedge \vec{e}_N$$

which is verified by

$$1 \wedge (\vec{e}_1 \wedge \ldots \wedge \vec{e}_N) = \vec{e}_1 \wedge \ldots \wedge \vec{e}_N$$

We thus obtain

(Hodge star applied to a positively oriented unit volume and its reverse)

$$\star\,(\vec{e}_1 \wedge \ldots \wedge \vec{e}_N) = 1 \tag{5.81a}$$
$$\star\,1 = \vec{e}_1 \wedge \ldots \wedge \vec{e}_N \tag{5.81b}$$

By applying linearity of Hodge star to (5.81), we can simply apply Hodge star to any N-vector, obtaining the associated signed hypervolume as a scalar value. Conversely, given a scalar signed hypervolume value, we can convert it to an N-vector by application of Hodge star:

(Hodge star converts N-vectors to the associated signed hypervolumes and back)

$$\star(\vec{x}_1 \wedge \ldots \wedge \vec{x}_N) = V(\vec{x}_1, \ldots, \vec{x}_N) \tag{5.82a}$$
$$\star\big(V(\vec{x}_1 \wedge \ldots \wedge \vec{x}_N)\big) = \vec{x}_1 \wedge \ldots \wedge \vec{x}_N \tag{5.82b}$$

We thereby restored/justified our previous convention of using the "$\star$" operation to convert between N-vectors and scalar signed hypervolume values.

Using (5.82) we can rewrite the algebraic Hodge star definition (5.75) as

("self-referring" algebraic definition of Hodge star for k-blades)

$$\zeta \wedge \star\zeta = \star\left(|\zeta|^2\right) \tag{5.83a}$$
$$|\star\zeta| = |\zeta| \tag{5.83b}$$

where we merged the conditions (5.75a) and (5.75c) into a single condition (5.83a). The application of Hodge star to non-decomposable k-vectors, as usual, is done by linearity. Alternatively, (5.76b) can be used to apply Hodge star to any k-vector, this being particularly convenient to apply in a positively-oriented orthonormal basis.

▷ We described the Hodge star definition (5.83) as "self-referring" because (5.83a) formally applies Hodge star to compute the value of the expression $\star\left(|\zeta|^2\right)$, while the Hodge star is only in the process of being defined. This self-reference is not a real problem, though, since we could agree to understand the expression $\star\left(|\zeta|^2\right)$ "in a purely notational sense" as an N-blade corresponding to a signed positively oriented hypervolume equal to $|\zeta|^2$.

We could have introduced some other notation to convert $|\zeta|^2$ to an N-blade, thereby avoiding self-reference in the Hodge star definition. But then we could drop this notation immediately afterwards, since, once Hodge star is defined, we simply can use Hodge star to denote the same conversion.

5.8.4 Self-inverse property

Applying Hodge star to basis vectors (of a positively oriented orthonormal basis) in 2D, we come to somewhat strange results:

$$\star\vec{e}_1 = \vec{e}_2 \qquad \text{since } \vec{e}_1 \wedge \star\vec{e}_1 = \vec{e}_1 \wedge \vec{e}_2 = \star 1$$

but

$$\star\vec{e}_2 = -\vec{e}_1 \quad \text{since } \vec{e}_2 \wedge \star\vec{e}_2 = \vec{e}_2 \wedge (-\vec{e}_1) = \vec{e}_1 \wedge \vec{e}_2 = \star 1$$

therefore $\star(\star\vec{e}_1) = \star\vec{e}_2 = -\vec{e}_1$, while intuitively we would have expected $\star(\star\vec{e}_1) = \vec{e}_1$. Something probably could be done about it, but we rather shouldn't. We should keep the Hodge star operation basis-independent and this is exactly what we are currently getting: in 2D, Hodge star simply turns vectors by 90° counterclockwise. Had we amended it so that $\star\vec{e}_1 = \vec{e}_2$ but $\star\vec{e}_2 = \vec{e}_1$, this would mean that Hodge star turns some of the vectors counterclockwise and some clockwise. This would either make Hodge star basis-dependent or make it have some "preferred directions". Neither is a desired property.

More generally, suppose $\zeta = \vec{e}_1 \wedge \ldots \wedge \vec{e}_k$ is a k-blade. By (5.78), $\star\zeta = \vec{e}_{k+1} \wedge \ldots \wedge \vec{e}_N$ is an $(N-k)$-blade and $\zeta \wedge \star\zeta = \vec{e}_1 \wedge \ldots \wedge \vec{e}_N = \star 1$, thereby satisfying (5.83). Now, if we are looking to applying Hodge star once again to $\star\zeta$, hoping to obtain ζ back, we will need to compute $\star\zeta \wedge \zeta$ in order to check (5.83). We can obtain $\star\zeta \wedge \zeta$ from $\zeta \wedge \star\zeta$ by the following process. First we take the factor $\vec{e}_{k+1}$ of the product $\zeta \wedge \star\zeta = \vec{e}_1 \wedge \ldots \wedge \vec{e}_N$ and shift it all the way to the left. This requires k swaps $\vec{e}_{k+1} \leftrightarrow \vec{e}_k$, $\vec{e}_{k+1} \leftrightarrow \vec{e}_{k-1}$, $\ldots$, $\vec{e}_{k+1} \leftrightarrow \vec{e}_1$. Now we shift the factor $\vec{e}_{k+2}$ all the way to the second position, again taking k swaps. Then we shift $\vec{e}_{k+3}$, etc., and finally shift $\vec{e}_N$, obtaining $\star\zeta \wedge \zeta$ in the end. We thereby have shifted $N-k$ factors, taking k swaps for each factor, giving $k(N-k)$ swaps in total. Now if $k(N-k)$ is even, we have the total product unchanged. If $k(N-k)$ is odd, the total product changes sign. We thereby obtain $\star\zeta \wedge \zeta = (-1)^{k(N-k)} \cdot \zeta \wedge \star\zeta$ and respectively

(Hodge star's self-inverse)

$$\star(\star\zeta) = (-1)^{k(N-k)} \cdot \zeta \tag{5.84}$$

Notice that $k(N-k)$ is odd iff N is even and k is odd, and this is when $\star\star\zeta = -\zeta$.

▷ Apparently (5.84) holds for any other basis k-blade, just instead of $\zeta = \vec{e}_1 \wedge \ldots \wedge \vec{e}_k$ we consider $\zeta = \vec{e}_{i_1} \wedge \ldots \wedge \vec{e}_{i_k}$. By linearity, the same property extends to arbitrary k-vectors.

To conclude the introduction of Hodge star, we will say that now it should be more clear why pseudovectors and pseudoscalars behave the way they do: they are actually k-vectors just pretending to be vectors or scalars, where the Hodge

star operation maps pseudovectors and pseudoscalars to the respective k-vectors and back. In 3D pseudovectors are mapped to 2-vectors and pseudoscalars to 3-vectors.

5.9 k-vectors as tensors

5.9.1 Bivectors as tensors

Looking at (5.29) we can notice that the components of the wedge product $\vec{x} \wedge \vec{y}$ have the form $x^i y^j - x^j y^i$, which in turn can be understood as a tensor index notation expression, yielding a rank 2 tensor with the following component matrix:

$$\begin{bmatrix} 0 & x^1y^2 - x^2y^1 & x^1y^3 - x^3y^1 \\ x^2y^1 - x^1y^2 & 0 & x^2y^3 - x^3y^2 \\ x^3y^1 - x^1y^3 & x^3y^2 - x^2y^3 & 0 \end{bmatrix} \tag{5.85}$$

Notice that (5.85) is an antisymmetric matrix, and so is the implied tensor

$$z^{ij} = x^i y^j - x^j y^i$$

where we have

$$z^{ij} = -z^{ji}$$

There is thereby a 1:1 correspondence between bivectors and antisymmetric rank 2 tensors:

$$\zeta = \vec{x} \wedge y \quad \longleftrightarrow \quad z^{ij} = x^i y^j - x^j y^i \tag{5.86a}$$

It is not difficult to realize that the same kind of 1:1 mapping between bivectors and antisymmetric rank 2 tensors is possible in any dimension count, not just 3D.

The tensor expression $x^i y^j - x^j y^i$ can be also rewritten in indexless notation as

$$\vec{x} \wedge \vec{y} \quad \longleftrightarrow \quad \vec{x} \otimes \vec{y} - \vec{y} \otimes \vec{x} \tag{5.86b}$$

Indeed, expanding (5.86b) in terms of basis vectors, we obtain

$$\begin{aligned} \vec{x} \wedge \vec{y} &= (x^i \vec{e}_i) \wedge (y^j \vec{e}_j) = x^i y^j \cdot (\vec{e}_i \wedge \vec{e}_j) \overset{(5.86b)}{\longleftrightarrow} x^i y^j \cdot (\vec{e}_i \otimes \vec{e}_j - \vec{e}_j \otimes \vec{e}_i) \\ &= x^i y^j \cdot (\vec{e}_i \otimes \vec{e}_j) - x^i y^j \cdot (\vec{e}_j \otimes \vec{e}_i) = x^i y^j \cdot (\vec{e}_i \otimes \vec{e}_j) - x^j y^i \cdot (\vec{e}_i \otimes \vec{e}_j) \\ &= (x^i y^j - x^j y^i) \cdot (\vec{e}_i \otimes \vec{e}_j) = z^{ij} \cdot (\vec{e}_i \otimes \vec{e}_j) \end{aligned} \tag{5.86c}$$

▷ It may seem like we swapped the factor order from $\vec{e}_j \otimes \vec{e}_i$ to $\vec{e}_i \otimes \vec{e}_j$ in the second line of (5.86c), which is not an allowed operation, since tensor product doesn't commute. But this is not what we actually did. What we did is we swapped the free summation index variables i and j in the second term.

The antisymmetrization of the components of z^{ij} occurring in (5.86a) by letting $z^{ij} = x^i y^j - x^j y^i$ has a shorthand notation by using square brackets:

$$z^{ij} = p^{[ij]} = p^{ij} - p^{ji} \tag{5.87a}$$

where $p^{ij} = x^i y^j$. Or, directly applying the brackets on $x^i y^j$

$$z^{ij} = x^{[i} y^{j]} = x^i y^j - x^j y^i \tag{5.87b}$$

Obviously, such notation shortcut works only if the antisymmetrized indices occur sequentially one after another in the expression, without other indices occurring in between.

Actually... maybe antisymmetrization should not be done exactly as in (5.87). The problem with (5.87) is that if we apply the antisymmetrization a second time on top of an already antisymmetrized tensor, the tensor's components will double. Indeed, since $z^{ij} = p^{ij} - p^{ji}$, it follows that

$$\begin{aligned} z^{[ij]} &= z^{ij} - z^{ji} = (p^{ij} - p^{ji}) - (p^{ji} - p^{ij}) = p^{ij} + p^{ij} - p^{ji} - p^{ji} \\ &= 2 \cdot (p^{ij} - p^{ji}) = 2z^{ij} \end{aligned}$$

This can be amended by dividing the expressions in (5.87) by 2 and redefining the antisymmetrization as:

(antisymmetrization of tensor in two indices)

$$z^{ij} = p^{[ij]} = \frac{1}{2}(p^{ij} - p^{ji}) \tag{5.88a}$$

and

$$z^{ij} = x^{[i}y^{j]} = \frac{1}{2}(x^i y^j - x^j y^i) \tag{5.88b}$$

To stay consistent to that, we also amend (5.86) obtaining

(reinterpretation of wedge product of two vectors as an antisymmetrized tensor product)

$$\zeta = \vec{x} \wedge \vec{y} \quad \longleftrightarrow \quad z^{ij} = \frac{1}{2}(x^i y^j - x^j y^i) \tag{5.89a}$$

and

$$\vec{x} \wedge \vec{y} \quad \longleftrightarrow \quad \frac{1}{2}(\vec{x} \otimes \vec{y} - \vec{y} \otimes \vec{x}) \tag{5.89b}$$

The formula (5.88) seems to be a better option than (5.87), since now antisymmetrization can be repeatedly applied an arbitrary number of times without changing the tensor's value, and it seems to be a more commonly accepted one.

Regardless of, whether we pick (5.87) or (5.88) (and respectively (5.86) or (5.89)), we thereby provided an equivalent representation of the bivector space as a subspace of rank 2 tensors. This subspace is arising from the entire space of rank 2 tensor by requiring the tensors to be antisymmetric.

▷ The readers are encouraged to check whether antisymmetric rank 2 tensors indeed forms a linear space.

5.9.2 3D 3-vectors as tensors

Now we shall try to generalize (5.86) and (5.89) to the triple exterior product. Let's pick (5.86) as a starting point, so that we don't need to deal with an extra normalization coefficient. Starting from there, we might obtain

$$\vec{x} \wedge \vec{y} \wedge \vec{z} = (\vec{x} \wedge \vec{y}) \wedge \vec{z} \quad \overset{(5.86)?}{\longleftrightarrow}$$

$$
\begin{aligned}
&(\vec{x}\otimes\vec{y}-\vec{y}\otimes\vec{x})\otimes\vec{z}-\vec{z}\otimes(\vec{x}\otimes\vec{y}-\vec{y}\otimes\vec{x})\\
&\quad=\vec{x}\otimes\vec{y}\otimes\vec{z}-\vec{y}\otimes\vec{x}\otimes\vec{z}-\vec{z}\otimes\vec{x}\otimes\vec{y}+\vec{z}\otimes\vec{y}\otimes\vec{x}
\end{aligned}
\tag{5.90a}
$$

or, with a different multiplication order

$$
\begin{aligned}
\vec{x}\wedge\vec{y}\wedge\vec{z}=\vec{x}\wedge(\vec{y}\wedge\vec{z})\quad&\overset{(5.86)}{\longleftrightarrow}?\\
&\vec{x}\otimes(\vec{y}\otimes\vec{z}-\vec{z}\otimes\vec{y})-(\vec{y}\otimes\vec{z}-\vec{z}\otimes\vec{y})\otimes\vec{x}\\
&\quad=\vec{x}\otimes\vec{y}\otimes\vec{z}-\vec{x}\otimes\vec{z}\otimes\vec{y}-\vec{y}\otimes\vec{z}\otimes\vec{x}+\vec{z}\otimes\vec{y}\otimes\vec{x}
\end{aligned}
\tag{5.90b}
$$

Neither of (5.90) looks right, and also their results are not equal. The respective versions of (5.86a) would be looking as

$$
\begin{aligned}
w^{ijk}&\overset{?}{=}x^iy^jz^k-y^ix^jz^k-z^ix^jy^k+z^iy^jx^k\\
&=x^iy^jz^k-x^jy^iz^k-x^jy^kz^i+x^ky^jz^i
\end{aligned}
\tag{5.91a}
$$

and

$$
\begin{aligned}
w^{ijk}&\overset{?}{=}x^iy^jz^k-x^iz^jy^k-y^iz^jx^k+z^iy^jx^k\\
&=x^iy^jz^k-x^iy^kz^j-x^ky^iz^j+x^ky^jz^i
\end{aligned}
\tag{5.91b}
$$

This doesn't look like anything similar to the expression generated by (5.25) and (5.26) which is

$$
V(\vec{x},\vec{y},\vec{z})=x_1y_2z_3-x_1y_3z_2+x_2y_3z_1-x_2y_1z_3+x_3y_1z_2-x_3y_2z_1 \tag{5.92}
$$

and thereby has 6 different terms rather than only 4.

Let's therefore start from scratch. The components of the tensor $w\longleftrightarrow\vec{x}\wedge\vec{y}\wedge\vec{z}$ should be antisymmetric in the same way how the product $\vec{x}\wedge\vec{y}\wedge\vec{z}$ is anticommutative. That is, swapping any pair of indices should result in sign inversion. This immediately forces all components with repeated indices to zero: $w^{iij}=0$, $w^{iji}=0$, $w^{jii}=0$ and $w^{iii}=0\ \forall i,j$. Only the components w^{ijk} with all 3 indices being different can be nonzero. But then, swapping any two indices should result in the sign change:

$$
w^{123}=-w^{213}=w^{231}=-w^{321}=w^{312}=-w^{132}
$$

It turns out that all nonzero components of w are having one and the same value, except that half of them are having a different sign!

Now in (5.86a) the components of the tensor z^{ij} were effectively having the value provided by the formula (5.7) (or, equivalently, by (5.6)). Similarly to that, the components of tensor w^{ijk} should have the value equal to the one provided by (5.92) (or, equivalently, by (5.25)). The formula (5.92) suggests

$$
w^{ijk}=x^iy^jz^k-x^iy^kz^j+x^jy^kz^i-x^jy^iz^k+x^ky^iz^j-x^ky^jz^i \tag{5.93a}
$$

(where we simply took (5.92) and replaced 1 by i, 2 by j and 3 by k), or, with a different order of factors in each term (but identical order of indices)

$$
w^{ijk}=x^iy^jz^k-x^iz^jy^k+z^ix^jy^k-y^ix^jz^k+y^iz^jx^k-z^iy^jx^k \tag{5.93b}
$$

which suggests

$$\vec{x}\wedge\vec{y}\wedge\vec{z} \quad\longleftrightarrow\quad \vec{x}\otimes\vec{y}\otimes\vec{z}-\vec{x}\otimes\vec{z}\otimes\vec{y}+\vec{z}\otimes\vec{x}\otimes\vec{y} \\ -\vec{y}\otimes\vec{x}\otimes\vec{z}+\vec{y}\otimes\vec{z}\otimes\vec{x}-\vec{z}\otimes\vec{y}\otimes\vec{x} \tag{5.94}$$

Equations (5.93) and (5.94) are somewhat unwieldy, therefore we shall reobtain them in a different form, starting with (5.6) and (5.25). Comparing (5.6) to (5.86) suggests a different way to write (5.86):

$$\vec{x}^{\,1}\wedge\vec{x}^{\,2} \quad\longleftrightarrow\quad \varepsilon_{ij}\cdot(\vec{x}^{\,i}\otimes\vec{x}^{\,j}) \tag{5.95a}$$

where we switched the notation for the multiplied vectors from $\vec{x}\wedge\vec{y}$ to $\vec{x}^{\,1}\wedge\vec{x}^{\,2}$, and where the indices i, j change in the range from 1 to 2. For (5.89) we respectively obtain

(wedge product of two vectors using 2nd rank Levi–Civita symbol)

$$\vec{x}^{\,1}\wedge\vec{x}^{\,2} \quad\longleftrightarrow\quad \frac{1}{2}\varepsilon_{ij}\cdot(\vec{x}^{\,i}\otimes\vec{x}^{\,j}) \tag{5.95b}$$

Similarly, comparing (5.25) to (5.94) suggests a different way to write (5.94):

$$\vec{x}^{\,1}\wedge\vec{x}^{\,2}\wedge\vec{x}^{\,3} \quad\longleftrightarrow\quad \varepsilon_{ijk}\cdot(\vec{x}^{\,i}\otimes\vec{x}^{\,j}\otimes\vec{x}^{\,k}) \tag{5.96a}$$

or, respectively

(wedge product of three vectors using 3rd rank Levi–Civita symbol)

$$\vec{x}^{\,1}\wedge\vec{x}^{\,2}\wedge\vec{x}^{\,3} \quad\longleftrightarrow\quad \frac{1}{6}\varepsilon_{ijk}\cdot(\vec{x}^{\,i}\otimes\vec{x}^{\,j}\otimes\vec{x}^{\,k}) \tag{5.96b}$$

where this time the factor is $\frac{1}{6}$ rather than $\frac{1}{2}$, since there are $3! = 6$ rather than $2! = 2$ terms involved in antisymmetrization. Respectively, (5.93a) can be simply rewritten (with the optionally added factor of $\frac{1}{6}$) as

(wedge product of vectors in terms of generalized Kronecker delta)

$$\vec{x}\wedge\vec{y}\wedge\vec{z} \quad\longleftrightarrow\quad w^{ijk} = x^{[i}y^{j}z^{k]} = \frac{1}{6}\delta^{ijk}_{lmn}x^{l}y^{m}z^{n} \tag{5.97}$$

where the square brackets around the three indices i, j, k denote the antisymmetrization across these indices, and where we need the symbol δ^{ijk}_{lmn} to be defined as follows:

1. $\delta^{123}_{123} = 1$
2. δ^{ijk}_{lmn} is antisymmetric in any pair of lower indices or any pair of upper indices.

The symbol δ^{ijk}_{lmn} is known as *generalized Kronecker delta* (in its particular form for 3 spatial dimensions and with 3 upper and lower indices respectively). Effectively it just gives the sign of the permutation from (i, j, k) to (l, m, n) (or back). E.g. $\delta^{123}_{123} = 1$, $\delta^{123}_{132} = \delta^{132}_{123} = -1$, $\delta^{122}_{123} = \delta^{123}_{122} = 0$, etc.

These ideas can be further generalized to other dimensions of the vector space and k-blades of higher values of k.

▷ In particular, we can rewrite (5.88b) as

$$z^{ij} = \frac{1}{2}\delta^{ij}_{kl}x^l y^l$$

5.9.3 Hodge star on tensors

It should be approximately clear by now how k-vectors can be reinterpreted in terms of antisymmetric tensors. It would be instructive, is to express the Hodge star operation in terms of antisymmetric tensors as well. In an orthonormal basis, component-wise Hodge star is effectively a "no op" (except that sometimes the components might change their sign), however we need to make sure that the right values are landing at the right tensor components. For simplicity and for the demonstrative purposes, we shall restrict our discussion to 3D space with a positively oriented orthonormal basis.

Goal 5.25. Work out the details of application of Hodge star to 3D k-vectors represented in antisymmetric tensor form, assuming a positively-oriented orthonormal basis. *Hint:* just write down the components.

Let's first consider the case of mapping from bivectors to vectors in 3D (mapping the exterior product to cross product).

Goal 5.25a. Work out the details of mapping bivectors to vectors in 3D.

Given a bivector

$$\zeta = \zeta^{12}\cdot(\vec{e}_1\wedge\vec{e}_2) + \zeta^{23}\cdot(\vec{e}_2\wedge\vec{e}_3) + \zeta^{31}\cdot(\vec{e}_3\wedge\vec{e}_1)$$

we are interested in its corresponding tensor representation z. By the earlier discussion in this section, the components z are given by

$$\begin{bmatrix} z^{11} & z^{12} & z^{13} \\ z^{21} & z^{22} & z^{23} \\ z^{31} & z^{32} & z^{33} \end{bmatrix} = \begin{bmatrix} 0 & \zeta^{12} & -\zeta^{31} \\ -\zeta^{12} & 0 & \zeta^{23} \\ \zeta^{31} & -\zeta^{23} & 0 \end{bmatrix}$$

(in case we don't normalize by $\frac{1}{2}$).

Let $\star z$ be the tensor representation of $\star\zeta$. The notation $\star z$ can also be alternatively and conveniently interpreted as applying Hodge star to the antisymmetric tensor z, thereby we are effectively in the process of introducing the Hodge star operation for antisymmetric tensors. Notice that, in our current specific case of ζ being a 3D bivector, $\star\zeta$ is just a vector and so should be $\star z$, therefore we simply have $\star z = \star\zeta$. The components of $\star z$ then should be

$$\begin{bmatrix} (\star z)^1 \\ (\star z)^2 \\ (\star z)^3 \end{bmatrix} = \begin{bmatrix} \zeta^{23} \\ \zeta^{31} \\ \zeta^{12} \end{bmatrix} = \begin{bmatrix} z^{23} \\ z^{31} \\ z^{12} \end{bmatrix} = \begin{bmatrix} -z^{32} \\ -z^{13} \\ -z^{21} \end{bmatrix} \tag{5.98}$$

Seems we have some options how to formally map z^{ij} to $(\star z)^k$, and it's not obvious if any of those are easy to notate, especially in a generic form. Actually one option, which looks kind of symmetric, would be to try something along the lines of

$$(\star z)^k = \varepsilon_{ij}{}^k z^{ij} \tag{5.99a}$$

where ${\varepsilon_{ij}}^k \triangleq \varepsilon_{ijk}$ (component-wise). We immediately notice, however, that for each k there are two nonzero terms in the right-hand side of (5.99a), e.g. we have

$$(\star z)^3 = {\varepsilon_{12}}^3 z^{12} + {\varepsilon_{21}}^3 z^{21} = z^{12} - z^{21}$$

which may double the value, compared to what we need. We either should have applied the $\frac{1}{2}$ normalization in our tensor representation, or we should apply it now:

$$(\star z)^k = \frac{1}{2}{\varepsilon_{ij}}^k z^{ij} \tag{5.99b}$$

▷ Remarkably, one doesn't actually need to have the tensor z^{ij} antisymmetrized in advance, since the formula (5.99) is performing the antisymmetrization (as a kind of a side-effect) anyway. So one can apply (5.99a) to compute the cross product $\vec{u} = \vec{x} \times \vec{y}$ directly as $u^k = {\varepsilon_{ij}}^k x^i y^j$ (in a positively oriented orthonormal basis).

Notice that $u^k = {\varepsilon_{ij}}^k x^i y^j$ is obtained by using (5.99a) rather than (5.99b), as if the normalization was already applied. This is because when bivectors are represented as antisymmetric tensors, each bivector component occurs twice among the tensor components, however the formula $u^k = {\varepsilon_{ij}}^k x^i y^j$ works on the tensor $x^i y^j$ which hasn't been antisymmetrized yet, and thus doesn't have "duplicated" components. Respectively there is no need to divide the result by 2.

Goal 5.25a completed. ■

Goal 5.25b. Work out the details of mapping vectors to bivectors in 3D.

So, suppose z is a rank 1 tensor, that is a vector. We want to find $\star z$ in terms of antisymmetric tensors (rather than in terms of k-vectors). Looking at (5.99) and considering (5.98), let's try to construct something analogous to (5.99), e.g.

$$(\star z)^{jk} = {\varepsilon_i}^{jk} z^i \tag{5.100a}$$

where ${\varepsilon_i}^{jk} \triangleq \varepsilon_{ijk}$ (component-wise). In particular, we get

$$(\star z)^{12} = {\varepsilon_3}^{12} z^3$$

This suggests that if, by our convention, we expect halved component values in the rank 2 tensor $(\star z)^{12}$, we should rather rewrite

$$(\star z)^{jk} = \frac{1}{2}{\varepsilon_i}^{jk} z^i \tag{5.100b}$$

Thus, given a 3D vector v and a 2nd rank 3D tensor u, related to each other via Hodge star (as we should recall, by (5.84), in 3D Hodge star is always its own inverse), we have

(Hodge star-mapping between vectors and rank 2 tensors in 3D, non-normalized component convention, positively oriented orthonormal basis)

$$v^i = \star(u^{jk}) = \frac{1}{2}{\varepsilon_{jk}}^i u^{jk} \qquad u^{jk} = \star(v^i) = {\varepsilon_i}^{jk} v^i \tag{5.101a}$$

and conversely

(Hodge star-mapping between vectors and rank 2 tensors in 3D, normalized component convention, positively oriented orthonormal basis)

$$v^i = \star(u^{jk}) = \varepsilon_{jk}{}^i u^{jk} \qquad u^{jk} = \star(v^i) = \frac{1}{2}\varepsilon_i{}^{jk} v^i \tag{5.101b}$$

Goal 5.25b completed. ■

Goal 5.25c. Work out the details of the remaining Hodge star application cases in 3D.

Following the ideas of (5.101), we obtain in a totally similar way

(Hodge star-mapping between scalars and rank 3 tensors in 3D, non-normalized component convention, positively oriented orthonormal basis)

$$v = \star(u^{ijk}) = \frac{1}{6}\varepsilon_{ijk} u^{ijk} \qquad u^{ijk} = \star v = \varepsilon^{ijk} v \tag{5.102a}$$

and

(Hodge star-mapping between scalars and rank 3 tensors in 3D, normalized component convention, positively oriented orthonormal basis)

$$v = \star(u^{ijk}) = \varepsilon_{ijk} u^{ijk} \qquad u^{ijk} = \star v = \frac{1}{6}\varepsilon^{ijk} v \tag{5.102b}$$

Goal 5.25c completed. ■

Goal 5.25 completed. ■

Summary of Chapter 5

- We have introduced the exterior (wedge) product as a general framework for the computation of areas, volumes and hypervolumes, where we derived the details of the exterior product operation essentially based just on the exterior product's additivity (plus a number of further minor assumptions).

- This general framework introduces a number of additional spaces (k-vector spaces) associated with the primary vector space, the ordinary vectors automatically also being 1-vectors and the scalars being 0-vectors.

- We have found that the exterior/wedge product is what essentially lies behind the idea of matrix determinants and derived a number of properties of determinants using the interpretation of determinants as volume deformations.

- We have found that k-vectors is what essentially lies behind the idea of pseudotensors.

Part II

Mostly calculus

Chapter 6

Instantaneous and additive quantities

Motivation-based construction of differentiation and integration framework basics. Derivative. Riemann and Riemann–Stieltjes integrals. Indefinite integral (antiderivative). Improper integrals. Differentials. Multiple dimensions. Mean value theorem. Common particular cases of generalized Stokes' theorem.

6.1 Derivative . . . 198
- 6.1.1 Motivation . . . 198
- 6.1.2 Consistency . . . 201
- 6.1.3 Differentiation of arbitrary functions . . . 203
- 6.1.4 Semantical interpretation . . . 204

6.2 Riemann integral . . . 205
- 6.2.1 Motivation . . . 205
- 6.2.2 Consistency . . . 206
- 6.2.3 Arbitrary integration limits . . . 209
- 6.2.4 Semantical interpretation . . . 210

6.3 Indefinite integral . . . 212
- 6.3.1 Antiderivative . . . 212
- 6.3.2 Fundamental theorem of calculus . . . 214
- 6.3.3 Differentiability of integrals . . . 215

6.4 Riemann–Stieltjes integral . . . 218

6.5 Differentials . . . 221
- 6.5.1 Differentials in differentiation context . . . 221
- 6.5.2 Differentials in integration context . . . 222
- 6.5.3 Formal definition . . . 223
- 6.5.4 Higher-order differentials . . . 225
- 6.5.5 Invariance of 1st-order differentials . . . 226

6.6 Improper integrals . . . 229

6.7 Partial derivatives . . . 233
- 6.7.1 Tangent plane . . . 233
- 6.7.2 Tangent lines in arbitrary directions . . . 236
- 6.7.3 Full derivative . . . 237
- 6.7.4 Partial derivative notation . . . 238
- 6.7.5 Gradient . . . 240

6.8 Multidimensional and iterated integrals 241
6.8.1 Area integrals in 2D 241
6.8.2 Unoriented areas in 2D 244
6.8.3 Iterated integrals 247
6.8.4 Area orientation in 2D 251
6.8.5 Volume and hypervolume integrals 252
6.8.6 Further integrals in multiple dimensions 254
6.8.7 Iterated improper integrals 257
6.9 Average value of a function 261
6.9.1 Integral form in 1D 261
6.9.2 Differentiation-based form in 1D 265
6.9.3 Multiple dimensions 266
6.10 Boundary vs. interior behavior of functions 268
6.10.1 Gauss's theorem for electrostatic field 268
6.10.2 Divergence 274
6.10.3 Gauss's theorem for arbitrary fields 277
6.10.4 Green's theorem 283
6.10.5 Stokes' theorem 288
6.10.6 Repeated application of nabla operator 291
Summary of Chapter 6 295

With the introduction of real numbers we become able to freely operate with various limits, as the majority of the limits we want to take finally begin converging to "valid" limiting values. This opens lots of new possibilities in regards to the treatment of various instantaneous and additive values associated with functions. These values come in the forms of derivatives and integrals, which constitute the topic of this chapter.

> ▷ This chapter is primarily intended as a review text, it assumes a preexisting knowledge of the calculus fundamentals, or, at the very least, a parallel study of a proper calculus course.

6.1 Derivative

6.1.1 Motivation

A widely used intuitive demonstration example for the introduction of the concept of the derivative is the idea of velocity. In this book we are going to choose a different illustration, namely the idea of flux of water, in a hope that it will work more intuitively, due to water amount being a naturally additive quantity. Let's therefore imagine water flowing through a pipe into a reservoir (Fig. 6.1). We want to construct a quantified assessment of "how quickly the water is filling the reservoir".

Motivational starter for Chapter 6

Develop a formal approach/framework to express the idea of an "instantaneous speed of change" of a variable and to support further handling of such speeds.

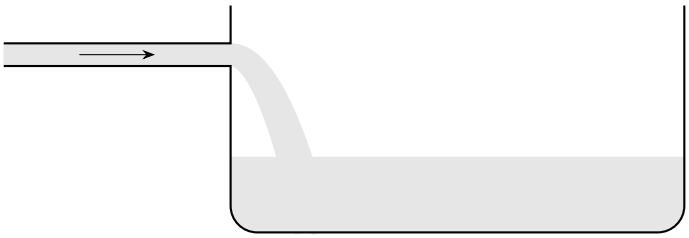


Figure 6.1: Water filling the reservoir.

Goal 6.1. Construct a framework for an exact quantified assessment of how quickly the (idealized) reservoir is being filled. *Hint:* we want to arrive at the idea of the "instantaneous flux" of water through the pipe.

Goal 6.1a. Construct an imaginary procedure for an approximate quantified assessment of the speed of filling the reservoir.

As a first thought experiment we can imagine holding a bucket at the point where the pipe opens into the reservoir for some predefined amount of time, so that we collect the water exiting the pipe into our bucket. Say, we hold the bucket for 10 seconds. Then we could use the amount of collected water β (measured e.g. in liters or in m^3) as a quantification of "how quickly the water is filling the reservoir": the more water is in the bucket, the quicker the reservoir is being filled. If we imagine two such pipe/reservoir setups running in parallel, and if one setup gives β_1 after 10 seconds of holding the bucket, and the other setup gives $\beta_2 = 2\beta_1$ after the same time, we can conclude that the water in the second setup is filling the reservoir "twice as quickly".

Now let's return to a single setup, and suppose that after $\Delta t_1 = 10s$ we got the water amount β_1 in our bucket. Clearly, after $\Delta t_2 = 20s$ we would get $\beta_2 = 2\beta_1$ instead (at least assuming our setup is stable and the water flows equally quickly at all times). Conversely, after $\Delta t_3 = 5s$ we would get $\beta_3 = \frac{1}{2}\beta_1$. Generally, β is linearly proportional to Δt. The coefficient of proportionality is given by the ratio $\frac{\beta}{\Delta t}$. We will refer to this coefficient as the *flux* of water through the pipe's opening:

$$I = \frac{\beta}{\Delta t}$$

Since β depends on t and Δt, a more correct notation would actually be

$$I = \frac{\beta(t, \Delta t)}{\Delta t} \tag{6.1}$$

Even though β depends on t and Δt, we should obtain one and the same value of I from (6.1) independently of t and Δt, at least as long as $\Delta t > 0$.[1]

Now suppose our setup is not that stable and the flux of water does vary with time. This means that we will be getting different values from (6.1) depending on t and Δt. Intuitively, at large Δt we are not going to get a very "precise" value of I. Instead the formula will effectively average (in some way, the details of which do not concern us at this point) the "true flux" (whatever the latter

[1] At small Δt the effects of the limited precision of measurement begin to show up, but we are going to ignore that in our idealized discussion.

means). We will think of (6.1) as giving us the *average flux* on the time interval $[t, t + \Delta t]$.

Goal 6.1a completed. ■

Goal 6.1b. Turn the just developed approximate assessment procedure into an (idealized) exact one.

Intuitively, the smaller is Δt, the "more precise" is the value given by (6.1). At sufficiently small Δt we will be effectively getting the "true flux" value (again, whatever that means). Since we don't know how small Δt needs to actually get (because we don't know how strongly the "true flux" is varying, nor did we specify how precise do we want the obtained flux value to be), in our thought experiment we should let Δt become arbitrarily small, thereby turning (6.1) into

$$I = \lim_{\Delta t \to +0} \frac{\beta(t, \Delta t)}{\Delta t} \tag{6.2}$$

(where the *right-sided limit* $\Delta t \to +0$ means that we are interested in an infinitely small, but still positive value of Δt).

The right-hand side of (6.1) was formally dependent on t and Δ. We ignored that dependency as long as we could assume a constant flux, but now we can't do so anymore. On the other hand, the right-hand side of (6.2) depends only on t, since the dependency on Δt is "eaten up" by the limit. We therefore can rewrite (6.2) with an explicit dependency on t in the left-hand side:

$$I(t) = \lim_{\Delta t \to +0} \frac{\beta(t, \Delta t)}{\Delta t} \tag{6.3}$$

Goal 6.1b completed. ■

The equation (6.3) gives us the "true" or *instantaneous* flux of water. Essentially, it already incorporates the key part of the idea of the derivative. However we still need to do a couple of further steps.

Goal 6.1c. Reexpress the previous result in terms of the amount of water in the reservior, thereby avoiding the need to "use" the bucket.

First, imagine that we are somehow capable of measuring the total amount of water Q in the reservoir. This amount varies with time, so we should write $Q(t)$. Then we can refrain from using the bucket and rely on the measurements of $Q(t)$ instead. Specifically we have

$$\beta(t, \Delta t) = Q(t + \Delta t) - Q(t) \tag{6.4}$$

where $\beta(t, \Delta t)$ is the amount of water which *would have been* collected into the bucket, had we continued using the bucket (but we don't). Indeed, the water coming in through the pipe is the only source of change of the total amount of water in the reservoir. Over the time range $[t, t + \Delta t]$ this water could go into the bucket, resulting in collecting the amount of $\beta(t, \Delta t)$, or it could go into the reservoir, increasing the total amount of water in the reservoir by exactly the same amount $\beta(t, \Delta t)$.

Substituting (6.4) into (6.3) yields

$$I(t) = \lim_{\Delta t \to +0} \frac{Q(t + \Delta t) - Q(t)}{\Delta t} \tag{6.5}$$

which still gives exactly the same instantaneous flux of water that was given by (6.3), just this time employing $Q(t)$ instead of $\beta(t, \Delta t)$.

Goal 6.1c completed. ■

Equation (6.5) looks almost exactly like the definition of the derivative.[2] But let's tune it further.

6.1.2 Consistency

Currently, the expression under the limit computes the average flux on the time interval $[t, t + \Delta t]$, which (since $\Delta t > 0$) is located to the right of the time moment t, at which we want to compute the instantaneous flux, that is we are computing the average over an infinitely short future interval. This seems quite unfair with respect to the past.

Goal 6.1d. Make (6.5) symmetric with respect to the time direction. Make sure that your proposed solution respects the present, and that contributions from the past and the future are consistent with each other.

We could make (6.5) fully symmetric by averaging over a symmetric interval $[t - \Delta t/2, t + \Delta t/2]$ instead:

$$I(t) = \lim_{\Delta t \to +0} \frac{Q(t + \Delta t/2) - Q(t - \Delta t/2)}{\Delta t}$$

This would kind of work and even be good enough, except that it completely ignores the value of Q at the central moment t (since $\Delta t > 0$ under the limit sign, and since we cannot allow $\Delta t = 0$ anyway, as it would immediately result in dividing zero by zero). If Q is slowly varying, this won't be a problem, but it would be better to keep the central moment t explicitly included into the formula, as it was in (6.5). If anything, this will allow the formula to automatically detect any "inacceptable features" of Q (such as a sudden jump, instead of varying slowly) occurring at the moment of interest t.

Therefore, instead of taking symmetric intervals $[t - \Delta t/2, t + \Delta t/2]$, we could allow "negative direction" intervals $[t - |\Delta t|, t]$ which would be implied in (6.5) at $\Delta t < 0$, in addition to the "positive direction" intervals $[t, t + \Delta t]$ occurring at $\Delta t > 0$. For such negative direction intervals the expression under the limit sign in (6.5) gives

$$\begin{aligned} \frac{Q(t + \Delta t) - Q(t)}{\Delta t} &= \frac{Q(t - |\Delta t|) - Q(t)}{-|\Delta t|} = \frac{Q(t) - Q(t - |\Delta t|)}{|\Delta t|} \\ &= \frac{\beta(t - |\Delta t|, |\Delta t|)}{|\Delta t|} \qquad (\text{for } \Delta t < 0) \end{aligned}$$

where, essentially, both numerator and denominator change sign from positive to negative, so the division result is again positive. We thus measure the average flux over the time range $[t - |\Delta t|, t]$. If the instantaneous flux doesn't change significantly over the total interval $[t - |\Delta t|, t + |\Delta t|]$ (and we assume it doesn't, provided Δt is small enough), then any measurement of the amount of flowed

[2] Actually it looks like the definition of the right-sided derivative, but we don't want that at this moment.

water over the time length $|\Delta t|$ will yield approximately one and the same value regardless of the specific starting point, as long as our measurement interval is still within $[t - |\Delta t|, t + |\Delta t|]$. Specifically, the measurements over $[t - |\Delta t|, t]$ and $[t, t + |\Delta t|]$ should yield approximately one and the same value:

$$\frac{Q(t) - Q(t - |\Delta t|)}{|\Delta t|} = \frac{\beta(t - |\Delta t|, |\Delta t|)}{|\Delta t|} \approx \frac{\beta(t, |\Delta t|)}{|\Delta t|} = \frac{Q(t + |\Delta t|) - Q(t)}{|\Delta t|}$$

As Δt becomes smaller, the approximate equality becomes more and more precise and in the limit we obtain one and the same value:

$$I(t) = \lim_{\Delta t \to +0} \frac{Q(t + \Delta t) - Q(t)}{\Delta t} = \lim_{\Delta t \to +0} \frac{Q(t) - Q(t - \Delta t)}{\Delta t}$$

which is simply equal to the ordinary bilateral limit

$$I(t) = \lim_{\Delta t \to 0} \frac{Q(t + \Delta t) - Q(t)}{\Delta t} \tag{6.6}$$

Goal 6.1d completed. ■

The expression (6.6) allows one, knowing how the total amount of water in the reservoir changes over time (this knowledge is encoded in the function $Q(t)$), to compute the instantaneous flux of water through the pipe's opening at any desired time moment (this is represented by the function $I(t)$). This expression, essentially, already contains the definition of $I(t)$ as the derivative of $Q(t)$. But to better match one of the standard derivative notations $\frac{\mathrm{d}Q}{\mathrm{d}t}$ we can still rewrite this expression in a slightly different form.

Let's introduce the notation

$$\Delta Q(t, \Delta t) = Q(t + \Delta t) - Q(t) \tag{6.7a}$$

thereby $\Delta Q(t, \Delta t)$ is the amount of water flowed through the pipe's opening over the time interval $[t, t + \Delta t]$, which thereby coincides with $\beta(t, \Delta t)$ (or, if $\Delta t < 0$, with $\beta(t - |\Delta t|, |\Delta t|)$). In this notation (6.6) becomes

$$I(t) = \lim_{\Delta t \to 0} \frac{\Delta Q(t, \Delta t)}{\Delta t} \tag{6.7b}$$

The right-hand side of (6.7b) is referred to as the *derivative* (at the point t) of the function $Q(t)$:

(definition of the derivative of a function of one argument)

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{\Delta Q(t, \Delta t)}{\Delta t} \tag{6.8}$$

▷ Other commonly used notations include $\frac{\mathrm{d}}{\mathrm{d}t}Q$, Q' or $\dot{Q}$ (the latter is usually restricted to the cases where t has time semantics). Since derivative itself is a function of the original function's argument, this argument can be notated explicitly, e.g. $Q'(t)$ or $\frac{\mathrm{d}Q}{\mathrm{d}t}(t)$.

Using the just introduced derivative notation, we can rewrite (6.7b) as

$$I(t) = \frac{\mathrm{d}Q}{\mathrm{d}t} \tag{6.9}$$

▷ The letters "d" in the numerator and denominator of $\frac{\mathrm{d}Q}{\mathrm{d}t}$ can be intuitively understood as special versions of the letter "Δ". Whereas "Δt" stands for a finite time difference or interval, "$\mathrm{d}t$" stands for an infinitely small time difference. Similarly, "$\mathrm{d}Q$" is an infinitely small version of the finite difference "ΔQ". The expressions $\mathrm{d}Q$ and $\mathrm{d}t$ are referred to as the *differential of* Q and *differential of* t respectively. We shall return to these expressions and their meanings in Subsection 6.5.

6.1.3 Differentiation of arbitrary functions

Given a function $f(x)$, the process of obtaining the derivative of $f(x)$ is referred to as *differentiation* of $f(x)$. With our previous pipe/reservoir illustration we were differentiating a (nonstrictly) increasing function $Q(t)$, since we were only talking of the water flowing through the pipe into the reservoir. Respectively we had $\Delta Q \geq 0\ \forall \Delta t > 0$ (and $\Delta Q \leq 0\ \forall \Delta t < 0$), therefore we would always obtain a nonnegative flux $I(t) \geq 0\ \forall t$.

There is however nothing stopping us from considering a function $f(t)$ which is not restricted to be increasing everywhere on $\mathbb{R}$, but is also allowed to decrease (e.g. in the pipe/reservoir example we could imagine having a reversible pump, which allows to pump the water both into and out of the reservoir). On the intervals where $f(t)$ is (nonstrictly) decreasing we would respectively get $\Delta f \leq 0$ $\forall \Delta t > 0$ and $\Delta f \geq 0\ \forall \Delta t < 0$, yielding a nonpositive derivative $f'(t) \leq 0$.

▷ Notice that a function being strictly increasing doesn't guarantee a strictly positive derivative. A simple counterexample is $f(t) = t^3$ which is strictly increasing everywhere but $f'(0) = 0$.

There is also nothing stopping us from letting the function to take vector (or matrix, etc.) values, where we effectively apply the idea of (6.7) to vector-valued (or matrix-valued, etc.) functions.

Goal 6.1e. Apply the obtained formal definition of the derivative to vector-valued functions of a real argument.

To differentiate a vector-function we simply differentiate each of its components:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \mathrm{d}x_1/\mathrm{d}t \\ \mathrm{d}x_2/\mathrm{d}t \\ \mathrm{d}x_3/\mathrm{d}t \end{bmatrix}$$

which is justified by

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= \lim_{\Delta t \to 0} \frac{\Delta \mathbf{x}}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbf{x}(t+\Delta t) - \mathbf{x}(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \begin{bmatrix} x_1(t+\Delta t) - x_1(t) \\ x_2(t+\Delta t) - x_2(t) \\ x_3(t+\Delta t) - x_3(t) \end{bmatrix} \\ &= \begin{bmatrix} \lim\big((x_1(t+\Delta t) - x_1(t))/\Delta t\big) \\ \lim\big((x_2(t+\Delta t) - x_2(t))/\Delta t\big) \\ \lim\big((x_3(t+\Delta t) - x_3(t))/\Delta t\big) \end{bmatrix} = \begin{bmatrix} \mathrm{d}x_1/\mathrm{d}t \\ \mathrm{d}x_2/\mathrm{d}t \\ \mathrm{d}x_3/\mathrm{d}t \end{bmatrix} \end{aligned}$$

Goal 6.1e completed. ■

Goal 6.1f. Apply the obtained formal definition of the derivative to complex-valued functions of a real argument.

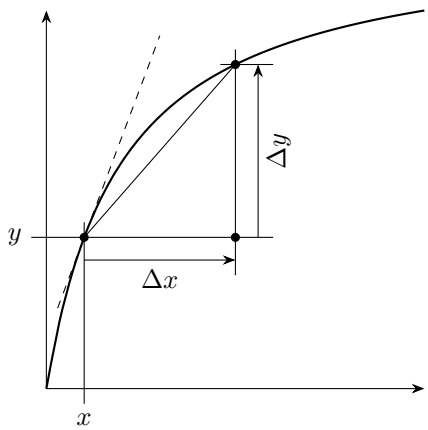


Figure 6.2: Derivative is equal to the tangent of the function's slope angle.

In a similar way, differentiation of a complex-valued function of a real argument independently differentiates its real and imaginary parts:

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\operatorname{Re} w + i \operatorname{Im} w) = \frac{\mathrm{d}}{\mathrm{d}t} \operatorname{Re} w + i \frac{\mathrm{d}}{\mathrm{d}t} \operatorname{Im} w \tag{6.10}$$

▷ Formally (6.10) applies to complex-valued functions of a complex argument as well, but it's not so useful, since in this case $\mathrm{d}t$ is complex and so are both terms in the right-hand side of (6.10).

Goal 6.1f completed. ■

6.1.4 Semantical interpretation

In the pipe/reservoir thought experiment we obtained the water flux as the result of differentiation of the total water amount. Whenever the function argument has time semantics, differentiation of that function yields the instantaneous speed of change of the function's value. If $x = f(t)$ is the x-coordinate of a point (varying over time t), then $f'(t)$ gives the x-velocity of that point. Respectively if $\mathbf{x}(t)$ is the 3D position of a point, then $\mathbf{x}'$ gives the instantaneous velocity of that point.

Goal 6.1g. Find out how derivative can be thought of visually in terms of properties of the function's graph.

Looking at Fig. 6.2 it's easy to realize that, graphically, differentiation yields the tangent of the function's slope angle. Indeed, the tangent of the slope of the secant line connecting the point at x, $y = f(x)$ to the point at $x + \Delta x$, $y + \Delta y = f(x + \Delta x)$ is given by $\Delta y/\Delta x$. In the limit $\Delta x \to 0$ the secant line is approaching the tangent line (given by the dashed line), respectively the tangent $\Delta y/\Delta x$ of the secant slope is approaching the tangent of the function graph's slope.

Goal 6.1g completed. ■

Goal 6.1 completed. ■

6.2 Riemann integral

6.2.1 Motivation

In Subsection 6.1, given the function $Q(t)$ describing how the total amount of water in the reservoir varies over time, we found the function $I(t)$ describing how the instantaneous flux of water varies over time. In this subsection we are going to be solving the opposite problem: given $I(t)$ describing how the instantaneous flux varies over time, find $Q(t)$ describing how the total amount varies over time.

Goal 6.2. Construct a framework which is the opposite of (6.8): given the flux of water as a function of time, we want to know the total amount of water in the reservoir as a function of time. *Hint:* use a limiting procedure, similar to the one we used to arrive at (6.8).

Goal 6.2a. Given a reservoir that is empty at the initial time moment $t = t_0$, find the amount of water $Q(t)$ at $t > t_0$.

Suppose at the initial time moment $t = t_0$ the reservoir is empty ($Q(t_0) = 0$), and we begin pumping water into the reservoir. Suppose further, like we did in the discussion of differentiation, that the flux of water is not varying too quickly (whatever that means). So let's take the flux value at $t = t_0$ and assume that it stays approximately the same over the first minute. The total amount of incoming water over this time interval $\Delta t = 1min$ can be estimated as $I(t_0)\Delta t$. By the same reasoning, the total amount of water incoming over the second minute can be estimated as $I(t_1)\Delta t$, where $t_1 = t_0 + \Delta t$. The amount corresponding to the third minute can be estimated as $I(t_2)\Delta t$ where $t_2 = t_1 + \Delta t$, etc. After N minutes we are going to have a total sum of N terms, each term corresponding to the water amount assessment for the respective time interval $[t_n, t_{n+1}]$:

$$Q(N\Delta t) = \sum_{n=0}^{N-1} I(t_n)\Delta t \quad \text{where } t_n = t_0 + n\Delta t \tag{6.11}$$

Fig. 6.3 illustrates.

We can increase the precision of the assessment by (6.11) by taking a smaller time step, in which case $I(t_n)$ will improve in their roles as assessments of the average flux on the respective time intervals (since the time intervals become smaller). E.g. instead of $\Delta t = 1min$ we could take $\Delta t = 1s$. Thereby we'd need 60 times more intervals to reach the same time moment. Generally, if we want to obtain the value of Q at time t (where $t > t_0$), we can decide how many intervals we want to use, and, given the desired number of intervals N, we respectively let $\Delta t = (t - t_0)/N$. Formula (6.11) thereby turns into

$$Q(t) = \sum_{n=0}^{N-1} I(t_n)\Delta t \quad \text{where } \Delta t = \frac{t - t_0}{N},\ t_n = t_0 + n\Delta t \tag{6.12}$$

The larger is N, the smaller is Δt, and the more precisely we are going to compute $Q(t)$. At $N \to \infty$ we have $\Delta t \to 0$ and the value obtained from (6.12)

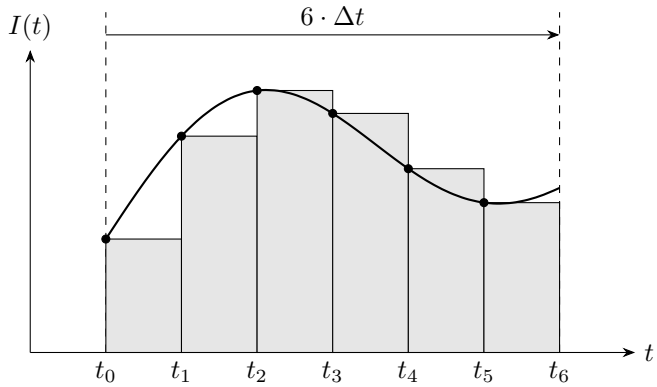


Figure 6.3: Approximate computation of the total amount of water flowed into the reservoir. Each term of (6.11) contributes the amount equal to the area of the respective rectangle.

can be considered as the true and exact value of $Q(t)$ at the time t:

$$Q(t) = \lim_{N\to\infty} \sum_{n=0}^{N-1} I(t_n)\Delta t \quad \text{where } \Delta t = \frac{t - t_0}{N},\ t_n = t_0 + n\Delta t \tag{6.13}$$

▷ Notice that the choice of t_n in (6.13) thereby depends on N and more strictly needs to be notated as $t_n(N)$ or something like that, but we don't do that for the sake of brevity of notation, understanding this dependency implicitly.

Goal 6.2a completed. ■

Formula (6.13) already contains the key part of the idea of the definite integral, but we still need to do some fine-tuning.

6.2.2 Consistency

We could notice that we don't have to take the instantaneous flux at the beginning of the interval $[t_n, t_{n+1}]$ as an assessment of the average flux on that interval. Instead we can take the instantaneous flux at some other time moment $\tau_n \in [t_n, t_{n+1}]$. So we could rewrite (6.13) as

$$Q(t) = \lim_{N\to\infty} \sum_{n=0}^{N-1} I(\tau_n)\Delta t \quad \text{where } \Delta t = \frac{t - t_0}{N},\ t_n = t_0 + n\Delta t,\ \tau_n \in [t_n, t_{n+1}] \tag{6.14}$$

Fig. 6.4 illustrates.

The change from (6.13) to (6.14) might look totally unnecessary from the practical standpoint, since we should expect the values delivered by both formulas to be identical. But (6.14) has an important consistency check embedded into it, not unlike the situation where we insisted, also for consistency checking reasons, that the derivative formula should explicitly include the value of the differentiated function at the central reference point. With (6.14) we expect that small variations in choices of τ_n should not strongly affect the value of

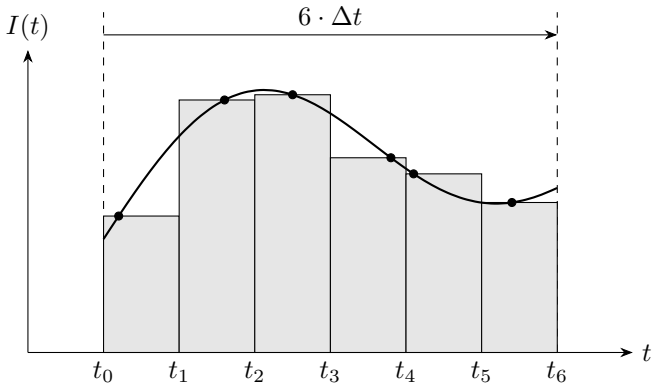


Figure 6.4: Computation of the total amount of water flowed into the reservoir using the sum in (6.14). Each term of the sum in (6.14) contributes the amount equal to the area of the respective rectangle.

the sum and become negligible in the limit. Should this not be the case, we might be obtaining different limit values depending on our choices of τ_n (or for some of the choices of τ_n the limit might diverge altogether), and then which of these different limit values gives the correct value of the total amount of water? Such inconsistencies, if occurring, mean that our function is too "exotic" to be a meaningful representation of water flux.[3]

The limit in (6.14) should be understood in the sense that, no matter which choice of τ_n we make (where this choice also needs to be done from scratch for each N, so it would be more formally correct to notate them as $\tau_n(N)$, but we omit the index N understanding it implicitly), the limit (6.14) converges to one and the same value. Only if that is the case, we consider the limit in (6.14) as converging and respectively $Q(t)$ obtained via (6.14) as having a defined value. This consistency check effectively makes sure that our assumption of $I(t)$ "changing slowly" indeed applies for a given $I(t)$, at least for sufficiently small Δt.

▷ The idea of arbitrary $\tau_n \in [t_n, t_{n+1}]$ in (6.14) is particularly captured in the idea of *Darboux sums*, where in the upper Darboux sum we choose τ_n at the maximum of $I(t)$ on $[t_n, t_{n+1}]$, and in the lower Darboux sum we choose τ_n at the minimum of $I(t)$ on $[t_n, t_{n+1}]$. Thereby the upper Darboux sum represents the maximum possible value of the right-hand side of (6.14) and the lower Darboux sum represents the minimum possible value of the right-hand side of (6.14). One small detail: since the minimum or a maximum of a continuous-argument function $I(t)$ on an interval $[t_n, t_{n+1}]$ may not exist, the minimum and maximum possible values of $I(t)$ are replaced by infimum and supremum.

There is one more consistency check that we need to embed into our formula. So far all intervals $[t_n, t_{n+1}]$ had one and the same length $\Delta t = (t - t_0)/N$. But they don't need to. It should be totally fine to use intervals of different lengths, as long as their lengths progressively become indefinitely small. Again, if we start getting inconsistent answers in this case, this indicates that something is wrong with our function $I(t)$ as a representation of the water flux. We therefore

[3]Technically we will say that the function is not *Riemann-integrable*.

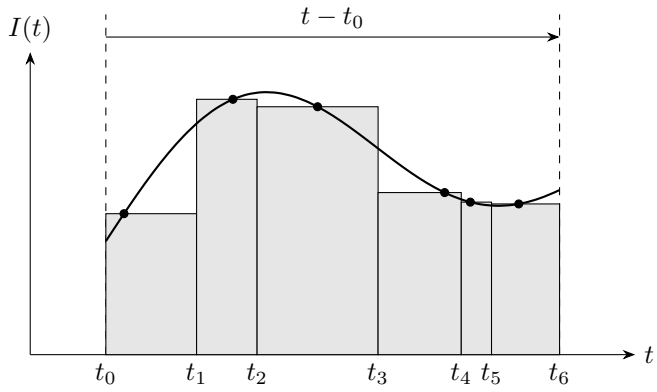


Figure 6.5: Computation of the total amount of water flowed into the reservoir using non-uniform partitioning of the time interval $[t_0, t]$ (where $t = t_6$).

introduce the concept of *partitioning* the total interval of interest $[t_0, t]$ into subintervals $[t_n, t_{n+1}]$ of possibly different lengths:

$$t_0 < t_1 < t_2 < \ldots < t_{n-1} < t_N = t$$

$$\Delta t_0 + \Delta t_1 + \ldots + \Delta t_{N-1} = t - t_0 \quad (\text{where } \Delta t_n = t_{n+1} - t_n > 0)$$

Fig. 6.5 illustrates.

In this case, however, we cannot simply let $N \to \infty$, because it could happen that the lengths of some of the subintervals do not approach zero. E.g. it could be that the first subinterval $[t_0, t_1]$ occupies exactly the first half of the entire interval $[t_0, t]$, while all remaining subintervals are equally spread in the second half of the entire interval $[t_0, t]$, their lengths being respectively equal to $(t - t_0)/2(N - 1)$. Therefore the lengths of all subintervals would approach zero, except for the length of the first subinterval, whose length stays constant at $(t - t_0)/2$.

To be able to specify that the lengths of all subintervals should approach zero, we introduce the idea of *partition's norm* Δ, which we define as the maximum of the subinterval lengths:

$$\Delta = \max\{\Delta t_n\} \tag{6.15}$$

Then instead of letting $N \to \infty$ we let $\Delta \to 0$:

$$Q(t) = \lim_{\Delta \to 0} \sum_{n=0}^{N-1} I(\tau_n)\Delta t_n \quad \text{where } \tau_n \in [t_n, t_{n+1}] \tag{6.16}$$

The limit in (6.16) is considered converging, only it converges to one and the same value regardless of our choices of τ_n and t_n.

▹ The sum under the limit sign in (6.16) is referred to as *Riemann sum*.

Essentially, (6.16) is already the definition of the *Riemann integral*:

(definition of Riemann integral)

$$\int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau = \lim_{\Delta\to 0}\sum_{n=0}^{N-1} I(\tau_n)\Delta t_n \tag{6.17}$$

except that so far it requires $t > t_0$ (while the full definition allows for arbitrary t), this shall be addressed a bit later.

▷ Notice that in the integral notation in the left-hand side of (6.17) we used τ at the position where the right-hand side uses τ_n, and we used $\mathrm{d}\tau$ in place of Δt_n. We cannot use the letter t as the argument of $I(\cdot)$ under the sign of the integral, since it's already being used to denote the *upper integration limit* on the integral sign. For the same reason we also shouldn't use the letter t instead of τ in the differential $\mathrm{d}\tau$, since Δt_n are not any kind of increments of the upper integration limit t, but rather the differences of the partitioning points t_n. On the other hand, using the letter τ in both $I(\tau)$ and $\mathrm{d}\tau$ under the integral sign is pretty intuitive, especially if we relabel $\Delta t_n = t_{n+1} - t_n$ in the right-hand side of (6.17) as $\Delta\tau_n = t_{n+1} - t_n$.

Notably, in the latter case in principle we could even redefine $\Delta\tau_n$ from $\Delta\tau_n = t_{n+1} - t_n$ to $\Delta\tau_n = \tau_{n+1} - \tau_n$ (and changing the upper summation limit in (6.17) to $N-2$), as it wouldn't change the obtained limiting value in (6.17), but this is not how we want to intuitively understand the Riemann integral, so we should not do that.

▷ In practice, the notation $\int_{t_0}^{t} I(t)\,\mathrm{d}t$, using the letter t both under the integral sign and in an integration limit, can be sometimes encountered, and in many cases it even wouldn't cause too much confusion, but formally it is incorrect.

Using the integral notation we can rewrite (6.16) as

$$Q(t) = \int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau \tag{6.18}$$

6.2.3 Arbitrary integration limits

Goal 6.2b. Generalize the previous results to the case of the initially non-empty reservoir, as well as to the case $t < t_0$.

The formula (6.16) and respectively (6.18) were assuming that initially the reservoir is empty: $Q(t_0) = 0$. If we formally choose a larger value of t_0, when the reservoir is not empty anymore, the amount of water $Q(t_0)$ already contained in the reservoir at the start of the integration obviously needs to be added to the amount of water that flowed into the reservoir after t_0, yielding the total value of

$$Q(t) = Q(t_0) + \int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau \tag{6.19}$$

Now that we have shifted t_0 to a later time point, t_0 is not "the beginning of history" anymore, and we could ask ourselves about the values that $Q(t)$ had prior to t_0. However the definition (6.17) requires $t > t_0$ and therefore cannot be used to obtain the information about $Q(t)$ at $t < t_0$, and respectively neither can the formula (6.19). We can however revert the roles of t and t_0 and treat

t as the initial time moment, which can be done by formally swapping t and t_0 in (6.19) yielding

$$Q(t_0) = Q(t) + \int_t^{t_0} I(\tau)\,\mathrm{d}\tau \quad (t_0 > t)$$

or

$$Q(t) = Q(t_0) - \int_t^{t_0} I(\tau)\,\mathrm{d}\tau \quad (t < t_0) \tag{6.20}$$

Comparing (6.20) to (6.19) suggests that we simply define the Riemann integral for the case $t < t_0$ as

(definition of Riemann integral for $t < t_0$)

$$\int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau \triangleq -\int_t^{t_0} I(\tau)\,\mathrm{d}\tau \tag{6.21}$$

so that (6.19) becomes applicable both for $t > t_0$ and $t < t_0$.

Finally by postulating that

(definition of Riemann integral for $t = t_0$)

$$\int_{t_0}^{t_0} I(\tau)\,\mathrm{d}\tau \triangleq 0 \tag{6.22}$$

we make (6.19) applicable for any t (within the domain of $I(t)$).

▷ The definition (6.17) can be made to accommodate arbitrary t (while staying compatible to (6.21) and (6.22)) by letting the partitioning subintervals $[t_n, t_{n+1}]$ "go in the negative direction". That is removing the requirement $t_n < t_{n+1}$ (thereby letting $t_{n+1} < t_n$ or even $t_{n+1} = t_n$). This lets all Δt_n, besides taking positive values, also take negative or zero values. Notice that in this case the partition norm definition (6.15) has to be adjusted to

$$\Delta = \max\{|\Delta t_n|\}$$

Also the interval notation $[t_n, t_{n+1}]$ in the requirement $\tau_n \in [t_n, t_{n+1}]$ becomes formally questionable for $t_n > t_{n+1}$, but that is a purely formal technical detail, which can be easily adjusted by understanding $[t_n, t_{n+1}]$ as $[t_{n+1}, t_n]$ in cases where $t_{n+1} > t_n$.

▷ Notably, formula (6.13) (which is essentially the definition of the Riemann integral without "consistency adjustments") already accommodates arbitrary t "out of the box". The same can be said about "the first consistency adjustment" (6.14).

Goal 6.2b completed. ■

6.2.4 Semantical interpretation

We constructed integration as the inverse of differentiation. By differentiating the amount $Q(t)$ of water in the reservoir we were obtaining the speed of change of $Q(t)$, which is equal to the flux $I(t)$ of water through the pipe. Conversely,

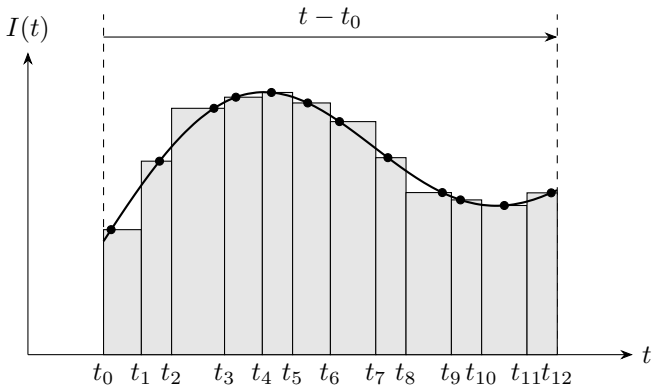


Figure 6.6: A finer-partition version of Fig. 6.5.

by integrating the flux $I(t)$ (the speed of change of $Q(t)$), we were obtaining $Q(t)$ back.

Generally, whenever the function's argument has time semantics, differentiating a quantity gives the speed of change of that quantity, and integrating this speed of change back gives the original quantity. E.g. if $x = f(t)$ gives the x-coordinate of a point, then $f'(t)$ gives the x-velocity of a point, and integrating the x-velocity gives back the x-coordinate. Respectively if $\mathbf{x}$ is a 3D position and $\mathbf{x}'$ is the corresponding 3D velocity, then integration of the velocity $\mathbf{x}'$ gives back the position $\mathbf{x}$.

If the argument doesn't have time semantics, then, as with the derivative, graphical interpretation could be particularly useful. By inverting the graphical interpretation of the derivative, we might come to the idea that integration of the tangent of the function's slope angle will give back the original function. This interpretation is correct, but has little to do with the way how we constructed the Riemann integral.

Looking at Fig. 6.5 and imagining the partition norm approaching zero, we can realize that the respective Riemann sums are simply approaching the area under the function's graph. This is illustrated by Fig. 6.6, which, compared to Fig. 6.5, has a smaller partition norm and respectively better approximates the area under the curve. Thus, the value of the Riemann integral can be visually thought of as the area under the function's graph.

In principle, the integrated function can be taking not only nonpositive, but also negative values, so that some of the terms $I(\tau_n)\Delta t_n$ will become negative. This motivates defining the area under the function's graph as a signed value, where the area becomes negative in the ranges where the function goes below zero (Fig. 6.7). Similarly, the area changes sign if the integration limits are swapped and Δt_n become negative.

▷ Integration can be seen as a generalization of multiplication to the case where "one of the multipliers is not constant". Indeed, suppose the water flux through the pipe is constant: $I(t) \equiv I$. In this case the amount of water going through the pipe over the time interval $[t_0, t_1]$ is simply $I \cdot (t_1 - t_0)$. However if the flux is varying, then, instead of multiplying I by $(t_1 - t_0)$, we need to integrate it over $[t_0, t_1]$.

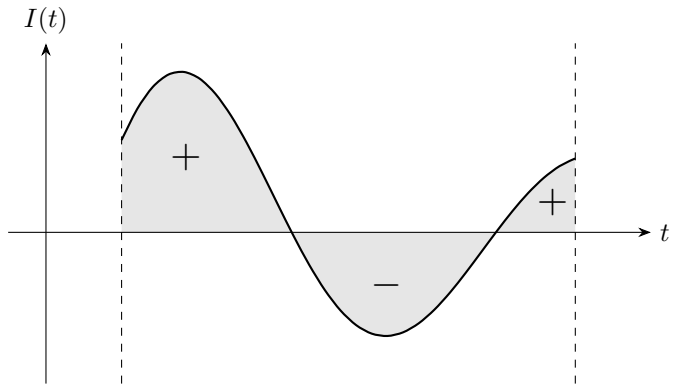


Figure 6.7: Positive and negative parts of the area "under" the curve.

This is closely related to the definition of multiplication as a repeated sum:

$$x \cdot y = \underbrace{y + y + \ldots + y}_{x \text{ terms}} \quad \text{where } y \in \mathbb{R},\ x \in \mathbb{N}$$

If y is varying with x, then instead of $y + y + \ldots + y$ we obtain $y_1 + y_2 + \ldots + y_x$. By generalizing from natural x to rational and further to real x, we essentially turn this sum into an integral $\int_0^x y(\xi)\,\mathrm{d}\xi$.

▷ Similarly to the differentiation, integration of vector-valued functions of a scalar argument can be still done according to the usual definition (6.17), effectively resulting in separate integration of each of the vector's components. E.g.

$$\int_{t_0}^{t} \mathbf{x}(\tau)\,\mathrm{d}\tau = \int_{t_0}^{t} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mathrm{d}\tau = \begin{bmatrix} \int_{t_0}^{t} x_1(\tau)\,\mathrm{d}\tau \\ \int_{t_0}^{t} x_2(\tau)\,\mathrm{d}\tau \end{bmatrix}$$

In the same way, complex-valued functions of a real argument can be integrated by integrating real and imaginary parts separately, which can be obtained not only by formal application of the definition of the Riemann integral, but also from the linearity of the integration with respect to the integrated function:

$$\int_{t_0}^{t} x(t)\,\mathrm{d}t = \int_{t_0}^{t} \big(\operatorname{Re} x(t) + i \operatorname{Im} x(t)\big)\,\mathrm{d}t = \int_{t_0}^{t} \operatorname{Re} x(t)\,\mathrm{d}t + i \int_{t_0}^{t} \operatorname{Im} x(t)\,\mathrm{d}t$$

Goal 6.2 completed. ■

6.3 Indefinite integral

6.3.1 Antiderivative

The formula (6.19) is the inverse of (6.9). The latter was allowing us to obtain $I(t)$ from $Q(t)$, the former allows us to obtain $Q(t)$ from $I(t)$.

Obviously, besides $I(t)$, the input data for (6.19) also contains the *initial value* $Q(t_0)$. From the perspective of (6.19) and (6.9) being mutually inverse, this extra piece of data is lost during differentiation in (6.9) (since adding or

removing a constant value to a function doesn't change its derivative), therefore we have to "manually" supply it back for the inverse operation.

But maybe sometimes we do not have this extra piece of data (or are not willing to supply it) and just want to obtain one (or all) of the possible functions $Q(t)$ such that $\frac{\mathrm{d}Q}{\mathrm{d}t} = I(t)$. This situation is captured in the idea of

(indefinite integral, a.k.a. antiderivative)

$$Q(t) = \int I(t)\,\mathrm{d}t \overset{\Delta}{\Longleftrightarrow} \frac{\mathrm{d}Q}{\mathrm{d}t} = I(t) \tag{6.23}$$

that is $Q(t)$ is an indefinite integral (antiderivative) of $I(t)$, if the derivative of $Q(t)$ is $I(t)$.

▷ Conversely, integrals with specified lower and upper integration limits are referred to as *definite integrals*.

▷ The notation $\int I(t)\,\mathrm{d}t$ is formally referring to the entire family of functions, which, upon differentiation, produce $I(t)$. Respectively, some care must be taken equating functions to indefinite integral. Normally such equalities require a free constant term to be present, e.g.

$$\int \cos t\,\mathrm{d}t = \sin t + C$$

but sometimes (somewhat informally) this term can be omitted, if we want to state that a particular function is an antiderivative of another function. E.g.

$$\int \cos t\,\mathrm{d}t = \sin t$$

can be understood as a statement that $\sin t$ is one of the antiderivatives of $\cos t$.

Thus, intuitively, the indefinite integral $\int Q(t)\,\mathrm{d}t$ can be thought of as a definite integral with unspecified initial time point t_0, as well as unspecified initial value $Q(t_0)$:

$$\int I(t)\,\mathrm{d}t = Q(t_0) + \int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau \quad \text{(where } t_0,\ Q(t_0) \text{ are arbitrary)}$$

It might seem that thereby the indefinite integral has two freedom degrees corresponding to the two free parameters t_0, and $Q(t_0)$, however we can quickly realize that t_0 doesn't contribute another freedom degree, its contribution being completely absorbed by $Q(t_0)$:

$$\begin{aligned} Q(t_1) + \int_{t_1}^{t} I(\tau)\,\mathrm{d}\tau &= Q(t_1) + \int_{t_1}^{t_0} I(\tau)\,\mathrm{d}\tau + \int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau \\ &= \left(Q(t_1) + \int_{t_1}^{t_0} I(\tau)\,\mathrm{d}\tau\right) + \int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau = Q(t_0) + \int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau \end{aligned}$$

Conversely, the contribution of $Q(t_0)$ can be absorbed by t_0, but sometimes only to a limited extent. E.g. consider the expression

$$10 + \int_{0}^{t} \cos t\,\mathrm{d}t$$

The antiderivative of cosine is sine. Therefore by shifting the starting point of the integration we can at most shift the value of the definite integral by ± 2 (in our case actually at most by ± 1, due to the integration starting at zero), therefore there is no way we can absorb a constant as large as 10 by shifting the integration's starting point.

Therefore, from the two freedom degrees we only need the one provided by $Q(t_0)$, which simply becomes a free additive constant, commonly notated as C:

$$\int I(t)\,\mathrm{d}t = \int_{t_0}^{t} I(\tau)\,\mathrm{d}\tau + C \tag{6.24}$$

6.3.2 Fundamental theorem of calculus

Goal 6.3. Express the definite Riemann integral in terms of antiderivative. *Hint:* find the specific value of the free additive constant C in (6.24) implied by a particular choice of the antiderivative function.

So let $f(t)$ be some (integrable) function and let

$$F(t) = \int f(t)\,\mathrm{d}t$$

be some antiderivative of $f(t)$, so that

$$F(t) = \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau + C$$

where we do not know or care about the value of C.

By formally letting $t = t_0$ we obtain

$$F(t_0) = \int_{t_0}^{t_0} f(\tau)\,\mathrm{d}\tau + C = 0 + C$$

therefore

$$C = F(t_0)$$

and

$$F(t) = \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau + C = \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau + F(t_0)$$

which is actually pretty much the formula (6.19). Subtracting $F(t_0)$ from both sides immediately yields

$$F(t) - F(t_0) = \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau$$

which, after a couple of notational changes, becomes

(fundamental theorem of calculus, a.k.a. Newton–Leibniz formula)

$$\int_{a}^{b} f(t)\,\mathrm{d}t = F(b) - F(a) \tag{6.25}$$

Somewhat remarkably, the existence of one side of (6.25) doesn't automatically imply the existence of the other side. That is, we can have a Riemann-integrable $f(t)$ which doesn't have an antiderivative $F(t)$ (in the strict sense of (6.23), which requires $f(t) = F'(t)$), and conversely we can have a differentiable $F(t)$ such that $f(t) = F'(t)$ is not Riemann integrable. The latter situation is rather an edge case, but the former is not.

Goal 6.3 completed. ■

6.3.3 Differentiability of integrals

We have just been suggesting one possible intuitive understanding of an indefinite integral as

$$\int f(t)\,\mathrm{d}t = \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau + C$$

which, by (6.23) means

$$\begin{aligned} f(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int f(t)\,\mathrm{d}t = \frac{\mathrm{d}}{\mathrm{d}t}\left(\int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau + C\right) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau + 0 \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau \end{aligned}$$

that is

(differentiation of a definite integral with respect to a varying upper limit)

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau = f(t) \tag{6.26}$$

Since differentiation is the inverse of integration, it might seem that (6.26) should hold all the time, but it turns out that it's not really the case. We are now about to construct a rather typical case where (6.26) fails.

In our pipe/reservoir setup we have been integrating the water flux $I(t)$ to obtain the amount of water $Q(t)$. It would be reasonable to assume that $I(t)$ is generally a continuous function, but it may also have jumps at few isolated points, where we suddenly open the valve on the pipe a bit more or a bit less.

> ▷ Of course it would be more realistic to assume that opening or closing the valve is not an instantaneous event, but takes a certain (however short) finite amount of time. But we could still use instantaneous jumps to represent idealized quick changes of the function's value. It is only essential (although not fully necessary) that there is a finite distance between each pair of neighboring jumps.

Intuitively, if $I(t)$ is fully continuous of $[a, b]$, the Riemann integral of $I(t)$ over $[a, b]$ should converge. Now, let's allow $I(t)$ to have a single jump at some point $c \in [a, b]$ (Fig. 6.8). We want to find out if having this jump affects the convergence of the Riemann sums. Intuitively, it shouldn't, since the jump is "infinitely narrow", therefore, as the partition norm decreases, the contribution of the partition's interval containing this jump will be smaller and smaller, eventually decreasing to zero.

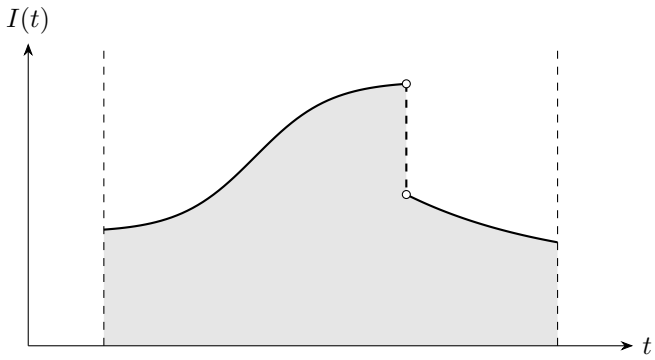


Figure 6.8: A continuous function with a jump and the area under the function's graph.

▷ On the other hand, a function having a jump should obviously result in a different value of the integral, since the function values before or after the jump (or both before and after) should be different from the case where the function is continuous at c. This might create some doubts about our argument that the jump doesn't affect the convergence, since at least it affects the value to which the Riemann sums converge. To address that doubt, at least on the intuitive level, we could split the integral over $[a, b]$ into two integrals:

$$\int_a^b I(t)\,\mathrm{d}t = \int_a^c I(t)\,\mathrm{d}t + \int_c^b I(t)\,\mathrm{d}t \tag{6.27}$$

Having a jump at c, the function $I(t)$ will also contain a jump at least on one of the intervals $[a, c]$ or $[b, c]$ (or on both if $I(c-0) \neq I(c) \neq I(c+0)$). However, since the jump is now contained at the interval's boundary, it doesn't really affect the value of the respective integral either.

The latter argument actually has value for intuitive understanding, but maybe not so much from a proof perspective. Strictly speaking, we didn't demonstrate that the equality (6.27) holds in case of a jump at c (or that it holds at all, for that matter). The earlier argument about a jump being "infinitely narrow" and thereby not contributing to the integral's value is actually a more rigorous one.

We can apply the same argument to the case where $I(t)$ has multiple jumps on $[a, b]$, as long as there is a finite number of those, obtaining

A sufficient condition for Riemann integral convergence: a Riemann integral of a bounded function, which has no more than a finite number of jump discontinuities, converges.

▷ Rigorously speaking, Riemann integral (of a bounded function) on $[a, b]$ converges if and only if the function is continuous almost everywhere on $[a, b]$ (that is everywhere except possibly a set of *measure zero*). Thus an infinite number of jumps (and/or "outliers", that is when $f(t+0) = f(t-0)$ but $f(t)$ has a different value) is possible as well, as long as the set of points, where the function jumps, is of measure zero.

Since differentiation and integration are mutually inverse operations (in the

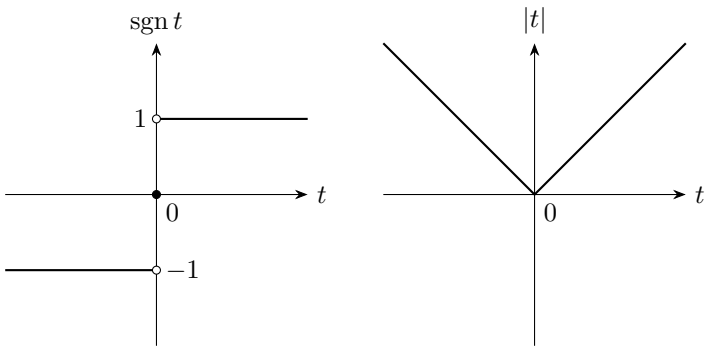


Figure 6.9: $\operatorname{sgn} t$ and its integral $|t|$.

sense of (6.23)), we would expect that for any integrable $I(t)$, the obtained function $Q(t)$ is differentiable. This however holds only for continuous $I(t)$. As soon as $I(t)$ contains a single discontinuous jump, $Q(t)$ is no longer differentiable. Indeed, let

$$I(t) = \operatorname{sgn} t = \begin{cases} -1 & \text{if } t < 0 \\ 0 & \text{if } t = 0 \\ 1 & \text{if } t = 0 \end{cases}$$

It is not difficult to realize that upon integration this yields $|t|$, in particular

$$Q(t) = \int_0^t I(\tau)\,\mathrm{d}\tau = \int_0^t \operatorname{sgn} \tau\,\mathrm{d}\tau = |t|$$

(where we formally allow t to also take negative values). Fig. 6.9 illustrates. We also have

$$\frac{\mathrm{d}}{\mathrm{d}t} Q(t) = \frac{\mathrm{d}}{\mathrm{d}t}|t| = \operatorname{sgn} t \quad \forall t \neq 0$$

however $Q(t) = |t|$ is not differentiable at $t = 0$, as the bilateral limit in (6.8) does not converge. The one-sided limits at $\Delta t \to +0$ and $\Delta t \to -0$, actually, do converge to $I(+0) = 1$ and $I(-0) = -1$ respectively. But, as we do not want to prefer future or the past in the definition of the derivative, the function $Q(t)$ is formally not differentiable at $t = 0$, thus the function $\tilde{I}(t) = \mathrm{d}Q/\mathrm{d}t$ is not defined at $t = 0$, while everywhere else we have $\tilde{I}(t) = I(t)\ \forall t \neq 0$.

The function $\tilde{I}(t)$, can be made integrable again by "manually" defining $\tilde{I}(0) = 0$, so that $\tilde{I}(t) = I(t)\ \forall t$. However, how would we know what $\tilde{I}(0)$ should be defined to, if we don't have any prior knowledge of the value $I(0)$, and are only given the function $Q(t)$? To answer this question, we can notice that *the specific value of* $\tilde{I}(0)$ *does not matter*, as long as $\tilde{I}(0)$ has some *finite* value. The argument is very similar to how we argued the convergence of a Riemann integral, despite the function having a jump. Similarly to how the behavior of the function at a jump point didn't contribute to the convergence since jump points are "infinitely narrow", the value of $\tilde{I}(0)$ doesn't contribute at all to the value of the integral, since this point is "infinitely narrow" (the respective term in the Riemann sum will decrease to zero, as the partition norm decreases to zero).

We can generalize the same argument to multiple points as

Independence of Riemann integral from a finite subset of function's values: the value of a Riemann integral is unaffected by changing the function's values on a finite set of points.

In that sense we can allow the function to have undefined values at a finite set of points, as long as we can treat these undefined values as finite ones.

▷ More precisely, similarly to the convergence of a Riemann integral, the value of a Riemann integral is unaffected by values of the integrated function on any *measure zero* subset of the integration interval.

With such amendments, the formula (6.26) doesn't exactly become applicable in case of a discontinuous $f(t)$. E.g. we cannot equate an undefined $\frac{\mathrm{d}}{\mathrm{d}t}|t|$ to $\operatorname{sgn} t$ at $t = 0$. But the converse formula (integration of a derivative) becomes applicable:

$$\int_{t_0}^{t} F'(\tau)\,\mathrm{d}\tau = F(t) + C = F(t) - F(t_0)$$

where we can notice that essentially it is the fundamental theorem of calculus (6.25). E.g.

$$\int_{t_0}^{t} \operatorname{sgn} t\,\mathrm{d}t = |t| - |t_0|$$

even though $F(t) = |t|$ is not an antiderivative of $f(t) = \operatorname{sgn} t$ in a fully strict sense of (6.23).

6.4 Riemann–Stieltjes integral

Suppose a force $\vec{F}$ is applied to a solid body, which, under this force, experiences a displacement $\Delta\vec{r}$ (where $\vec{r}$ is the body position). The force thereby performs the work $W = \vec{F} \cdot \Delta\vec{r}$. Now, the latter formula is applicable if $\vec{F}$ is constant. However what if $\vec{F}$ is varying over time: $\vec{F} = \vec{F}(t)$?

A similar problem occurs if the displacement $\Delta\vec{r}$ is the result of a non-uniform motion (in particular, if the trajectory in not linear), although in that case $W = \vec{F} \cdot \Delta\vec{r}$ *happens* to deliver a correct answer if $\vec{F}$ is constant. Fundamentally, however, the formula $W = \vec{F} \cdot \Delta\vec{r}$ is supposed to work "for granted" only in case of uniform (or, at least, maybe linear) motion.

Goal 6.4. Extend the previously developed Riemann integration framework to accommodate the case of computing the work of a varying force applied to a non-uniformly moving body. Stick to the one-dimensional case for simplicity. *Hint:* adjust the Riemann sums to accommodate the case $\vec{F} \cdot \Delta\vec{r}$, where the integrated function $\vec{F}(t)$ and the integration variable $\vec{r}(t)$ are both functions of yet another variable t.

Apparently, in this case we could split the time interval in question (say, it's $[t_0, t]$) into smaller subintervals $[t_n, t_{n+1}]$. If the subintervals are small enough, we can approximate the work on each of these subintervals as

$$\widetilde{\Delta W_n} = \vec{F}(\tau_n) \cdot \Delta\vec{r}_n$$

where $\tau_n \in [t_n, t_{n+1}]$ and $\Delta\vec{r}_n = \vec{r}(t_{n+1}) - \vec{r}(t_n)$, so that

$$W \approx \sum_{n=0}^{N-1} \widetilde{\Delta W_n} = \sum_{n=0}^{N-1} \left(\vec{F}(\tau_n) \cdot \Delta\vec{r}_n\right)$$

At this point, for simplicity of the further argument, let's assume a one-dimensional case, so that we replace $\vec{F}$ by its projection F onto the only coordinate axis x (note that thereby F stands for the signed projection value and not for the absolute magnitude $|\vec{F}|$), and we replace $\vec{r}$ simply by x:

$$W \approx \sum_{n=0}^{N-1} \widetilde{\Delta W_n} = \sum_{n=0}^{N-1} F(\tau_n)\Delta x_n \tag{6.28}$$

where $\Delta x_n = x(t_{n+1}) - x(t_n)$.

We now wish to take the limit of the right-hand side of (6.28), expecting to thereby obtain an exact value of W. Now this right-hand side looks almost identical to the right-hand side of the definition of the Riemann integral (6.17), except that this time we have Δx_n in place of Δt_n. The expression obtained in the limit is referred to as

(Riemann–Stieltjes integral)

$$\int_{\tau=t_0}^{t} F(\tau)\,\mathrm{d}x(\tau) = \lim_{\Delta\to 0} \sum_{n=0}^{N-1} F(\tau_n)\Delta x_n \tag{6.29}$$

where the partitioning is still done with respect to t, but the integration is done over x. At $x(t) \equiv t$ the Riemann–Stieltjes integral turns into a Riemann integral.

▷ The sum under the limit sign in the right-hand side of (6.29) is referred to as *Riemann–Stieltjes sum.*

Goal 6.4 completed. ■

Even though by construction we obtained a Riemann–Stieltjes integral as a solution for our problem of computing the work, the answer can be reexpressed in terms of a Riemann integral.

Goal 6.5. Reexpress Riemann-Stieltjes integral in terms of Riemann integral. *Hint:* express Δx_n in terms of Δt_n.

Let $x'(t)$ be the body's velocity. In that case Δx_n can be approximated as

$$\Delta x_n \approx x'(\tau_n)\Delta t_n$$

These approximations become exact at $\Delta t_n \to 0$ and therefore the right-hand side of (6.29) can be rewritten as

$$\lim_{\Delta\to 0} \sum_{n=0}^{N-1} F(\tau_n)x'(\tau_n)\Delta t_n = \int_{t_0}^{t} F(\tau)x'(\tau)\,\mathrm{d}\tau$$

It follows that

(Riemann–Stieltjes integral expressed in terms of Riemann integral)

$$\int_{\tau=t_0}^{t} F(\tau)\,\mathrm{d}x(\tau) = \int_{t_0}^{t} F(\tau)x'(\tau)\,\mathrm{d}\tau \tag{6.30}$$

The formula (6.30) works for all "reasonable" $x(t)$, but may fail in more exotic cases. E.g. let $x(t)$ be the (non-centered) Heaviside step function

$$x(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{1}{2} & \text{if } t = 0 \\ 1 & \text{if } t > 0 \end{cases}$$

which contains a jump from 0 to 1 at $t = 0$. Assuming $F(t)$ is continuous at $t = 0$, the Riemann–Stieltjes integral (6.29) will converge to $F(0)$, while the Riemann integral in the right-hand side of (6.30) is even not defined, since $x'(0)$ does not exist.

▷ The integral in the right-hand side of (6.30) can be formally made computable for an $x(t)$ containing a discontinuous step by using *Dirac delta* function introduced in Section 8.13, although in that case it is no longer a Riemann integral in the strict original sense.

▷ The expression of the same form as the one in the right-hand side of (6.30) is encountered twice in the integration by parts formula

$$\int_{t_0}^{t} f(\tau)g'(\tau)\,\mathrm{d}t = f(\tau)g(\tau)\Big|_{\tau=t_0}^{t} - \int_{t_0}^{t} g(\tau)f'(\tau)\,\mathrm{d}\tau$$

which makes this formula look very elegant in terms of Riemann–Stieltjes integrals:

$$\int_{\tau=t_0}^{t} f(\tau)\,\mathrm{d}g(\tau) = f(\tau)g(\tau)\Big|_{\tau=t_0}^{t} - \int_{\tau=t_0}^{t} g(\tau)\,\mathrm{d}f(\tau)$$

which looks even more elegant by omitting the integration limits, yielding the indefinite Riemann–Stieltjes integrals:

$$\int f(\tau)\,\mathrm{d}g(\tau) = f(\tau)g(\tau) - \int g(\tau)\,\mathrm{d}f(\tau)$$

or, by also omitting the explicit notation of the functions' arguments:

(integration by parts formula in terms of indefinite Riemann–Stieltjes integrals)

$$\int f\,\mathrm{d}g = fg - \int g\,\mathrm{d}f \tag{6.31}$$

▷ The formula (6.30) provides the basis for integration variable replacement in a Riemann integral. Suppose $x(t)$ is a monotone function on $[t_0, t]$. Then the Riemann–Stieltjes integral in the left-hand side of (6.30) can be written as the Riemann integral with respect to x instead:

$$\int_{\tau=t_0}^{t} F(\tau)\,\mathrm{d}x(\tau) = \int_{x_0}^{x_1} F(\tau(x))\,\mathrm{d}x$$

where $x_0 = x(t_0)$, $x_1 = x(t)$ and $\tau(x)$ is the function inverse of $x(\tau)$. It follows that

$$\int_{x_0}^{x_1} F(\tau(x))\,\mathrm{d}x = \int_{\tau=t_0}^{t} F(\tau)\,\mathrm{d}x(\tau) = \int_{t_0}^{t} F(\tau)x'(\tau)\,\mathrm{d}\tau$$

or

$$\int_{x_0}^{x_1} \Phi(x)\,\mathrm{d}x = \int_{t_0}^{t} \Phi(x(\tau))x'(\tau)\,\mathrm{d}\tau$$

where $\Phi(x) = F(\tau(x))$.

Goal 6.5 completed. ■

6.5 Differentials

The symbols $\mathrm{d}t$, $\mathrm{d}f$ etc. encountered in the notation of derivatives (e.g. $\mathrm{d}f/\mathrm{d}t$) or integrals (e.g. $\int f(t)\,\mathrm{d}t$) are referred to as *differentials*. So $\mathrm{d}t$ is referred to as a *differential of* t, $\mathrm{d}f$ as a differential of f, etc. The differentials are emerging from the respective finite differences Δt, Δf etc., and therefore can be intuitively understood as infinitely small versions of the same differences. We cannot however freely operate on infinitely small values in the same way how we operate on finite values, since, technically, infinitely small values are considered equal to zero and normally need to be treated as such. In order to be considered different from zero, infinitely small values need to appear in the context of limits. Therefore differentials can appear only in the context of limits implied by the context. E.g. in $\mathrm{d}f/\mathrm{d}t$ the implied context is $\Delta t \to 0$, in $\int f(t)\,\mathrm{d}t$ the implied context is $\max|\Delta t_n| \to 0$ etc.

6.5.1 Differentials in differentiation context

Having agreed to implicitly understand the existence of such context, we can start performing various operations on differentials, as if they were ordinary values. Starting with

$$\frac{\mathrm{d}f}{\mathrm{d}t} = f'(t)$$

(where the implied context is $\Delta t \to 0$, which we in this case can also somewhat informally write as $\mathrm{d}t \to 0$), we can multiply both sides by $\mathrm{d}t$ obtaining

$$\mathrm{d}f = f'(t)\,\mathrm{d}t \tag{6.32a}$$

which for the time being can be used as the definition of the differential $\mathrm{d}f$ of a function f. The equality (6.32a) should be understood as a limiting version of the approximate equality

$$\Delta f \approx f'(t)\Delta t \tag{6.32b}$$

which in the limit turns into the strict equality (6.32a), the symbols Δf and Δt being replaced by $\mathrm{d}f$ and $\mathrm{d}t$. Notice that technically Δt is a free independent variable (since the context is defined by letting Δt approach zero), while Δf depends on both the point of interest t (at which we compute $f'(t)$) as well as on the value of Δt, which can be reflected by explicitly notating Δf as a function of these two values: $\Delta f(t, \Delta t)$. We can thereby more explicitly and precisely rewrite (6.32b) as

$$\Delta f(t, \Delta t) = f'(t)\Delta t + o(\Delta t) \tag{6.32c}$$

where $o(\Delta t)$ is the "error term" appearing for finitely large Δt. The "error term" $o(\Delta t)$ thus accommodates all higher-order infinitesimals of the right-hand side of (6.32c).

Similarly to how we obtained (6.32a) from $\mathrm{d}f/\mathrm{d}t = f'(t)$, the "derivative of the sum" formula

$$\frac{\mathrm{d}(f+g)}{\mathrm{d}t} = (f+g)' = f' + g' = \frac{\mathrm{d}f}{\mathrm{d}t} + \frac{\mathrm{d}g}{\mathrm{d}t}$$

by multiplying both sides by $\mathrm{d}t$ turns into

(differential of a sum is equal to the sum of differentials)

$$\mathrm{d}(f+g) = \mathrm{d}f + \mathrm{d}g \tag{6.33}$$

where the context is again given by $\Delta t \to 0$.

Similarly, the "derivative of the product"

$$\frac{\mathrm{d}(fg)}{\mathrm{d}t} = (fg)' = fg' + gf' = f\frac{\mathrm{d}g}{\mathrm{d}t} + g\frac{\mathrm{d}f}{\mathrm{d}t}$$

turns into

(differential of a product)

$$\mathrm{d}(fg) = f\,\mathrm{d}g + g\,\mathrm{d}f \tag{6.34}$$

etc.

6.5.2 Differentials in integration context

Equality (6.32a) suggests the possibility to write the (definite or indefinite) integrals of the form $\int \mathrm{d}f$ which are thereby understood as

$$\int_a^b \mathrm{d}f \overset{(6.32a)}{=} \int_a^b f'(t)\,\mathrm{d}t \overset{(6.25)}{=} f(b) - f(a) \tag{6.35a}$$

Actually, the left-hand side of (6.35a) can be interpreted in the sense of the Riemann–Stieltjes integral

$$\int_a^b \mathrm{d}f = \int_a^b 1\,\mathrm{d}f = \int_a^b 1(t)\,\mathrm{d}f(t) \overset{(6.29)}{=} f(b) - f(a) \tag{6.35b}$$

where $1 = 1(t)$ stands for a function equal to 1 everywhere. The conclusion of (6.35a) thereby still holds even in cases when $f(t)$ has jump-type discontinuities and $f'(t)$ doesn't formally exist. This suggests that the idea of a differential $\mathrm{d}f$ is not restricted to the cases of differentiable functions (where $\mathrm{d}f$ are infinitely small), but also applies to functions with jump discontinuities (where $\mathrm{d}f$ can be finitely large, since Δf_n in the respective Riemann–Stieltjes sum might never go to zero).

We thus obtain a general formula

(Riemann–Stieltjes integral without an integrated function)

$$\int \mathrm{d}f = f(t) + C \tag{6.36}$$

▷ Notice that (6.34) after applying integration to both sides turns into the Riemann–Stieltjes integral version of the integration by parts formula:

$$fg \overset{(6.36)}{=} \int \mathrm{d}(fg) = \int f\,\mathrm{d}g + \int g\,\mathrm{d}f$$

So, one doesn't really need to remember the integration by parts formula, it suffices to remember (6.34) and (6.31) follows trivially.

6.5.3 Formal definition

So far we have used (6.32a) as a somewhat intuitive definition of the differential $\mathrm{d}f$. Let's try to develop something more formal.

A more precise version of (6.32a) is given by (6.32c), which we could therefore use to construct a more formal definition of the differential. The symbol Δf appearing in the left-hand side of (6.32c) is turned into $\mathrm{d}f$ by discarding the "error term" $o(\Delta t)$ in the right-hand side of (6.32c), and simultaneously we need to re-notate Δt as $\mathrm{d}t$. So the symbols $\mathrm{d}f$ and $\mathrm{d}t$ seem to emerge in two different ways:

- The differential $\mathrm{d}t$ of the *free variable* t of the implied context $\Delta t \to 0$ is simply understood as Δt.
- The differential $\mathrm{d}f$ of a function $f(t)$ of the free variable t of the implied context $\Delta t \to 0$ is understood as the principal part $f'(t)\Delta t$ (excluding the "error term" $o(\Delta t)$) of the right-hand side of (6.32c).

Since $\Delta t = \mathrm{d}t$ by definition, it follows that the principal part $f'(t)\Delta t$ of the right-hand side of (6.32c) can be rewritten as $f'(t)\,\mathrm{d}t$, and we formally arrive at (6.32a): $\mathrm{d}f = f'(t)\,\mathrm{d}t$. The differential $\mathrm{d}f$ thus depends on the two variables t and $\mathrm{d}t$, thereby formally making it a function of these two variables:

(differential of a differentiable function)

$$\mathrm{d}f(t, \mathrm{d}t) = f'(t)\,\mathrm{d}t \tag{6.37}$$

On the other hand, we mentioned that equation (6.35b) implicitly extends the idea of a differential of a differentiable function to functions containing jump discontinuities. At a jump discontinuity point of a function f, (6.32c) turns into

$$\Delta f(t, \Delta t) = \begin{cases} \big(f(t+0) - f(t)\big) + f'(t+0)\,\Delta t + o(\Delta t) & \text{if } \Delta t > 0 \\ \big(f(t-0) - f(t)\big) + f'(t-0)\,\Delta t + o(\Delta t) & \text{if } \Delta t < 0 \end{cases}$$

(where Δt cannot be equal to zero under the implied context $\Delta t \to 0$, because taking the limit assumes $\Delta t \neq 0$). Respectively the principal part of the right-hand side turns into

(differential of a jump-discontinuous function)

$$\mathrm{d}f(t,\mathrm{d}t) = \begin{cases} \big(f(t+0)-f(t)\big) + f'(t+0)\,\mathrm{d}t & \text{if } \mathrm{d}t > 0 \\ \big(f(t-0)-f(t)\big) + f'(t-0)\,\mathrm{d}t & \text{if } \mathrm{d}t < 0 \end{cases} \tag{6.38}$$

Notice that if f is differentiable at t, (6.38) coincides with (6.37).

So far we have introduced two ideas of a differential: one applicable to functions, one to the free independent variable. We can merge both ideas together by simply defining the differential of a variable y as the principal part of the difference $\Delta y(t,\Delta t)$, where t is the free variable implied by the context $\Delta t \to 0$. The variable y can depend on t, in which case we can simultaneously understand y as a function of t (notationally expressed as $y = y(t)$), in which case (6.37) or (6.38) applies, e.g. by (6.37) we would have $\mathrm{d}y(t,\mathrm{d}t) = y'\,\mathrm{d}t$. Or, if we need the differential of the free variable t itself, we can consider t to be a function of itself ($t = t(t)$). This function is differentiable at all points, its derivative being equal to 1 everywhere, so (6.37) applies, formally yielding

(differential of the free variable of the implied context)

$$\mathrm{d}t(t,\mathrm{d}t) = t'(t)\,\mathrm{d}t = 1\cdot\mathrm{d}t = \mathrm{d}t$$

If the variable y is not dependent on t, nor does it coincide with t, effectively it means that y behaves as a constant function in the context of a varying t: $y(t) = \text{const}$. We can therefore still apply (6.37), obtaining

(differential of an independent variable)

$$\mathrm{d}y(t,\mathrm{d}t) = y'(t)\,\mathrm{d}t = 0\cdot\mathrm{d}t \equiv \mathrm{d}t$$

We thus obtain a universal formal definition of the differential:

Formal definition of a differential of a variable: the differential $\mathrm{d}y(t,\mathrm{d}t)$ of y is the principal (linear or affine) part of $\Delta y(t,\Delta t)$:

$$\Delta y(t,\Delta t) = \mathrm{d}y(t,\Delta t) + o(\Delta t) \tag{6.39a}$$

where t is the free independent variable of the implied limiting context $\Delta t \to 0$.

By treating Δt as an infinitely small value $\Delta t = \mathrm{d}t$, we typically obtain infinitely small values of Δy, which by (6.39a) also implies infinitely small values of $\mathrm{d}y(t,\Delta t) = \mathrm{d}y(t,\mathrm{d}t)$. Somewhat less formally, we can rewrite (6.39a) as

Mostly formal definition of a differential of a variable: the differential $\mathrm{d}y(t,\mathrm{d}t)$ of y is the principal (linear or affine) part of $\Delta y(t,\mathrm{d}t)$:

$$\Delta y(t,\mathrm{d}t) = \mathrm{d}y(t,\mathrm{d}t) + o(\mathrm{d}t) \tag{6.39b}$$

where t is the free independent variable of the implied limiting context $\Delta t \to 0$, and $\mathrm{d}t$ is understood as an infinitely small value Δt.

6.5.4 Higher-order differentials

The differential symbol "d" itself can be understood as an operator, which takes a function of one variable and delivers a function of two variables. So e.g., being applied to a function $f(t)$, it turns the function $f(t)$ of one variable t into the function $\mathrm{d}f(t, \mathrm{d}t)$ of variables t and $\mathrm{d}t$.

Let's try applying the operator "d" two times in a row. That is, given the function $f(t)$, we apply the operator "d" once, obtaining $\mathrm{d}f = f' \,\mathrm{d}t$, and then apply it once again to the result of the first application:

$$\begin{aligned} \mathrm{d}^2 f = \mathrm{d}(\mathrm{d}f) = \mathrm{d}(f' \cdot \mathrm{d}t) &\overset{(6.34)}{=} \mathrm{d}(f') \cdot \mathrm{d}t + f' \cdot \mathrm{d}(\mathrm{d}t) = (f'' \,\mathrm{d}t) \cdot \mathrm{d}t + f' \cdot \mathrm{d}^2 t \\ &= f'' \,\mathrm{d}t^2 + f' \,\mathrm{d}^2 t \end{aligned} \tag{6.40}$$

Now $\mathrm{d}t$ doesn't depend on t, therefore

$$\mathrm{d}^2 t = \mathrm{d}(\mathrm{d}t) = (\mathrm{d}t)' \cdot \mathrm{d}t = 0 \cdot \mathrm{d}t = 0$$

and we obtain

(2nd-order differential of a differentiable function)

$$\mathrm{d}^2 f = f'' \,\mathrm{d}t^2 \tag{6.41}$$

justifying the notation $\mathrm{d}^2 f / \mathrm{d}t^2$ for the second derivative of f.

The derivation of (6.41) has been a bit nonchalant, let's dive into some details. Let's start with the fact that after the first application of the operator "d" to the function f we obtain a function of two variables $\mathrm{d}f(t, \mathrm{d}t)$. There is some formal difficulty applying the operator "d" once again, since the operator expects a function of one variable. However the operator "d" cares about the function's dependency on the free variable t, intuitively we just want to know how does $\mathrm{d}f(t, \mathrm{d}t)$ change when we go from the point at t to the point at $t + \mathrm{d}t$. From this perspective the second argument $\mathrm{d}t$ is just a fixed-value parameter and can be ignored. The multiplier $\mathrm{d}t$ is thereby treated as a constant with respect to t, therefore it simply can be factored out from under the derivative: $(\mathrm{d}f)' = (f' \cdot \mathrm{d}t)' = f'' \cdot \mathrm{d}t$, therefore $\mathrm{d}(\mathrm{d}f) = (\mathrm{d}f)' \cdot \mathrm{d}t = f'' \cdot \mathrm{d}t^2$.

Essentially we did the same in (6.40): we imagined that the second argument of $\mathrm{d}f(t, \mathrm{d}t)$ is formally a function of t, that is $\mathrm{d}t = \tau(t)$, so $\mathrm{d}f(t, \mathrm{d}t)$ can be considered as an (infinitesimally-valued) function of t as well: $\varphi(t) = \mathrm{d}f(t, \tau(t))$. We can apply the operator "d" to this function, obtaining

$$\mathrm{d}^2 f = \mathrm{d}\varphi(t, \mathrm{d}t) = \big(\mathrm{d}f(t, \tau(t))\big)' \cdot \mathrm{d}t$$

where we further have

$$\begin{aligned} \big(\mathrm{d}f(t, \tau(t))\big)' = \big(f'(t) \cdot \tau(t)\big)' &= f''(t) \cdot \tau(t) + f'(t) \cdot \tau'(t) \\ &= f''(t) \cdot \mathrm{d}t + f'(t) \cdot 0 = f''(t) \cdot \mathrm{d}t \end{aligned}$$

since, from the perspective of dependency of t, $\tau(t) = \text{const}$, and (6.41) follows.

Differentials of orders higher than 2 are defined by iteratively applying the same steps to the definition of the 2nd-order differential, obtaining

(n-th differential of a differentiable function)

$$\mathrm{d}^n f = f^{(n)}\,\mathrm{d}t^n \tag{6.42}$$

(where $f^{(n)}$ denotes the n-th derivative of f).

▷ Formula (6.42) suggests that Taylor series for an infinitely small argument offset $\mathrm{d}t$ can be written as

(Taylor series in terms of differentials)

$$f(t+\mathrm{d}t) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(t)\,\mathrm{d}t^n = \sum_{n=0}^{\infty} \frac{\mathrm{d}^n f}{n!} \tag{6.43}$$

6.5.5 Invariance of 1st-order differentials

We have been somewhat nonchalantly using 1st-order differentials almost as if they were ordinary variables in equations like (6.33) and (6.34), assuming that the limiting context is given by $\Delta t \to 0$ where t is the argument of both f and g. There is actually quite a large freedom in the understanding of the implied limiting context.

Suppose we're having several variables y_n, whose values are changing as

$$y_n = f_n(t) \tag{6.44a}$$

We thereby have

$$\frac{\mathrm{d}y_n}{\mathrm{d}t} = f_n'(t) \tag{6.44b}$$

It would be natural to assume $\Delta t \to 0$ as the limiting context, in which case t is the free variable. To make our discussion as generic as possible, we shall formally include both the case where the free variable t is different from all variables y_n, as well as the case where it coincides with one of y_n (e.g. $y_2 = f_2(t) = t$, therefore $f_2'(t) \equiv 1$). In either case we have

$$\mathrm{d}y_n = f_n'(t)\,\mathrm{d}t \tag{6.44c}$$

Instead of using t as the free variable, we could write y_n as functions of some other free variable τ, with the implied context $\Delta\tau \to 0$:

$$y_n = \varphi_n(\tau) \tag{6.45a}$$

$$\frac{\mathrm{d}y_n}{\mathrm{d}\tau} = \varphi'(\tau) \tag{6.45b}$$

$$\mathrm{d}y_n = \varphi_n'(\tau)\,\mathrm{d}\tau \tag{6.45c}$$

Again, τ can be different from all y_n, or it could coincide with one of y_n.

What is important, however, is that we assume that the descriptions (6.44) and (6.45) are equivalent in the sense that they describe the same set of dependencies between y_n. Specifically, if a particular set of values of y_n is attained at some particular value of t, there is a corresponding value of τ at which exactly

the same set of values y_n is attained. Furthermore, small variations in t should correspond to small variations of τ and vice versa. More precisely, t and τ must be related via some locally invertible continuous (actually, differentiable) function

$$t = g(\tau) \tag{6.46a}$$

and

$$y_n = f_n(t) = f_n(g(\tau)) = \varphi_n(\tau) \tag{6.46b}$$

We are thus having two different implied contexts $\Delta t \to 0$ and $\Delta\tau \to 0$ with their respective free variables t and τ, and we want to find out, how differentials $\mathrm{d}y_n$ change when switching from one context to the other.

By (6.46) we have

$$\mathrm{d}t = g'(\tau)\,\mathrm{d}\tau \tag{6.47a}$$

and

$$\mathrm{d}y_n = f_n'(t)\,\mathrm{d}t = \varphi_n'(\tau)\,\mathrm{d}\tau = f_n'(g(\tau))g'(\tau)\,\mathrm{d}\tau \tag{6.47b}$$

We can notice from (6.47b), that, if a particular value of $\mathrm{d}y_n$ is attained for a particular value of $\mathrm{d}t$, exactly the same value of $\mathrm{d}y_n$ would be attained for $\mathrm{d}\tau = \mathrm{d}t/g'(\tau)$, that is if $\mathrm{d}t$ and $\mathrm{d}\tau$ are related via (6.47a). And this happens simultaneously for all n, since (6.47a) is independent of n.

Furthermore, by (6.47b), all y_n are proportional to $\mathrm{d}t$ (the coefficients of the proportionality being $f_n'(t)$), but they are also proportional to $\mathrm{d}\tau$ (the coefficients of the proportionality being $\varphi_n'(\tau) = f_n'(g(\tau))g'(\tau)$). Thus, the ratios $\mathrm{d}y_n/\mathrm{d}y_m$ must be identical in both limiting contexts $\Delta t \to 0$ and $\Delta\tau \to 0$:

(invariance of 1st order differentials)

$$\frac{\mathrm{d}y_n(t,\mathrm{d}t)}{\mathrm{d}y_m(t,\mathrm{d}t)} = \frac{\mathrm{d}y_n(\tau,\mathrm{d}\tau)}{\mathrm{d}y_m(\tau,\mathrm{d}\tau)} \quad \forall \mathrm{d}t, \mathrm{d}\tau,\ t = g(\tau) \tag{6.48}$$

where t and τ are assumed to correspond to the same state (as described by (6.46a)).

This means that

Invariance of 1st order differentials (alternative form): all homogeneous equations with respect to 1st order differentials as the unknowns are invariant under a change of the implied limiting context.

Indeed, simultaneously multiplying all differentials in such equation by one and the same nonzero constant yields a fully equivalent equation. E.g., given a linear homogeneous equation $\mathrm{d}y = 2\,\mathrm{d}x + 3\,\mathrm{d}z$, we can multiply all differentials by 2 yielding an equivalent equation $2\,\mathrm{d}y = 4\,\mathrm{d}x + 6\,\mathrm{d}z$. Similarly, given $\mathrm{d}z^2 = 2\,\mathrm{d}x^2 + 3\,\mathrm{d}y^2$, multiplying all differential by 2 again yields an equivalent equation $4\,\mathrm{d}z^2 = 8\,\mathrm{d}x^2 + 12\,\mathrm{d}y^2$. Etc. On the other hand, an equation, relating 1st order differentials to each other but being not homogeneous with respect to those,

would be highly suspicious. We should only "add or compare apples to apples and oranges to oranges", adding or comparing a 1st order differential to anything but another 1st order differential should raise some questions.[4]

▷ Notice that the rules (6.33) and (6.34) are homogeneous linear equations with respect to the differentials.

The property (6.48) might look trivial, but actually it's not. In particular, the same property doesn't hold for higher-order differentials. E.g. let $y = x^2$ and consider the point $x = 1$, $y = 1$. Treating x as the free variable we obtain at this point

$$\begin{aligned} \mathrm{d}y &= y'(x)\,\mathrm{d}x = 2x\,\mathrm{d}x = 2\,\mathrm{d}x \\ \mathrm{d}^2y &= y''(x)\,\mathrm{d}x^2 = 2\,\mathrm{d}x^2 \\ \mathrm{d}^2x &= x''(x)\,\mathrm{d}x^2 = 0\,\mathrm{d}x^2 = 0 \end{aligned}$$

On the other hand, treating y as the free variable, we have $x = \sqrt{y}$ and

$$\begin{aligned} \mathrm{d}x &= x'(y)\,\mathrm{d}y = \frac{1}{2\sqrt{y}}\,\mathrm{d}y = \frac{1}{2}\,\mathrm{d}y \\ \mathrm{d}^2x &= x''(y)\,\mathrm{d}y^2 = -\frac{1}{4y^{3/2}}\,\mathrm{d}y^2 = -\frac{1}{4}\,\mathrm{d}y^2 = -\left(\frac{1}{2}\mathrm{d}y\right)^2 = -\,\mathrm{d}x^2 \\ \mathrm{d}^2y &= y''(y)\,\mathrm{d}y^2 = 0\,\mathrm{d}y^2 = 0 \end{aligned}$$

We can see that, regardless of which variable is the free one, $\mathrm{d}y$ is always twice as large as $\mathrm{d}x$. However d^2x and d^2y, depending on the choice of the free variable, take totally different values. In particular, if x is the free variable, then $\mathrm{d}^2x = 0$, while if y is the free variable, then $\mathrm{d}^2x = -\,\mathrm{d}x^2 \neq 0$. A similar situation occurs for d^2y.

The invariance of 1st-order differentials is directly related to the chain rule for differentiation (which is what we applied in the derivation of (6.48), specifically in (6.47b)). We can now do the reverse and formally reobtain the chain rule by manipulating the differentials. Given $x = g(t)$ and $y = f(x)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(g(t)) = \frac{\mathrm{d}y}{\mathrm{d}t}$$

Multiplying both the numerator and the denominator by $\mathrm{d}x$,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(g(t)) = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t}$$

Now $\mathrm{d}y/\mathrm{d}x$ is f' evaluated at $x = g(t)$, while $\mathrm{d}x/\mathrm{d}t = g'(t)$ and we obtain

[4] One common exception to that is comparing a natural power of a 1st-order differential to a higher-order differential (e.g. $\mathrm{d}^2y = 2\,\mathrm{d}t^2$), but higher-order differentials are not invariant under a change of the implied limiting context anyway, as we are about to find out.

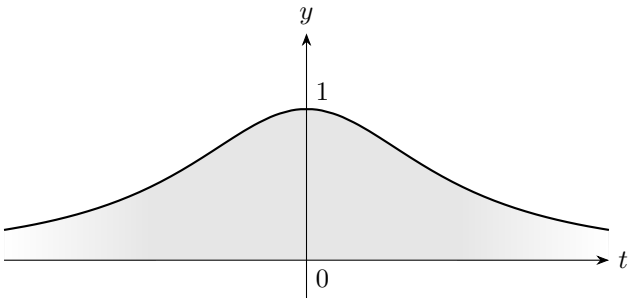


Figure 6.10: The graph of the function $y = 1/(1+t^2)$ is infinitely long, but the area under the graph is finite.

(chain rule for differentiation)

$$\frac{\mathrm{d}}{\mathrm{d}t} f(g(t)) = \frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t} = f'(g(t)) \cdot g'(t) \tag{6.49}$$

where $y = f(x)$, $x = g(t)$.

6.6 Improper integrals

Until now we have been implicitly assuming that the integration limits are taking finite values. However, suppose we are having an infinitely long curve, e.g. $f(t) = 1/(1+t^2)$ defined on $t \in (-\infty, +\infty)$ and want to compute the area under the entire curve? Our first thought could be that an infinitely long area would be infinitely large, but that's not necessarily the case. E.g. the antiderivative of $f(t) = 1/(1+t^2)$ is $F(t) = \arctan t + C$, so

$$\int_a^b \frac{\mathrm{d}t}{1+t^2} = \arctan b - \arctan a \tag{6.50a}$$

and it kind of follows that

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}t}{1+t^2} = \arctan(+\infty) - \arctan(-\infty) = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi \tag{6.50b}$$

Fig. 6.10 illustrates.

This motivates the introduction of the idea of

(improper integral on an open interval (a, b)***)***

$$\int_a^b f(t)\,\mathrm{d}t = \lim_{\substack{\alpha \to a \\ \beta \to b}} \int_\alpha^\beta f(t)\,\mathrm{d}t \tag{6.51a}$$

We could also have an integral which is improper at only only one of the two integration limits:

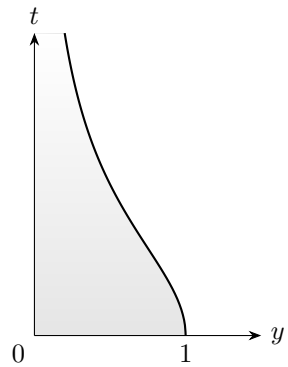


Figure 6.11: The graph of the function $t = \sqrt{1/y - 1}$ is infinitely high, but the area under the graph is finite and is equal to half of the area under the graph in Fig. 6.10.

(improper integral on a semiopen interval $[a, b)$***)***

$$\int_a^b f(t)\,\mathrm{d}t = \lim_{\beta \to b} \int_a^\beta f(t)\,\mathrm{d}t \tag{6.51b}$$

or

(improper integral on a semiopen interval $(a, b]$***)***

$$\int_a^b f(t)\,\mathrm{d}t = \lim_{\alpha \to b} \int_\alpha^b f(t)\,\mathrm{d}t \tag{6.51c}$$

Replacing the function $y = 1/(1 + t^2)$ from (6.50b) with its inverse:

$$t = \sqrt{\frac{1}{y} - 1} \tag{6.52}$$

we thereby mirror the graph of the function with respect to the line $y = t$. Upon mirroring, the area under the graph, however, stays the same. Therefore the area under the graph of (6.52) is equal to one half of the area under the graph of $y = 1/(1+t^2)$, where we need to halve the result of (6.50b), since only one half of the area from Fig. 6.10 is left after the function inversion. Fig. 6.11 illustrates.

We therefore expect to have:

$$\int_0^1 \sqrt{\frac{1}{y} - 1}\,\mathrm{d}y = \frac{\pi}{2} \tag{6.53}$$

This motivates the application of the formulas (6.51) not only to infinite integration limits but also to finite limits at which the function value approaches infinity.

▷ Apparently if both the integration limit and the function value are infinitely large, the integral diverges anyway (possibly unless evaluated in the Cauchy principal value sense), so we usually consider either the integration limit being infinite or the function value being infinite at the integration limit, but not both simultaneously.

▷ There is no difference in the notation between a usual proper integral taken over a closed interval $[a, b]$ or an improper integral taken over an open interval (a, b) or a semiopen interval $[a, b)$ or $(a, b]$. Notably, if both the limit and the function value are finite, there is no difference between proper and improper integration at this limit, since the antiderivative is a continuous function.

Typically one would need to figure out from the context, whether improper integration is desired. Usually one can simply infer it from the fact that an integration limit or the function value is infinite. In case of a finite limit, one can also explicitly indicate that improper integration is desired by using the integration limits of the form $a + 0$ or $b - 0$, e.g.

$$\int_{+0}^{1} \frac{\mathrm{d}t}{\sqrt{t}} = 2\sqrt{t}\Big|_{t=+0}^{1} = 2\sqrt{t}\Big|_{t=0}^{1} = 2$$

Cauchy principal value

A doubly improper integral is commonly defined as a sum of two one-sided improper integrals. In particular:

$$\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = \int_{-\infty}^{c} f(t)\,\mathrm{d}t + \int_{c}^{\infty} f(t)\,\mathrm{d}t$$

where the result must turn out to be independent of the choice of the finite point c. However, we believe it's more intuitive to define such integral as a double limit as we did in (6.51a). Either way, the idea is that both limits have to be approached independently, only then is the integral considered as converging.

Sometimes the integral doesn't converge when the two integration limits are approached independently, but does converge if the limits are approached simultaneously, e.g. with

$$\int_{-\infty}^{\infty} t\,\mathrm{d}t = 0$$

which is trivial, or a less trivial case

$$\int_{-\infty}^{\infty} \frac{t\,\mathrm{d}t}{1+t^2} = 0$$

This is referred as convergence in the *Cauchy principal value* sense. A number of different notations are in use, in this book we employ:

(improper integral in the Cauchy principal value sense)

$$\text{p.v.} \int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = \lim_{\alpha\to+\infty} \int_{-\alpha}^{\alpha} f(t)\,\mathrm{d}t$$

or simply state in words that Cauchy principal value convergence is implied.

Notably, Cauchy principal value integral can be rewritten as an ordinary improper integral

$$\begin{aligned}\text{p.v.}\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t &= \lim_{\alpha\to+\infty}\int_{-\alpha}^{\alpha} f(t)\,\mathrm{d}t = \lim_{\alpha\to+\infty}\int_{0}^{\alpha}\big(f(t)+f(-t)\big)\,\mathrm{d}t \\ &= \int_{0}^{+\infty}\big(f(t)+f(-t)\big)\,\mathrm{d}t\end{aligned}$$

▷ If a doubly improper integral converges in the standard sense, it's not too difficult to realize that it also converges to the same limit in the Cauchy principal value sense (since both integration limits simultaneously approaching their target values is a particular case of them approaching the target values independently). So Cauchy principal value sense doesn't affect the value of an already converging integral.

▷ Sometime Cauchy principal value sense could be used not as a means to amend convergence problems (although that might be a nice byproduct), but simply because the integrals immediately appear in the Cauchy principal value sense by their construction and/or derivation.

The limit order

The understanding of improper integrals as limits of proper integrals suggested by (6.51) is quite a natural way to generalize the idea presented by (6.50), however that's not the only possible understanding. Recall that the integral itself is a limit of a sum (as the partition norm approaches zero). And then, in an improper integral, we apply another limit (according to (6.51)) on top of it. But we could also consider the opposite order of the two limits. Specifically, in (6.50b) we could first construct a Riemann sum by partitioning the infinite interval $(-\infty,\infty)$ and then take the limit of this Riemann sum as partition norm approaches zero:

(an alternative theoretically possible definition of an improper integral)

$$\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = \lim_{\Delta\to 0}\sum_{n=-\infty}^{\infty} f(\tau_n)\Delta t_n \tag{6.54}$$

This might feel even more in line with the original definition of the Riemann integral. Rather than adding another limiting step on top of a normal Riemann integral, we simply make the definition of the Riemann integral accommodate infinite intervals.

The Riemann sum thereby contains an infinite number of terms from the very beginning (assuming the partition norm is finite). Thus the Riemann sum itself needs to be understood as a limit of a finite number of terms, and it is this limit which becomes the inner limit of the two (the outer limit being the usual limit contained in the definition of the Riemann integral). This in turn brings up the question of convergence of such infinite Riemann sum. In particular, conditional convergence might be a thing, since in such case the Riemann sum is no longer unambiguously defined, but rather starts depending on the summation order.

This stops feeling too convincing. However, at least we could consider using (6.54) whenever the Riemann sum under the limit sign in the right-hand side converges absolutely.

The definition (6.54) therefore looks theoretically more consistent, but also potentially more cumbersome to handle in case of convergence issues. If instead of infinite integration limits we are dealing with infinite function values, such as in (6.53), the idea of (6.54) stops working altogether, as there will be always choices of points τ_n making the Riemann sum arbitrarily large, therefore the limit in (6.54) doesn't converge at all.

On the other hand, considering the way how we constructed the Riemann integral in Section 6.2 to begin with, we can notice that $f(t)$ having no value or attaining arbitrarily large values on the integration interval is hardly consistent with the initial idea that we used for Riemann integration. E.g. it probably doesn't make sense to imagine an infinitely large flux of water. This may suggest that in case of improper integrals arising due to function value becoming infinitely large, such as in (6.53), the idea of the Riemann integral cannot be consistently employed to begin with and one might want to come up with some other approach to computing an area under the curve.

> ▷ One such idea could be Lebesgue integration. The standard definition (6.51) of an improper integral can be also considered as something which is quite different from the standard Riemann integral, even though it builds on top of the latter.

We leave the discussion of the definition (6.54) without any conclusive statement. The purpose of this discussion was solely to raise the awareness of an alternative approach to define an integral over an infinite range.

6.7 Partial derivatives

Until now we have been focusing on one-dimensional functions $y = f(t)$ of a one-dimensional argument t. Suppose we are now having a function $z = f(x, y)$ of two arguments x and y. We wish to apply the ideas of differentiation to such function.

Goal 6.6. Extend the previously developed concept of derivative to scalar functions of multiple arguments, or, equivalently, to scalar functions of a vector argument. *Hint:* start with a function of 2 arguments and use the geometric interpretation of derivative of a scalar function of one scalar argument as a tangent line.

6.7.1 Tangent plane

Having two arguments suggests that we cannot assume time semantics, at least not for both arguments simultaneously. Outside of the time semantics of the argument, the idea of the derivative $f'(x)$ can be thought of as determining the slope of the function graph's tangent line. With two arguments, the graph of the function $f(x, y)$ is no longer a line but a surface over a 2D argument plane, therefore instead of a tangent line we should have a tangent plane. For simplicity of the argument, let's first focus on a tangent plane at the point $z = f(0, 0)$. Fig. 6.12 provides an example.

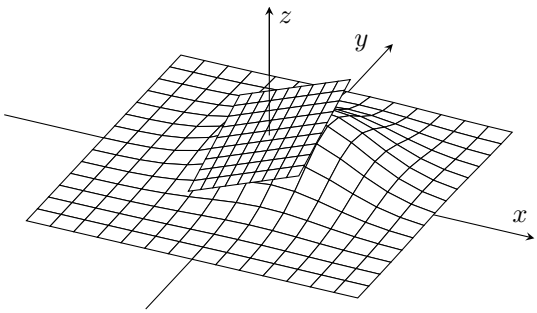


Figure 6.12: A (shifted) 2D Gaussian function and its tangent plane at $x = 0$, $y = 0$.

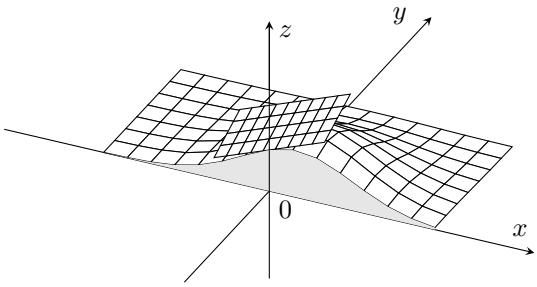


Figure 6.13: A slice of the graph from Fig. 6.12 by the xz plane.

Goal 6.6a. Given $f(x, y)$, develop a way to find the equation of its tangent plane, specifying z as an explicit function of x and y. *Hint:* construct two tangent lines lying in that plane.

A plane is uniquely defined by specifying two intersecting lines within that plane. We can choose one such line going strictly over the x argument axis (thereby lying in the xz coordinate plane) and the other such line going strictly the y argument axis (thereby lying in the yz coordinate plane). Let's consider these two lines one by one.

Starting with the line lying in the xz coordinate plane, we can temporarily disregard the y argument of $f(x, y)$, fixing it to $y = 0$. This reduces the function $f(x, y)$ to a function $f_x(x) = f(x, 0)$ of a single argument x. The graph of $z = f_x(x)$ lying in the coordinate plane xz is simply a slice of the graph of $z = f(x, y)$ (Fig. 6.13). Since we're looking to have a tangent plane in the 3D space of the graph of $z = f(x, y)$, the line we want to choose in the xz plane should be a tangent line to the graph of $z = f_x(x)$. Fig. 6.14 illustrates.

Similarly, in the plane yz we want to choose the tangent line to the graph of $z = f_y(y) = f(0, y)$. These two tangent lines are thus determined by $f'_x(0)$ and $f'_y(0)$.

Now, from a slightly different perspective, a plane in 3D is given by the equation $z = ax + by + c$ (unless this plane is vertical, but in our case we assume it isn't, similarly to how we assume that the tangent line to $y = f(x)$ is not vertical, in order to have a finite derivative value). Substituting $x = y = 0$ immediately gives $z = 0a + 0b + c = c$, but at this point we should have

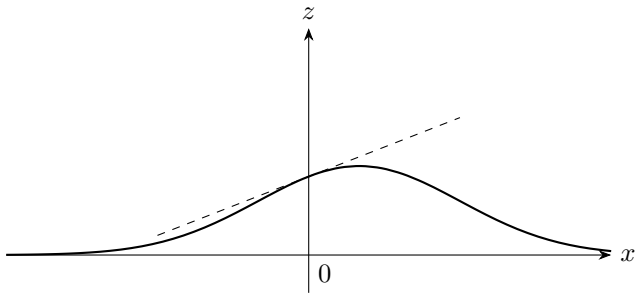


Figure 6.14: The xz plane part of Fig. 6.13, essentially containing the graph of the function $z = f_x(x)$ and its tangent line.

$ax + by + c = f(x, y)$, therefore $c = f(0, 0)$.

In order to consider the plane $z = ax + by + c$ in the coordinate plane xz, we let $y = 0$, thereby turning the plane $z = ax+by+c$ into the line $z = ax+c$, which must be the tangent line to $f_x(x)$ at $x = 0$ (Fig. 6.14). But the latter tangent line is given by $z = f'_x(0)x + f_x(0)$. Now we already have $f_x(0) = f(0, 0) = c$, while the first term of $z = f'_x(0)x + f_x(0)$ suggests that $a = f'_x(0)$. In a totally analogous way, by considering the coordinate plane yz, we obtain $b = f'_y(0)$. Thus we obtained

$$z = ax + bx + c = f'_x(0)x + f'_y(0)y + f(0, 0) \tag{6.55a}$$

as the equation of the tangent plane to $f(x, y)$ at $x = 0$, $y = 0$.

Let's now try to generalize (6.55a) to a tangent plane at an arbitrary point x, y. Instead of the function $f_x(x) = f(x, 0)$ we'll need to introduce the function $f_x(x) = f(x, y)$, and instead of the function $f_y(y) = f(0, y)$ we are going to have $f_y(y) = f(x, y)$:

$$z = ax + bx + c = f'_x(x)x + f'_y(x)y + f(x, y) \tag{6.55b}$$

We can immediately see that the notation $f_x(x)$ and $f_y(y)$ becomes rather clumsy, as the value $f_x(x)$ also depends on y and the value of $f_y(y)$ also depends on x. So instead of $f_x(x)$ we would rather write $f_x(x, y)$ and instead of $f_y(y)$ we would rather write $f_y(x, y)$. But wait, those functions are simply identical to $f(x, y)$, we don't need a different notation for those, except when we specify their derivatives in (6.55b). We are therefore simply going to notate the respective derivatives as $f'_x(x, y)$ and $f'_y(x, y)$:

$$z = ax + bx + c = f'_x(x, y)x + f'_y(x, y)y + f(x, y) \tag{6.55c}$$

Apparently, we have just introduced the idea of

(partial derivatives)

$$f'_x(x, y) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x} \tag{6.56a}$$

$$f'_y(x, y) = \lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} \tag{6.56b}$$

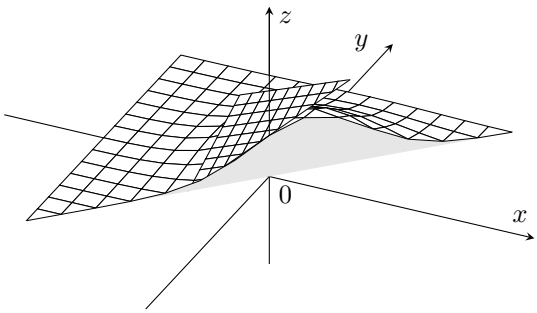


Figure 6.15: A slice of the graph from Fig. 6.12 taken in the 45° direction.

We also simultaneously obtained the tangent plane equation (6.55c).

Goal 6.6a completed. ■

6.7.2 Tangent lines in arbitrary directions

Let's temporarily return to the tangent plane at $z = f(0,0)$. Instead of considering tangent lines in the xz or yz coordinate plane, we could consider any other plane containing the z axis and try to construct a tangent line to $f(x,y)$ within that plane (still at $x = 0$, $y = 0$). Fig. 6.15 illustrates.

Goal 6.6b. Construct the tangent line going in any other given direction than just along the x or the y axis. *Hint:* this line will be lying within the earlier obtained tangent plane.

Let α be the "azimuth angle" of this plane. We could introduce a new auxiliary coordinate axis u lying in the direction of α within the xy plane, yielding

$$x(u) = u\cos\alpha \tag{6.57a}$$
$$y(u) = u\sin\alpha \tag{6.57b}$$

as the conversion equations from the auxiliary coordinate u to the main coordinates x and y. Choosing the multipliers to be equal to the cosine and the sine of α ensures that the axis u uses the same "units of measure" as the x, y, and, most importantly, the z axis.

We now wish to construct the function $f_u(u)$, which would be the counterpart of the functions $f_x(x) = f(x,0)$ and $f_y(y) = f(0,y)$ that we constructed earlier. Apparently

$$f_u(u) = f(x(u), y(u)) = f(u\cos\alpha, u\sin\alpha)$$

We also wish to construct the tangent line to $f_u(u)$ at $u = 0$. Again, this tangent line must lie in the tangent plane to $f(x,y)$, given by the equation (6.55c). Substituting (6.57) into (6.55c) yields

$$\begin{aligned} z &= f'_x(0,0)\cdot t\cos\alpha + f'_y(0,0)\cdot t\sin\alpha + f(0,0) \\ &= \big(f'_x(0,0)\cdot\cos\alpha + f'_y(0,0)\cdot\sin\alpha\big)\cdot t + f(0,0) \end{aligned} \tag{6.58a}$$

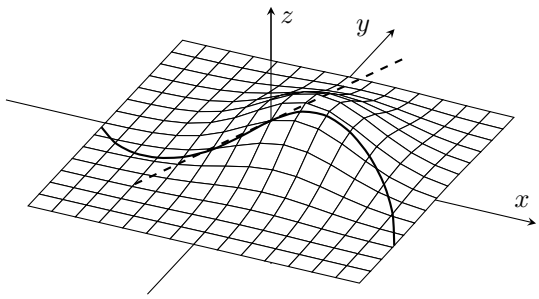


Figure 6.16: A parametric curve $x = x(t)$, $y = y(t)$ going through the graph of the function from Fig. 6.12 and the resulting tangent line (dashed).

The slope of the tangent line going in the direction α is thereby equal to the coefficient of the first-order term in (6.58), which is

$$f'_x(0,0) \cdot \cos\alpha + f'_y(0,0) \cdot \sin\alpha$$

On the other hand, the slope of the tangent line must be equal to $f'_u(0)$. It follows that

$$f'_u(0) = f'_x(0,0) \cdot \cos\alpha + f'_y(0,0) \cdot \sin\alpha \tag{6.59}$$

Goal 6.6b completed. ■

6.7.3 Full derivative

Now that we have learned to construct tangent lines to $z = f(x, y)$ lying in arbitrary directions with respect to the xy plane, we can imagine that one of those tangent lines must also be a tangent line to the parametric 3D curve

$$\begin{aligned} x &= x(t) \\ y &= y(t) \\ z &= f_t(t) = f(x(t), y(t)) \end{aligned}$$

where $x(t)$, $y(t)$ are arbitrary differentiable functions. We are still considering the tangent lines going through $x = 0$, $y = 0$, $z = f(0, 0)$, therefore, for now, we assume $x(0) = 0$, $y(0) = 0$. Fig. 6.16 illustrates.

Goal 6.6c. Construct the tangent line to the just specified parametric 3D curve.

This idea is to choose the auxiliary u axis going in the same direction as the parametric 2D curve $x(t)$, $y(t)$ in the xy plane, which is achieved at α such that

$$\cos\alpha = \frac{x'(0)}{\sqrt{(x'(0))^2 + (y'(0))^2}} \qquad \sin\alpha = \frac{y'(0)}{\sqrt{(x'(0))^2 + (y'(0))^2}}$$

therefore (6.59) turns into

$$f'_u(0) = \frac{f'_x(0,0)x'(0) + f'_y(0,0)y'(0)}{\sqrt{(x'(0))^2 + (y'(0))^2}}$$

At $t = 0$, the variables x and y, expressed as functions of t, are however $\sqrt{(x'(0))^2 + (y'(0))^2}$ times faster than the same variables, expressed as functions of u. Indeed

$$\frac{\mathrm{d}x}{\mathrm{d}u} = \cos\alpha \qquad \frac{\mathrm{d}y}{\mathrm{d}u} = \sin\alpha$$

while

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= x'(0) = \sqrt{(x'(0))^2 + (y'(0))^2} \cdot \cos\alpha = \sqrt{(x'(0))^2 + (y'(0))^2} \cdot \frac{\mathrm{d}x}{\mathrm{d}u} \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= y'(0) = \sqrt{(x'(0))^2 + (y'(0))^2} \cdot \sin\alpha = \sqrt{(x'(0))^2 + (y'(0))^2} \cdot \frac{\mathrm{d}y}{\mathrm{d}u}\end{aligned}$$

where we also have

$$\mathrm{d}u = \sqrt{(x'(0))^2 + (y'(0))^2} \cdot \mathrm{d}t$$

Therefore $f'_t(0)$ must be $\sqrt{(x'(0))^2 + (y'(0))^2}$ times larger than $f'_u(0)$:

$$f'_t(0) = f'_x(0,0) \cdot x'(0) + f'_y(0,0) \cdot y'(0)$$

Finally, realizing that our previous argument is not restricted to $x = 0$, $y = 0$ and can be equally made at any other point x, y, we can rewrite the obtained formula for arbitrary x, y, obtaining

$$f'_t(t) = f'_x(x(t), y(t)) \cdot x'(t) + f'_y(x(t), y(t)) \cdot y'(t) \tag{6.60}$$

Formula (6.60) gives the full derivative of $f_t(t) = f(x(t), y(t))$ expressed in terms of partial derivatives f'_x and f'_y.

> ▷ Notice that our derivation of (6.60) was far from being rigorous. We were concentrating on intuitive aspects of (6.60), while being quite handwavy with a number of details, as well as taking a number of implicit assumptions. The formula (6.60) itself is correct, but holds under certain conditions, which we shall not discuss in this text, advising the readers to check other, more standard, texts on calculus.

Goal 6.6c completed. ■

6.7.4 Partial derivative notation

Let's look again at (6.60). The $f'_t(t)$ in the left-hand side could be notated as $\mathrm{d}z/\mathrm{d}t$, since $z = f_t(t)$. Similarly, $f'_x(x(t), y(t))$ could be notated as $\mathrm{d}z/\mathrm{d}x$, since $z = f_x(x(t), y(t)) = f(x(t), y(t))$, and $x'(t)$ could be notated as as $\mathrm{d}x/\mathrm{d}t$. Similarly, $f'_y(x(t), y(t))$ and $y'(t)$ could be notated as $\mathrm{d}z/\mathrm{d}y$ and $\mathrm{d}y/\mathrm{d}t$ respectively. Summing up, we might rewrite (6.60) as

$$\frac{\mathrm{d}z}{\mathrm{d}t} \stackrel{?}{=} \frac{\mathrm{d}z}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\mathrm{d}z}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}t} \tag{6.61}$$

The equation (6.61) looks pretty suspicious. In particular, we could cancel ("chain-rule style") the common factors $\mathrm{d}x$ in the first term and the common factors $\mathrm{d}y$ in the second term, yielding

$$\frac{\mathrm{d}z}{\mathrm{d}t} \stackrel{?}{=} \frac{\mathrm{d}z}{\mathrm{d}t} + \frac{\mathrm{d}z}{\mathrm{d}t} \tag{6.62}$$

which obviously cannot hold, unless $\mathrm{d}z/\mathrm{d}t = 0$.

The problems with (6.61) are having, however, a purely notational source. When dealing with full differentials, such as $\mathrm{d}z$, $\mathrm{d}x$, etc., we are assuming a specific set of dependencies between all involved variables. In the derivation of (6.60) we have assumed $x = x(t)$, $y = y(t)$ and $z = f(x, y) = f(x(t), y(t))$ as the dependencies. It is in this context, that (6.60) holds. However, in this very context, $\mathrm{d}z/\mathrm{d}x$ doesn't denote the derivative f'_x. The notation $\mathrm{d}z/\mathrm{d}x$ would have worked if z depended solely on x. But in our case z also depends on y, and y does implicitly depend on x, since both x and y are connected via the variable t, that they both are depending on. Therefore $\mathrm{d}z$ does not only include the variation of z due to the change of the first argument of $f(\cdot, \cdot)$, but also the variation of z due to the change of the second argument of $f(\cdot, \cdot)$, since the two arguments cannot change independently. For the same reason, $\mathrm{d}z/\mathrm{d}y$ doesn't denote the derivative f'_y. Thus (6.61) is not a correct notation of (6.60).

We might try to amend the problem by rewriting (6.61) as

$$\frac{\mathrm{d}z}{\mathrm{d}t} \stackrel{?}{=} \frac{\mathrm{d}f}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\mathrm{d}f}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}t}$$

where $\mathrm{d}f/\mathrm{d}x$ could explicitly hint at the fact that we are not considering the dependencies between different variables, but are solely interested in the derivative of f as a function, all dependencies being ignored. This could be still confusing at many occasions, in particular since otherwise $\mathrm{d}f$ and $\mathrm{d}z$ would be often considered equivalent notations, e.g. in case of $z = f(x)$. Also, it is relatively common to reuse the variable names to denote the functions, e.g. we have done exactly that a bit earlier by writing $x = x(t)$, $y = y(t)$.

For the reasons like that, there is a special partial derivative notation introduced instead of trying to use $\mathrm{d}f/\mathrm{d}x$ for this purpose. Instead of $\mathrm{d}f\mathrm{d}x$ we write $\partial f/\partial x$, so that (6.61) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t), y(t)) = \frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} \tag{6.63}$$

(which is just another notation of (6.60)). The partial differential ∂x cannot be cancelled with the full differential $\mathrm{d}x$, and neither can it be done for the respective differentials of y, thereby we don't get to weird results like (6.62).

Generalizing (6.63) to a function of N different variables we obtain

(expansion of full derivative in terms of partial derivatives)

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x_1(t), \ldots, x_N(t)) = \frac{\partial f}{\partial x_1}\frac{\mathrm{d}x_1}{\mathrm{d}t} + \ldots + \frac{\partial f}{\partial x_N}\frac{\mathrm{d}x_N}{\mathrm{d}t} = \sum_{n=1}^{N} \frac{\partial f}{\partial x_n}\frac{\mathrm{d}x_n}{\mathrm{d}t} \tag{6.64}$$

Multiplying (6.64) by $\mathrm{d}t$ we obtain

(full differential in terms of partial derivatives)

$$\mathrm{d}f = \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \ldots + \frac{\partial f}{\partial x_N}\mathrm{d}x_N = \sum_{n=1}^{N} \frac{\partial f}{\partial x_n}\mathrm{d}x_n \tag{6.65}$$

6.7.5 Gradient

The right-hand side of (6.63) can be reexpressed as a dot product of two vectors

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t), y(t)) = \begin{bmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \end{bmatrix} \cdot \begin{bmatrix} \dfrac{\mathrm{d}x}{\mathrm{d}t} \\ \dfrac{\mathrm{d}y}{\mathrm{d}t} \end{bmatrix} = \operatorname{grad} f \cdot \mathbf{r}' \tag{6.66}$$

where

$$\mathbf{r}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$$

therefore its derivative is

$$\mathbf{r}'(t) = \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix}$$

and where

$$\operatorname{grad} f = \begin{bmatrix} \dfrac{\partial f}{\partial x} \\ \dfrac{\partial f}{\partial y} \end{bmatrix}$$

The vector $\operatorname{grad} f$ is referred to as the *gradient of* f (evaluated at the point t). In N dimensions (and a Euclidean space),

(n-dimensional gradient in orthonormal coordinates)

$$\operatorname{grad} f(x_1, \ldots, x_N) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \ldots & \dfrac{\partial f}{\partial x_N} \end{bmatrix}^{\mathrm{T}} \tag{6.67}$$

Since we already kind of introduced $\mathbf{r}(t) = [x_1(t), \ldots, x_N(t)]^{\mathrm{T}}$, we can renotate $f(x_1(t), \ldots, x_N(t))$ as $f(\mathbf{r}(t))$ obtaining a different form of (6.66)

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{r}(t)) = \operatorname{grad} f \cdot \mathbf{r}' \tag{6.68}$$

Equation (6.68) states that the growth speed of $f(\mathbf{r}(t))$ is equal to $\operatorname{grad} f \cdot \mathbf{r}'$, which is thereby maximized (among different $\mathbf{r}'$ of the same absolute magnitude) when $\mathbf{r}'$ is pointing in the same direction as $\operatorname{grad} f$. The vector $\operatorname{grad} f$ thereby points into the direction of the fastest growth of $f(\mathbf{r})$, the absolute magnitude of $\operatorname{grad} f$ determining how quickly this growth happens with respect to the change of the function's argument $\mathbf{r}$.

▷ Actually, by letting $\mathbf{r}(t) = \mathbf{n} \cdot t$, where $|\mathbf{n}| = 1$, we turn $\mathbf{r}(t)$ into a straight line going through the origin in the direction $\mathbf{n}$ at a unit speed, since $|\mathbf{r}'| = |\mathbf{n}| \equiv 1$. In this case the variable t functions as yet another coordinate (of the same scale as the main coordinates $x_1, \ldots, x_n$). If $\mathbf{n}$ is pointing in the same direction as $\operatorname{grad} f$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{r}(t)) = \operatorname{grad} f \cdot \mathbf{r}' = \operatorname{grad} f \cdot \mathbf{n} = |\operatorname{grad} f| \cdot |\mathbf{n}| \cdot \cos 0 = |\operatorname{grad} f|$$

It follows that $|\operatorname{grad} f|$ simply gives the tangent of the function's slope in the direction of its fastest growth.

Conversely, the dot product $\operatorname{grad} f \cdot \mathbf{r}'$ is minimized when both vectors point in opposite directions. Thereby the direction of the fastest decreasing is opposite to the direction of the fastest growth and the slope tangent is given by $-|\operatorname{grad} f|$.

The right-hand side of (6.67) can be formally rewritten as

$$\begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \cdots & \dfrac{\partial f}{\partial x_N} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \dfrac{\partial}{\partial x_1} & \cdots & \dfrac{\partial}{\partial x_N} \end{bmatrix}^{\mathrm{T}} \cdot f = \nabla \cdot f$$

where

(nabla operator in orthonormal coordinates)

$$\nabla = \begin{bmatrix} \dfrac{\partial}{\partial x_1} & \cdots & \dfrac{\partial}{\partial x_N} \end{bmatrix}^{\mathrm{T}} \tag{6.69}$$

is a vector operator. Thereby (6.68) becomes

(differentiation chain rule for a scalar function of a vector argument)

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{r}(t)) = \nabla f \cdot \mathbf{r}' \tag{6.70}$$

or, in the differential form,

$$\mathrm{d}f = \nabla f \cdot \mathbf{r}' \, \mathrm{d}t$$

or

(differential of a scalar function of a vector argument)

$$\mathrm{d}f = \nabla f \cdot \mathrm{d}\mathbf{r} \tag{6.71}$$

Goal 6.6 completed. ■

▷ The gradient formula (6.67) and the nabla operator definition (6.69) assume an orthonormal basis, otherwise the expressions become more complicated. Some further details on that can be found in Section 4.14.

6.8 Multidimensional and iterated integrals

6.8.1 Area integrals in 2D

We have learned how we can differentiate the function $f(\mathbf{r})$ by using the concepts of partial derivatives and gradient. Let's now look into integration of $f(\mathbf{r})$, where, for simplicity, we shall start it two dimensions, that is the argument $\mathbf{r}$ being a two-dimensional vector.

We have found that integration of a one-dimensional function $f(x)$ defines the (signed) area under the function graph. We could generalize the integration of $f(x)$ to integration of $f(\mathbf{r})$ in two dimensions, where it would be natural to require that the integral defines the (signed) volume under the function's graph (where the latter, as we should recall, is a surface in a 3D space, the two dimensions representing the argument $\mathbf{r}$, the third dimension representing the function's value). Respectively, instead of integrating over a one-dimensional

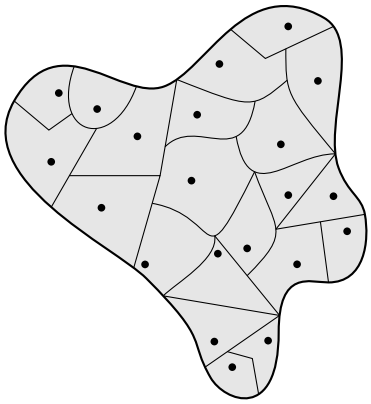


Figure 6.17: Partitioning of a connected 2D shape. Black dots represent the points $\mathbf{r}_n$ of the Riemann sum (6.72).

interval $[a, b]$ of the argument axis, we'll need to integrate over a *compact connected* shape of the argument plane (the compact connected shape being a generalization of a 1D-interval to 2D).

> ▷ A compact shape is a bounded shape which includes its own boundary. A connected shape is a shape, any two points of which can be connected by a continuous path lying fully within the shape.

Goal 6.7. Generalize 1D Riemann integrals to 2D. *Hint:* construct the Riemann sums, as usual, just with respect to the 2D area.

As a 2D generalization of the Riemann partitioning of a closed 1D interval $[a, b]$, suppose we are given a compact connected 2D shape S, which we partition into smaller subshapes (Fig. 6.17). Now we can construct Riemann sums

$$\sum_n f(\mathbf{r}_n)\Delta S_n \tag{6.72}$$

where ΔS_n are the (positive-valued) areas of the partitioning subshapes (so that $\sum_n \Delta S_n$ gives the total area of S), and $\mathbf{r}_n$ are selected points lying within the respective partitioning subshapes.

It might seem a good idea to define the partition norm as $\Delta = \max \Delta S_n$, but then we could realize that this will allow for the partitioning subshapes to have non-decreasing linear dimensions even though $\Delta \to 0$. Indeed, a decreasing ΔS_n can e.g. be an area of a rectangle of a fixed width and a decreasing height. We don't want this, as the function's value can still vary within a large range over such rectangle, while we want the function's value on each partitioning subshape to be constrained to progressively smaller value ranges, as $\Delta \to 0$. Therefore we should define the partition norm as $\Delta = \max \Delta_n$, where Δ_n are the diameters of the partitioning subshapes (a diameter of an arbitrary-form shape is defined as the maximum possible distance between two points belonging to the shape). With this partition norm Δ, we introduce

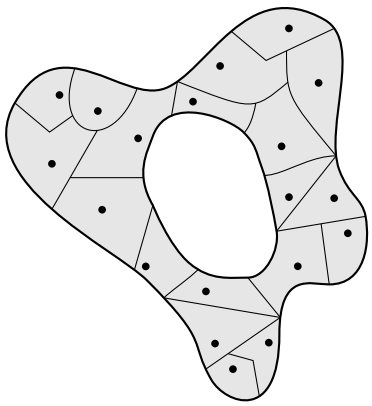


Figure 6.18: Partitioning of a connected 2D shape of genus 1.

(definition of 2D area integral in the Riemann sense)

$$\int_S f(\mathbf{r})\,\mathrm{d}S = \lim_{\Delta\to 0}\sum_n f(\mathbf{r}_n)\Delta S_n \tag{6.73}$$

▷ In some texts the partitioning subshapes may be supposed to assume rectangular (or square) forms with the sides being parallel to the coordinate axes. This however creates preferred directions in space. Therefore we would rather allow the partitioning subshapes to assume arbitrary forms. This, however, in turns opens the question of definition and/or computation of the area of an arbitrary-form shape. We shall address this question in Subsection 6.8.2, for now assuming that we are somehow able to compute the areas of arbitrary-form shapes.

Notably, S doesn't have to be a simply connected shape, in can be a connected shape of any genus. Fig. 6.18 illustrates Riemann partitioning of a connected shape of genus 1. Upon further thinking we can realize that S doesn't even need to be connected. We can create a Riemann partition across a union of multiple disconnected shapes (Fig. 6.19).

Apparently in the case of the situation in Fig. 6.19 we have

(integral over a union of two disconnected areas)

$$\int_{S_1\cup S_2} f(\mathbf{r})\,\mathrm{d}S = \int_{S_1} f(\mathbf{r})\,\mathrm{d}S + \int_{S_2} f(\mathbf{r})\,\mathrm{d}S \tag{6.74}$$

which holds if $S_1 \cap S_2 = \varnothing$, or, more precisely, if the area of $S_1 \cap S_2$ is zero. More generally,

(integral over a union of two areas)

$$\int_{S_1\cup S_2} f(\mathbf{r})\,\mathrm{d}S = \int_{S_1} f(\mathbf{r})\,\mathrm{d}S + \int_{S_2} f(\mathbf{r})\,\mathrm{d}S - \int_{S_1\cap S_2} f(\mathbf{r})\,\mathrm{d}S \tag{6.75}$$

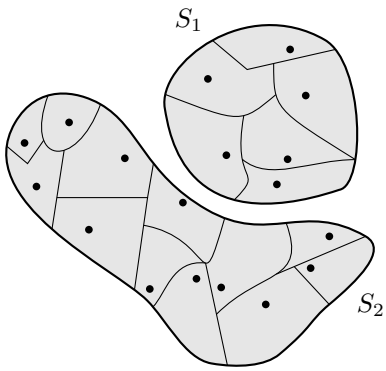


Figure 6.19: Partitioning of a disconnected 2D shape $S = S_1 \cup S_2$. All subshapes are included into a single Riemann sum.

▷ The notation $\int_S f(\mathbf{r})\,\mathrm{d}S$ in combination with the formula (6.74) can be also applied to integrals in one dimension, writing the latter in the form:

$$\int_I f(t)\,\mathrm{d}t$$

where I could be a connected or a disconnected one-dimensional interval. E.g. for a disjoint interval $I = [a, b] \cup [c, d]$, where $c > b$, we would have

$$\int_{[a,b]\cup[c,d]} f(t)\,\mathrm{d}t = \int_{[a,b]} f(t)\,\mathrm{d}t + \int_{[c,d]} f(t)\,\mathrm{d}t$$

6.8.2 Unoriented areas in 2D

In the introduction of the area integrals in Subsection 6.8.1 we assumed that we are somehow able to compute areas of 2D shapes on a plane. But how actually can we do this? How can we even approach the idea of defining the area?

Goal 6.7a. Find a way to define the area of an arbitrary-form shape. *Hint:* use the materials of Subsection 6.8.1.

As a first observation, let's notice that by letting $f(\mathbf{r}) \equiv 1$ in the definition (6.73) of the area integral we make the integration results simply equal to the area of the shape:

$$\int_S \mathrm{d}S = \lim_{\Delta\to 0} \sum_n \Delta S_n = \lim_{\Delta\to 0} S = S \tag{6.76}$$

where

$$S = \sum_n \Delta S_n$$

denotes the total area of the shape S.

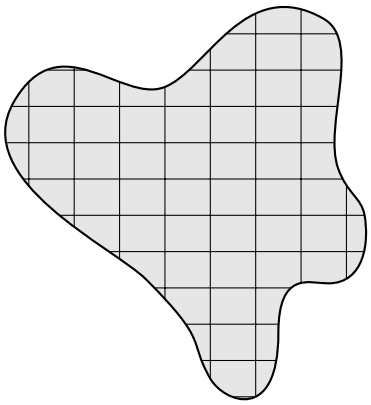


Figure 6.20: Uniform rectangular partitioning of a 2D shape, except at the boundary, where the partitioning subshapes have to assume partially non-rectangular form.

▷ Thereby S denotes both the shape itself and its area. Which of the two meanings is implied in each particular case, should be clear from the context.

Equation 6.76 suggests to define the area of a 2D shape S simply as the value of the integral $\int_S \mathrm{d}S$. Except that this definition would be self-referring, since, in order to compute the Riemann sums (6.72), we need to be able to compute the areas ΔS_n of the partitioning subshapes.

The problem can be solved by realizing that the value yielded by a converging area integral $\int_S \mathrm{d}S$ should not depend on the specific forms of the partitioning subshapes and we can choose these shapes to be of whatever form we like (as long as we can be sure that the area integral itself converges independently of the form of the partitioning subshapes). It would be convenient to let the partitioning subshapes be equally-sized rectangles (or even equally-sized squares). The area of a rectangle can be easily computed as the product of its sides, therefore we can compute ΔS_n for all subshapes, except those lying at the boundary of S. Indeed, we cannot force the subshapes lying at the boundary to assume rectangular shapes, since they, at least partially, need to follow the shape of the boundary of S. Fig. 6.20 illustrates.

However we can notice that the total area under the subshapes at the boundary is approaching zero as the partition norm decreases to zero, therefore we can replace these subshapes with larger rectangular subshapes, or, alternatively discard them altogether (Fig. 6.21). The total area under these rectangular areas at the boundary is approaching zero as well, therefore including them into the Riemann sum or omitting them doesn't affect the limiting value of the sum. Actually, the partitioning doesn't even need to be uniform like in Fig. 6.21. The example in Fig. 6.22 illustrates rectangular partitioning with a number of irregularities.

Thereby we can formally define the area of an arbitrary-form shape in terms of the formula for the area of a rectangle:

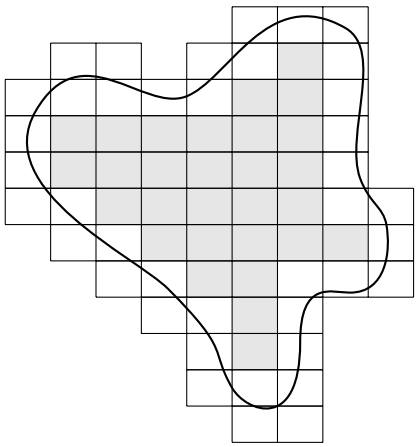


Figure 6.21: Fully uniform rectangular partitioning of a 2D shape. The rectangles at the boundary may be included into or omitted from the Riemann sum.

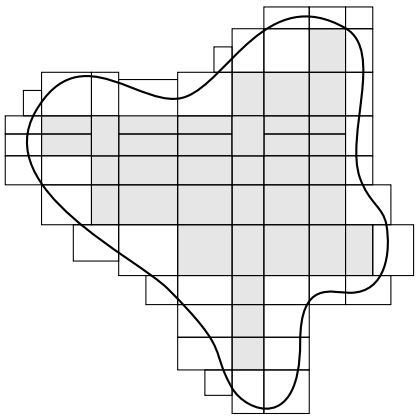


Figure 6.22: Nonuniform rectangular partitioning of a 2D shape. The rectangles at the boundary may be included into or omitted from the Riemann sum.

(area of an arbitrary-form shape)

$$S = \int_S \mathrm{d}S = \lim_{\substack{\Delta_x \to 0 \\ \Delta_y \to 0}} \sum_{n,m} \Delta x_n \Delta y_m \tag{6.77}$$

where the rectangles $\Delta x_n \times \Delta y_m$ are contained inside of or covering the shape S, as in Fig. 6.21.

▷ Conceptually, however, the unsigned area (on a 2D plane) should be defined as an additive continuous nonnegative quantity, associated with 2D shapes. The "units of measure" are specified by postulating the area of a unit square to be unity. Using additivity and continuity we can further infer the formula for the area of a rectangle (or even a parallelogram, as in Chapter 5). By the same additivity and continuity properties, we essentially infer the formula (6.77).

Thereby we are now able to find the areas of arbitrary-form shapes and respectively our definition of the area Riemann integral (6.73) can accommodate arbitrary-form partitioning subshapes, which is what we needed to make it invariant with respect to the choice of the coordinate system.

Furthermore, by pretty much the same argument, not only can we use rectangular partitions to compute the areas of 2D shapes, but we also can use rectangular partitions to compute the Riemann sums in the definition of the area Riemann integral (6.73). Indeed, assuming $f(\mathbf{r})$ is bounded on S, the contributions of the partitioning subshapes lying on the boundary of S approach zero and can be ignored.

▷ If $f(\mathbf{r})$ is not bounded, then we need to be talking of an improper integral in the first place anyway (since S, by assumption, is compact).

▷ Using rectangular partitions in the definition of the area Riemann integral (6.73) is not really necessary, the definition should be formally perfectly applicable assuming arbitrary-form partitions, but is much easier to deal with with rectangular partitions.

Goal 6.7a completed. ■

Goal 6.7 completed. ■

6.8.3 Iterated integrals

In Subsection 6.8.2 we developed a way to build 2D Riemann partitions using rectangular partitioning subshapes. We also mentioned that rectangular partitions can be used both to compute the area of the shape S as well as a 2D area integral over this shape. Computing the area of a rectangular shape S using Riemann partitions sounds like an unnecessary complication, since the area is simply equal to the product of the sides, however using this kind of partitions to compute a 2D area integral over such shape sounds promising.

Computing 2D area integrals using rectangular partitions is already relatively easy, but becomes particularly easy if the shape S itself is a rectangle oriented along the coordinate axes, and this is what we shall look at now.

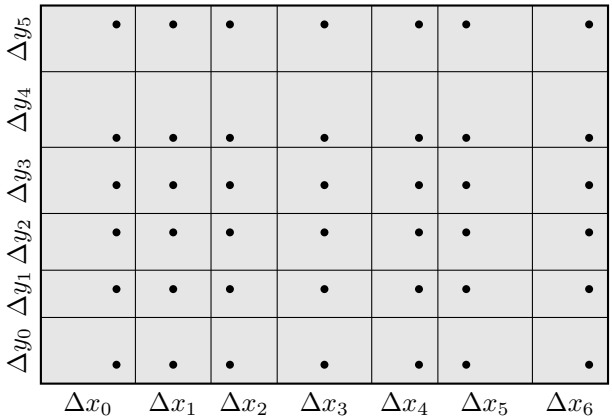


Figure 6.23: Partitioning of a rectangle. Black dots represent the points $\mathbf{r}_{mk} = [\xi_m \ \upsilon_k]^{\mathrm{T}}$ at which the function is evaluated in the Riemann sum.

Goal 6.8. Find a straightforward way to compute a 2D area integral over a rectangular area aligned with the coordinate axis directions.

So suppose that $S = [a, b] \times [c, d]$ is a rectangle whose sides are parallel to the coordinate axes x and y, so that the points within S satisfy $x \in [a, b]$ and $y \in [c, d]$. We can therefore construct partitionings of S as Cartesian products of the partitionings of the intervals $[a, b]$ and $[c, d]$.

Suppose

$$\begin{aligned} x_0 &= a \\ x_1 &= x_0 + \Delta x_0 \\ &\vdots \\ x_M &= x_{M-1} + \Delta_{M-1} = b \end{aligned}$$

is a partition of $[a, b]$ with $\xi_m \in [x_m, x_{m+1}]$ $\forall m$ and suppose

$$\begin{aligned} y_0 &= c \\ y_1 &= y_0 + \Delta y_0 \\ &\vdots \\ y_K &= y_{K-1} + \Delta_{K-1} = d \end{aligned}$$

is a partition of $[c, d]$ with $\upsilon_k \in [y_k, y_{k+1}]$ $\forall k$. Then we can construct a Cartesian product of these two partitions, consisting of partitioning rectangles $[x_m, x_{m+1}] \times [y_k, y_{k+1}]$ and points $\mathbf{r}_{mk} = [\xi_m \ \upsilon_k]^{\mathrm{T}}$ belonging to the respective rectangles. Fig. 6.23 illustrates.

We thereby obtained a rectangular partitioning of S. Let $\Delta S_n = \Delta x_m \Delta y_k$ be the areas of the partitioning rectangles, and let $\mathbf{r}_n$ be the previously introduced points $\mathbf{r}_{mk}$, just alternatively indexed, where the index n simply goes over all partitioning rectangles. This partitioning is a subset of possible partitionings of S required by the definition (6.73) and therefore the respective Riemann sums

should converge to the same limit, and we can compute the area integral as

$$\begin{aligned}
\int_S f(\mathbf{r})\,\mathrm{d}S &= \lim_{\Delta\to 0}\sum_n f(\mathbf{r}_n)\Delta S_n = \lim_{\substack{\Delta_x\to 0\\ \Delta_y\to 0}}\sum_{m,k} f(\xi_m,\upsilon_k)\Delta x_m\Delta y_k \\
&= \lim_{\substack{\Delta_x\to 0\\ \Delta_y\to 0}}\sum_k\left(\sum_m f(\xi_m,\upsilon_k)\Delta x_m\right)\Delta y_k \\
&= \lim_{\Delta y\to 0}\sum_k\left(\lim_{\Delta x\to 0}\sum_m f(\xi_m,\upsilon_k)\Delta x_m\right)\Delta y_k \\
&= \lim_{\Delta y\to 0}\sum_k\left(\int_a^b f(x,\upsilon_k)\,\mathrm{d}x\right)\Delta y_k \\
&= \int_c^d\left(\int_a^b f(x,y)\,\mathrm{d}x\right)\mathrm{d}y = \int_c^d\int_a^b f(x,y)\,\mathrm{d}x\,\mathrm{d}y \qquad (6.78\text{a})
\end{aligned}$$

(where in the latest expression we simply omitted the parentheses, understanding them implicitly).

Conversely, we could take summation over y rather than x as the nested limit/sum, yielding the opposite formula

$$\int_S f(\mathbf{r})\,\mathrm{d}S = \int_a^b\int_c^d f(x,y)\,\mathrm{d}y\,\mathrm{d}x \qquad (6.78\text{b})$$

We can summarize the obtained results in a concise form as

(rectangular area integral via iterated integrals)

$$\int\limits_{[a,b]\times[c,d]} f(\mathbf{r})\,\mathrm{d}S = \int_c^d\int_a^b f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b\int_c^d f(x,y)\,\mathrm{d}y\,\mathrm{d}x \qquad (6.79)$$

The formula (6.79) holds pretty much whenever the area integral in the left-hand side converges, since the Riemann sums implied by the iterated integrals in the second and third parts of (6.79) are essentially particular cases of the Riemann sum implied by the area integral in the left-hand side (as demonstrated by (6.78a)). Now the area integral will converge if $f(x,y)$ is continuous on $S = [a,b]\times[c,d]$, but it also, not unlike the one-dimensional case, can accommodate discontinuities. E.g. there can be a finite number of jump discontinuities, where each discontinuity can be a whole one-dimensional line (as long as these lines do not get indefinitely close to each other or to themselves).

▷ In (6.79) let

$$F(b,y) = \int_a^b f(x,y)\,\mathrm{d}x \qquad \frac{\partial}{\partial b}F(b,y) = f(b,y)$$

in which case (6.79) takes the form

$$\int_c^d F(b,y)\,\mathrm{d}y = \int_a^b\int_c^d \frac{\partial}{\partial x}F(x,y)\,\mathrm{d}y\,\mathrm{d}x$$

Differentiating both sides with respect to b,

$$\frac{\partial}{\partial b}\int_c^d F(b,y)\,\mathrm{d}y = \int_c^d \left(\frac{\partial}{\partial x}F(x,y)\bigg|_{x=b}\right)\mathrm{d}y$$

Renaming b to x, we obtain the formula

(differentiation with respect to a parameter under an integral sign)

$$\frac{\partial}{\partial x}\int_c^d F(x,y)\,\mathrm{d}y = \int_c^d \frac{\partial}{\partial x}F(x,y)\,\mathrm{d}y \tag{6.80}$$

Thus, (6.80) is a slightly different, almost equivalent, form of the second equality of (6.79). Respectively, the second equality of (6.79) can be understood as *integration under the integral sign*, where the outside integral with respect to x (in the right-hand side) can be pulled inside the integral with respect to y.

▷ By intuitively understanding the integral in (6.80) as an infinite sum, we can see (6.80) as a generalization of the formula

$$\frac{\mathrm{d}}{\mathrm{d}x}\sum_{n=1}^{N} f_n(x) = \sum_{n=1}^{N}\frac{\mathrm{d}}{\mathrm{d}x}f_n(x)$$

Similarly, the second equality in (6.79) can be understood as a generalization of

$$\int_a^b \left(\sum_{n=1}^{N} f_n(x)\right)\mathrm{d}x = \sum_{n=1}^{N}\int_a^b f_n(x)\,\mathrm{d}x$$

Goal 6.8 completed. ■

Goal 6.9. Develop a practical way to compute area integrals over arbitrary shapes. *Hint:* first try to extend (6.79), then consider splitting the shape into simpler parts.

The formula (6.79) can be generalized to shapes of somewhat more complicated forms by turning the inner integration limits into functions of the outer integration variable. E.g.

$$\int_S f(\mathbf{r})\,\mathrm{d}S = \int_c^d \int_{a(y)}^{b(y)} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{6.81}$$

Fig. 6.24 illustrates. Notice that, as the shape stops being rectangular, the partitioning subshapes at the boundaries stop being rectangular as well, but, as discussed earlier, it doesn't really matter, since the total area covered by these boundary subshapes is approaching zero anyway.

▷ As usual, by letting $f(\mathbf{r}) = f(x,y) \equiv 1$ in (6.81) we can use (6.81) to compute the area of the shape, as the integral simply turns into $\int_S \mathrm{d}S$.

As the shape becomes even more complicated we might not be able to express the area integral in the form (6.81) anymore. In this case we might split the complex shape into simpler-form smaller shapes, applying iterated integration to each of these shapes in turn, and then combine the results using (6.74).

Goal 6.9 completed. ■

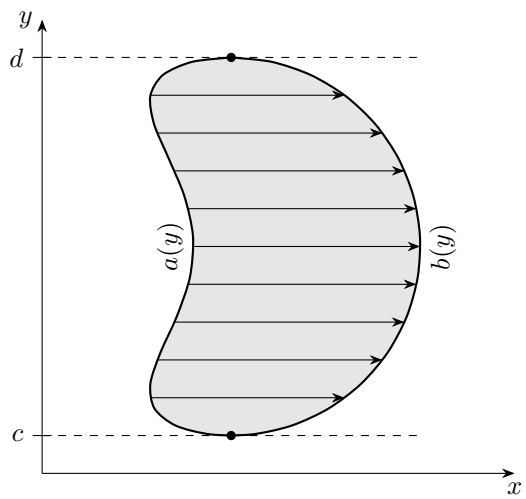


Figure 6.24: Iterated integration by (6.81). The inner integral is computed along each horizontal line, then the obtained results are integrated with respect to y over $[c, d]$.

6.8.4 Area orientation in 2D

Suppose we swap the integration limits a and b in (6.79): Apparently, thereby the integration result changes the sign:

$$\int_c^d \int_b^a f(x, y)\,\mathrm{d}x\,\mathrm{d}y = -\int_c^d \int_a^b f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

This is the same effect as swapping the integration limits in a one dimensional integral:

$$\int_b^a f(t)\,\mathrm{d}t = -\int_a^b f(t)\,\mathrm{d}t$$

which we can intuitively understand as "integrating in the opposite direction". That is, one-dimensional integration "has a direction". This direction can be positive, if in $\int_a^b f\,\mathrm{d}t$ we have $b > a$, or it can be negative if $a > b$.

One-dimensional integration is done over a one-dimensional interval. Therefore we could also refer to the "integration direction" as the "orientation of the respective interval". E.g. we would refer to the interval $[0, 1]$ as "positively oriented" and to the interval $[1, 0]$ as "negatively oriented". Respectively, we can associate a "signed length" with each of these intervals, the lengths being given by the respective integrals:

$$\int_0^1 \mathrm{d}x = 1 \qquad \int_1^0 \mathrm{d}x = -1$$

In the same way, we could associate an "orientation" property with each 2D shape, and we should do so in a way which is compatible with (6.79). In particular, we should consider $[0, 1] \times [0, 1]$ as a positively oriented square, since

its area is given by

$$\int_{[0,1]\times[0,1]} \mathrm{d}S = \int_0^1 \int_0^1 \mathrm{d}x\,\mathrm{d}y = 1$$

On the other hand, $[1,0]\times[0,1]$ is a negatively oriented square:

$$\int_{[1,0]\times[0,1]} \mathrm{d}S = \int_0^1 \int_1^0 \mathrm{d}x\,\mathrm{d}y = -1$$

The square $[1,0]\times[1,0]$ is again positively oriented

$$\int_{[1,0]\times[1,0]} \mathrm{d}S = \int_1^0 \int_1^0 \mathrm{d}x\,\mathrm{d}y = \int_1^0 (-1)\,\mathrm{d}y = 1$$

Generally, the area of a rectangle $[a,b]\times[c,d]$ is thereby given by (6.79) regardless of whether $a<b$ or $a>b$ and $c<d$ or $c>d$:

$$\int_{[a,b]\times[c,d]} \mathrm{d}S = \int_c^d \int_a^b \mathrm{d}x\,\mathrm{d}y = (b-a)(c-d)$$

If a shape is given in more abstract terms, e.g. simply as a picture, we'll have to endow this picture with an additional piece of information specifying the orientation of this shape. Pretty much the same as, being given an unoriented 1D interval (maybe also, as a picture), we'll have to specify its orientation explicitly as well. So e.g. for the shape in Fig. 6.17 we would additionally specify if the shape is positively or negatively oriented, respectively obtaining a positive or a negative area of the shape. The areas of the partitioning subshapes in Fig. 6.17 would respectively also change their signs, thereby affecting the sign of the obtained integration result.

A particularly visual and intuitive way to specify the orientation of a shape is by specifying the orientation of its boundary. This possibility is due to Green's theorem (or Stokes' theorem for an oriented surface in 3D), which relates both orientations to each other. Both theorems are discussed later in this chapter.

> ▷ The idea of the signed area of an oriented shape, which we just introduced, is pretty much the same as the idea of the signed area discussed in Chapter 5. It doesn't deprecate the idea of the (nonnegative) unoriented area. Rather, both ideas coexist.

6.8.5 Volume and hypervolume integrals

In a similar way to how we integrated $f(\mathbf{r})$ over a 2D area, in a 3D setup ($\mathbf{r}$ being a 3D vector) we could integrate $f(\mathbf{r})$ over the volume of a 3D shape V:

(definition of volume integral in the Riemann sense)

$$\int_V f(\mathbf{r})\,\mathrm{d}V = \lim_{\Delta\to 0} \sum_n f(\mathbf{r}_n)\Delta V_n \tag{6.82}$$

where V is the shape (but also can denote the shape's volume) and ΔV_n are the volumes of the partition elements. If V is a rectangular parallelepiped (box), the 3D counterpart of (6.79) applies:

(rectangular 3D volume integral via iterated integrals)

$$\int\limits_{[a,b]\times[c,d]\times[g,h]} f(\mathbf{r})\,\mathrm{d}V = \int_g^h \int_c^d \int_a^b f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \tag{6.83}$$

(where in the right-hand side we specified only one of the possible $3! = 6$ orders of integration). For volumes of more complicated shapes the 3D counterparts of (6.81) and (6.74) apply.

The generalization to hypervolumes in higher dimension counts is obvious. In particular, (6.82) is applicable in any dimension count.

Goal 6.10. Figure out a way to replace the integration variable in (hyper)volume integrals, similarly to how we can replace the integration variable t in the 1D integral $\int_a^b f\,\mathrm{d}t$ with some other variable τ, related to t as $t = g(\tau)$. *Hint:* use the materials of Chapter 5.

Suppose $f(\mathbf{r})$ is a function which we integrate over the (hyper)volume of a shape V:

$$\int_V f(\mathbf{r})\,\mathrm{d}V \tag{6.84}$$

Suppose V is an image of another shape W under a transformation $\mathbf{g}(\cdot)$:

$$\mathbf{r} = \mathbf{g}(\mathbf{q})$$

where

$$\mathbf{q} \in W \iff \mathbf{r} \in V$$

We wish to express the integral (6.84) in terms of integration over W instead of integration over V.

Let $\mathrm{d}W$ denote the (hyper)volume differential of W. The partitioning of W can use arbitrary-form subshapes. In particular, we could choose them to be (hyper)parallelepipeds. Thus, we could see $\mathrm{d}W$ as a hypervolume of an infinitely small hyperparallelepiped. Assuming $\mathbf{g}(\mathbf{q})$ is differentiable, the image of the latter infinitely small hyperparallelepiped under $\mathbf{g}(\cdot)$ is another infinitely small hyperparallelepiped of hypervolume $\mathrm{d}V$. Indeed, by (6.65) each of the components of $\mathbf{g}(\mathbf{q})$ is a linear function in an infinitely small neighborhood of $\mathbf{q}$:

$$\begin{aligned} g_n(\mathbf{q}+\mathrm{d}\mathbf{q}) &= g_n(\mathbf{q}) + \mathrm{d}g_n(\mathbf{q}) = g_n(\mathbf{q}) + \sum_k \frac{\partial g_n}{\partial q_k}\,\mathrm{d}q_k \\ &= g_n(\mathbf{q}) + \begin{bmatrix} \frac{\partial g_n}{\partial q_1} & \cdots & \frac{\partial g_n}{\partial q_N} \end{bmatrix} \cdot \mathrm{d}\mathbf{q} = g_n(\mathbf{q}) + \nabla g_n \cdot \mathrm{d}\mathbf{q} \end{aligned}$$

(where we discarded the $o(\mathrm{d}\mathbf{q})$ term for simplicity of notation, and where the first dot multiplication stands for the product of a row matrix with a column matrix, while the second dot multiplication is understood as a dot product of two column vectors). Combining all g_n to a single column matrix $\mathbf{g}$, we obtain

$$\mathbf{g}(\mathbf{q}+\mathrm{d}\mathbf{q}) = \mathbf{g}(\mathbf{q}) + \begin{bmatrix} \frac{\partial g_1}{\partial q_1} & \cdots & \frac{\partial g_1}{\partial q_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_N}{\partial q_1} & \cdots & \frac{\partial g_N}{\partial q_N} \end{bmatrix} \mathrm{d}\mathbf{q} = \mathbf{g}(\mathbf{q}) + \mathbf{J} \cdot \mathrm{d}\mathbf{q}$$

Therefore

$$\mathrm{d}\mathbf{r} = \mathbf{g}(\mathbf{q} + \mathrm{d}\mathbf{q}) - \mathbf{g}(\mathbf{q}) = \mathbf{J} \cdot \mathrm{d}\mathbf{q}$$

The matrix

$$\mathbf{J} = \begin{bmatrix} \frac{\partial g_1}{\partial q_1} & \cdots & \frac{\partial g_1}{\partial q_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_N}{\partial q_1} & \cdots & \frac{\partial g_N}{\partial q_N} \end{bmatrix}$$

is known as *Jacobian matrix* or simply *Jacobian*. By (5.61)

$$\mathrm{d}V = \det \begin{bmatrix} \frac{\partial g_1}{\partial q_1} & \cdots & \frac{\partial g_1}{\partial q_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_N}{\partial q_1} & \cdots & \frac{\partial g_N}{\partial q_N} \end{bmatrix} \cdot \mathrm{d}W = \det \mathbf{J} \cdot \mathrm{d}W$$

The determinant $\det \mathbf{J}$ is referred to as the *Jacobian determinant* or also simply the *Jacobian* of the transformation $\mathbf{g}(\mathbf{q})$.

We therefore obtain the formula

(change of hypervolume integration variable)

$$\int_V f(\mathbf{r})\,\mathrm{d}V = \int_W \Big(f(\mathbf{g}(\mathbf{q})) \cdot \det \mathbf{J} \cdot \mathrm{d}W \Big) \tag{6.85}$$

▷ In formula (6.85) it is a bit cumbersome to remember that functions of $\mathbf{r}$ need to be integrated with respect to $\mathrm{d}V$ and functions of $\mathbf{q}$ need to be integrated with respect to $\mathrm{d}W$. This problem didn't exist in one dimension, as the function argument and the integration variable are the same, e.g. $\int f(t)\,\mathrm{d}t$. Notice that we cannot simply write $\mathrm{d}\mathbf{r}$, since that would be a differential of position, not of hypervolume, and we cannot write $f(V)$ since we need to supply a position vector, not a volume, as the argument of $f(\cdot)$.

As a way around that problem, sometimes the volume differential $\mathrm{d}V$ is notated as $\mathrm{d}^3\mathbf{r}$ (or, generally, in N dimensions, $\mathrm{d}^N\mathbf{r}$). This is particularly common in physics texts. Notice that $\mathrm{d}^3\mathbf{r}$ in this case doesn't denote the 3rd-order differential of $\mathbf{r}$, but a volume differential, something which (in agreement with (6.83)) can also be loosely thought of as the product of 3 differentials $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ (although the latter implies a rectangular parallelepiped, while a volume differential theoretically can assume arbitrary shapes).

Goal 6.10 completed. ■

6.8.6 Further integrals in multiple dimensions

Instead of always integrating functions with respect to the "differential of the hypervolume" (which, in smaller dimension counts, includes differentials of area, volume, length and position), we could consider a much wider family of integrals. First, we shall notice that the expression under the hypervolume integral in the left-hand side of (6.82) can be seen as a differential of some scalar-valued function F:

$$f(\mathbf{r})\,\mathrm{d}V = \mathrm{d}F$$

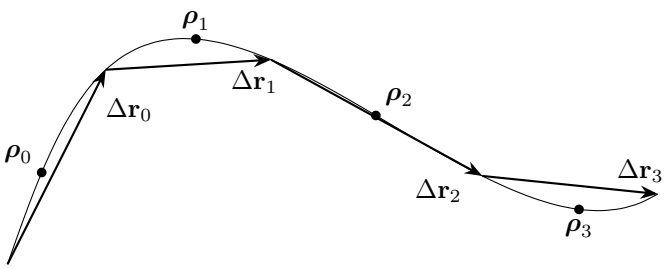


Figure 6.25: Secant line-based Riemann partitioning for the path integral (6.86).

respectively the hypervolume integral can be rewritten as

$$\int_V \mathrm{d}F$$

This observation suggests that we can apply a hypervolume integral (including, but not limited to area, volume, length and path integrals as its partial cases) to anything which looks like a differential of a function.

Path integrals

E.g. we could take the following path integral over some curve C on a 2D plane (or in 3D, or in a higher dimensional space):

(Riemann path integral)

$$\int_C \mathbf{f}(\mathbf{r}) \cdot \mathrm{d}\mathbf{r} = \lim_{\Delta \to 0} \sum_n \big(\mathbf{f}(\boldsymbol{\rho}_n) \cdot \Delta\mathbf{r}_n\big) \tag{6.86}$$

where the Riemann sum splits the curve C into a sequence of small displacements $\Delta\mathbf{r}_n$ (Fig. 6.25). The secant vectors $\Delta\mathbf{r}_n$ build a sequence of approximations of the curve, converging to the curve itself in the limit $\Delta \to 0$.

▷ It might have been even more consistent to define $\Delta\mathbf{r}_n$ in (6.86) not as secant vectors (which is what we suggested by Fig. 6.25), but as tangent vectors, where we let the length of $\Delta\mathbf{r}_n$ be equal to the length of the respective curve segment, which is being approximated by $\Delta\mathbf{r}_n$. Fig. 6.26 illustrates. Notice that, in order to avoid a self-referring definition, we still have to use the secant-vector interpretation of (6.86) for the computation of the segment lengths.

Apparently, normally (6.86) would converge to one and the same value in both interpretations.

We could notice that the expression $\mathbf{f}(\mathbf{r}) \cdot \mathrm{d}\mathbf{r}$ under the integral sign in (6.86) looks like a differential of some scalar-valued function F. We could even figure out some specific details about F (although we don't really have to). Specifically, we could notice that

$$\mathrm{d}F \overset{(6.71)}{=} \nabla F \cdot \mathrm{d}\mathbf{r} = \mathbf{f}(\mathbf{r}) \cdot \mathrm{d}\mathbf{r}$$

thus F needs to satisfy $\mathbf{f} = \nabla F$, where we implicitly assumed that F is a function of $\mathbf{r}$ in order to be able to apply (6.71).

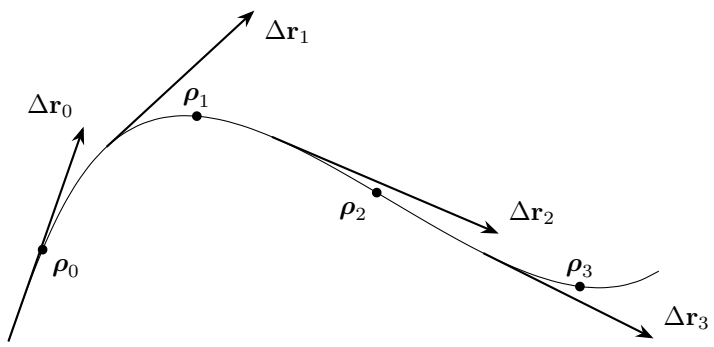


Figure 6.26: Tangent line-based Riemann partitioning for the path integral (6.86).

If C is a closed curve, such function $F(\mathbf{r})$ might not exist.[5] However F doesn't need to be a function of $\mathbf{r}$. E.g. it can be a function of some other variable t, which parametrizes the curve C (with t going from t_0 to t_1), so that $\mathbf{r} = \mathbf{r}(t)$ is also a function of the same variable, and we have

$$\mathrm{d}F(t) = \mathbf{f}(\mathbf{r}(t)) \cdot \mathrm{d}\mathbf{r}(t)$$

This suggests that the Riemann path integral (6.86) can be rewritten as an ordinary Riemann–Stieltjes integral with respect to t, which then further transforms into a Riemann integral with respect to the same variable:

$$\int_{t_0}^{t_1} \mathrm{d}F(t) = \int_{t_0}^{t_1} \mathbf{f}(\mathbf{r}(t)) \cdot \mathrm{d}\mathbf{r}(t) \overset{(6.30)}{=} \int_{t_0}^{t_1} \mathbf{f}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) \cdot \mathrm{d}t \tag{6.87}$$

Surface integrals

Similarly, in 3D we could imagine the integral

(Riemann integral over a 3D surface)

$$\int_S \vec{f}(\vec{r}) \cdot \mathrm{d}\vec{S} \tag{6.88}$$

where $\vec{r}$ is changing within an oriented surface S (the vector $\mathrm{d}\vec{S}$ being the oriented differential of the surface area). Like $\Delta\mathbf{r}_n$ in (6.86), $\mathrm{d}\vec{S}$ in (6.88) are limiting versions of finite oriented secant plane area elements $\Delta\vec{S}$, approximating the surface S.

▷ We switched to the "geometric" vector notation $\vec{r}$, $\vec{f}$ and $\mathrm{d}\vec{S}$ (compared to the previously used notation $\mathbf{r}$, $\mathbf{f}$ etc., which in this book typically suggests column matrix interpretation) in (6.88) to illustrate that one can have more abstract entities under the sign of the integral. Also the notation $\mathrm{d}\mathbf{S}$ might be looking a bit confusing, since using a bold capital letter might suggest a matrix rather than a vector.

[5] This happens if the integral (6.86) over this closed curve converges to a nonzero value, implying that $\mathbf{f}(\mathbf{r})$ describes a non-conservative field.

The same integral (6.88) in the Riemann–Stieltjes form can be written as

$$\int_{v_0}^{v_1}\int_{u_0}^{u_1} \vec{f}(u,v)\cdot \mathrm{d}\vec{S}(u,v)$$

where u, v are the free variables of the parametric equation $\vec{r}=\vec{r}(u,v)$ of the surface, and the oriented area differential is given by

$$\mathrm{d}\vec{S}(u,v)=\frac{\partial\vec{r}}{\partial u}\,\mathrm{d}u\times\frac{\partial\vec{r}}{\partial v}\mathrm{d}v=\left(\frac{\partial\vec{r}}{\partial u}\times\frac{\partial\vec{r}}{\partial v}\right)\cdot\mathrm{d}u\,\mathrm{d}v$$

(which is obtained similarly to the derivation of (6.85)).

▷ More correctly, $\vec{f}(u,v)\cdot\mathrm{d}\vec{S}(u,v)$ is a triple vector product

$$\vec{f}(u,v)\cdot\mathrm{d}\vec{S}(u,v)=\star\left(\frac{\partial\vec{r}}{\partial u}\,\mathrm{d}u\wedge\frac{\partial\vec{r}}{\partial v}\mathrm{d}v\wedge\vec{f}(u,v)\right)=\star\left(\vec{f}(u,v)\wedge\frac{\partial\vec{r}}{\partial u}\wedge\frac{\partial\vec{r}}{\partial v}\right)\cdot\mathrm{d}u\,\mathrm{d}v$$

The readers may wish to review the material of Chapter 5 in connection with the above formula.

Vector/tensor-valued (hyper)volume integrals

We could also imagine cases where the differential under the integral sign is a vector (or a tensor), e.g.

$$\int_V \vec{f}(\vec{r})\,\mathrm{d}V$$

but these are less common.

6.8.7 Iterated improper integrals

In formula (6.79) one of the integrals over $[a,b]$ and $[c,d]$ (or both) could be improper. In such case, when swapping the relative order of the two integrals, we, instead of swapping the relative order of two implied limits (corresponding to the two respective partition norms approaching zero), are changing the relative order of a larger number of implied limits. The extra limits are coming from the improper lower and upper limits of the integrals. E.g.

$$\begin{aligned}&\int_0^{+\infty}\int_0^{+\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y\\&\quad=\lim_{d\to+\infty}\lim_{\Delta_y\to 0}\sum_n\lim_{b\to+\infty}\lim_{\Delta_x\to 0}\left(\sum_m f(x_m,y_n)\Delta x_m\Delta y_n\right)\end{aligned}$$

compared to

$$\int_0^d\int_0^b f(x,y)\,\mathrm{d}x\,\mathrm{d}y=\lim_{\Delta_y\to 0}\sum_n\lim_{\Delta_x\to 0}\left(\sum_m f(x_m,y_n)\Delta x_m\Delta y_n\right)$$

where b and d are finite.

Respectively, the conditions, under which we can swap the relative order of the two integrals without affecting the result, thereby become more fragile, and the equality

$$\int_0^{+\infty}\int_0^{+\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^{+\infty}\int_0^{+\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x \tag{6.89a}$$

is holding in a smaller number of cases than

$$\int_0^{d}\int_0^{b} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^{b}\int_0^{d} f(x,y)\,\mathrm{d}y\,\mathrm{d}x \tag{6.89b}$$

Finding the exact conditions, under which (6.89a) holds, is clearly outside the focus of this book. But we can still take a look at a number of intuitive aspects of this problem.

Absolutely integrable functions

We shall assume that the function $f(x,y)$ is sufficiently "nice" (e.g. continuous), so that (6.89b) holds for any positive b and d, and want to find out whether the property (6.89b) continues to hold in the limit $b \to +\infty$, $d \to +\infty$. We can start by drawing a parallel between (6.89) and the formulas

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} a_{mn} = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} a_{mn} \tag{6.90a}$$

$$\sum_{n=0}^{N}\sum_{m=0}^{M} a_{mn} = \sum_{m=0}^{M}\sum_{n=0}^{N} a_{mn} \tag{6.90b}$$

The formula (6.90b) apparently holds always (and we just ensured that its counterpart (6.89b) also holds always, by simply including this condition into our assumptions). Now we would be interested, when does (6.90a) hold, and then maybe (6.89a) would hold under a similar set of conditions.

The left- and right-hand sides of (6.90a) perform infinite summation of a countable total number of terms (by the discussion in Section 2.8 we should remember that $\aleph_0^2 = \aleph_0$) in two different summation orders. Changing the summation order of an infinite number of terms is generally possible under the condition of absolute summability. Respectively, we could hope that (6.89a) holds under the condition of absolute integrability of $f(x,y)$:

$$\int_0^{+\infty}\int_0^{+\infty} |f(x,y)| \cdot \mathrm{d}x\,\mathrm{d}y < +\infty$$

Essentially, this is the statement of *Fubini's theorem*.

▷ Strictly speaking, changing the order of summation in (6.90a) is not directly comparable to changing the order of the terms of an absolutely converging series, because in (6.90a) we are not adding all a_{mn} term by term (in which case the usual theorem about unconditional convergence of an absolutely convergence series applies), but are adding the values of the nested sums. So, rigorously speaking, (6.90a) requires a separate proof (which is Fubini's theorem for series).

Uniform convergence: one of the integrals is improper

An alternative idea is based on the possibility to swap two limits under the condition of the uniform convergence of the inner limit of the two.

▷ Uniform convergence is often seen as some kind of "advanced" kind of convergence, the pointwise convergence being intuitively felt as the "default" one. In reality it is, however, the other way around: uniform convergence should be considered as default. Indeed, let's formally write the standard ε-N statement for a sequence of functions $\{f_n\}$ converging to the function f, as if it was just a number sequence $\{a_n\}$ converging to the number a:

$$\forall \varepsilon > 0 \implies \exists N(\varepsilon)\colon \forall n > N(\varepsilon) \implies \|f_n - f\| < \varepsilon$$

where $\|f_n - f\|$ occurs in place of $|a_n - a|$. The expression $\|f_n - f\|$ is supposed to give the distance of f_n from f, pretty much like $|a_n - a|$ gives the distance from a_n to a. The norm expression $\|f_n - f\|$ could be defined in a number of different ways, but the most intuitive in the current context is probably simply

$$\|f_n - f\| = \max |f_n(t) - f(t)|$$

(or, more formally correctly, $\sup |f_n(t) - f(t)|$). However we thereby obtain exactly the statement of the uniform, rather than pointwise, convergence. Pointwise convergence occurs when we (in some cases quite nonchalantly) write $|f_n(t) - f(t)|$ instead of $\|f_n - f\|$, which implicitly assumes some specific point t where the condition holds.

If the inner sum in the left-hand side of (6.90a) converges uniformly with respect to n, we can swap the implied limits $N \to \infty$ and $M \to \infty$, thereby obtaining the right-hand side of (6.90a). This suggests that (6.89a) also might hold, in case the inner integral on either side of the equation is known to be uniformly converging. Here it's worth to first look at the case, which is intermediate between (6.89a) and (6.89b):

$$\int_0^d \int_0^{+\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^{+\infty} \int_0^d f(x,y)\,\mathrm{d}y\,\mathrm{d}x \tag{6.91a}$$

(only one of the two integrals being improper), and its discrete-sum counterpart

$$\sum_{n=0}^{N} \sum_{m=0}^{\infty} a_{mn} = \sum_{m=0}^{\infty} \sum_{n=0}^{N} a_{mn} \tag{6.91b}$$

Let

$$s(M,n) = \sum_{m=0}^{M} a_{mn}$$
$$S(M,N) = \sum_{n=0}^{N} \sum_{m=0}^{M} a_{mn} = \sum_{n=0}^{N} s(M,n)$$

Then the equation (6.91b) can be rewritten as

$$\sum_{n=0}^{N} \Big(\lim_{M\to\infty} s(M,n) \Big) = \lim_{M\to\infty} \sum_{n=0}^{N} s(M,n) = \lim_{M\to\infty} S(M,N)$$

In the left-hand side we are having simply a finite sum of several limits, and the formula must hold whenever the limits exist.

With (6.91a) there are, however, two limits involved. Indeed, let

$$F(b,y) = \int_0^b f(x,y)\,\mathrm{d}x$$
$$\Phi(b,d) = \int_0^d \int_0^b f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^d F(b,y)\,\mathrm{d}y$$

Then (6.91a) becomes

$$\int_0^d \left(\lim_{b\to+\infty} F(b,y)\right)\mathrm{d}y = \lim_{b\to+\infty}\int_0^d F(b,y)\,\mathrm{d}y = \lim_{b\to+\infty}\Phi(b,d) \tag{6.92}$$

or, re-notating b as x,

(integration under a limit sign)

$$\int_0^d \left(\lim_{x\to+\infty} F(x,y)\right)\mathrm{d}y = \lim_{x\to+\infty}\int_0^d F(x,y)\,\mathrm{d}y \tag{6.93}$$

So we swap the implicit limit $\Delta_y \to 0$ occurring in the integration and the explicit limit $x \to +\infty$. By *Moore–Osgood theorem*, in order to swap the limits, we need to know that the inner limit at $x \to +\infty$ in the left-hand side of (6.93) is converging uniformly with respect to y, that is the improper integral $\int_0^{+\infty} f(x,y)\,\mathrm{d}x$ is converging uniformly. Alternatively, we need to know that the inner integral in the right-hand side of (6.93) is converging uniformly (as its partition norm decreases to zero: $\Delta_y \to 0$) with respect to x.

▷ By intuitively understanding the integral in (6.93) as an infinite sum, we can see (6.93) as a generalization of the formula

$$\sum_{n=1}^{N}\left(\lim_{x\to+\infty} f_n(x)\right) = \lim_{x\to+\infty}\left(\sum_{n=1}^{N} f_n(x)\right)$$

▷ The equation (6.93) can be rewritten as

$$\int_0^d \left(\lim_{x\to+\infty}\frac{\partial}{\partial y}\Phi(x,y)\right)\mathrm{d}y = \lim_{x\to+\infty}\Phi(x,d)$$

Differentiating both sides with respect to d,

$$\lim_{x\to+\infty}\left(\frac{\partial}{\partial y}\Phi(x,y)\bigg|_{y=d}\right) = \frac{\partial}{\partial d}\lim_{x\to+\infty}\Phi(x,d)$$

or, re-notating d as y,

(differentiation under a limit sign)

$$\lim_{x\to+\infty}\frac{\partial\Phi}{\partial y}(x,y) = \frac{\partial}{\partial y}\lim_{x\to+\infty}\Phi(x,y) \tag{6.94}$$

Thus (6.94) is a slightly different, almost equivalent, form of (6.93).

Uniform convergence: both integrals are improper

Getting back to (6.89a) we can use (6.92) to rewrite its left-hand side as

$$\int_0^{+\infty}\int_0^{+\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \lim_{d\to+\infty}\int_0^d \left(\lim_{b\to+\infty} F(b,y)\right)\mathrm{d}y \overset{(6.92)}{=} \lim_{d\to+\infty}\lim_{b\to+\infty}\Phi(b,d)$$

and similarly, its right-hand side can be rewritten as

$$\int_0^{+\infty}\int_0^{+\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x \overset{(6.92)}{=} \lim_{b\to+\infty}\lim_{d\to+\infty}\Phi(b,d)$$

Thus, besides the applicability of (6.92) to each of the sides of (6.89a) (which e.g. would be more or less ensured if the inner of the two improper integrals converges uniformly), we need to make sure that we can swap the limits of $\Phi(b,d)$:

$$\lim_{d\to+\infty}\lim_{b\to+\infty}\Phi(b,d) = \lim_{d\to+\infty}\lim_{b\to+\infty}\Phi(b,d)$$

which may again require the application of the Moore–Osgood theorem.

Similar considerations can be applied to improper integrals at finite integration limits.

6.9 Average value of a function

6.9.1 Integral form in 1D

Suppose we want to compute the "arithmetic average value" of a function $f(t)$ on a given interval $[a,b]$. The problem is, obviously, that it's trivial to compute an arithmetic average of a finite set of values, but $f(t)$ has infinitely many values on $[a,b]$. Let's think how should we approach this problem.

Goal 6.11. Generalize the idea of the arithmetic average of a finite set of numbers to the arithmetic average of the values a function is taking on a given interval.

We could start by picking a finite number N of points τ_n regularly spaced on $[a,b]$ (Fig. 6.27) and computing their arithmetic average as an approximation of the average value of $f(t)$ on $[a,b]$:

$$\langle f\rangle_{[a,b]} \approx \frac{1}{N}\sum_{n=0}^{N-1} f(\tau_n) \quad \text{where } \tau_n = \frac{b-a}{N-1}n \tag{6.95}$$

and then letting $N\to\infty$:

$$\langle f\rangle_{[a,b]} = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} f(\tau_n) \quad \text{where } \tau_n = \frac{b-a}{N-1}n \tag{6.96}$$

It's not difficult to notice that (6.96) is essentially a Riemann integral of $f(t)$ over $[a,b]$. Indeed, the partitioning subintervals are $[\tau_n,\tau_{n+1}]$ with the respective function evaluation on each subinterval being done at the point $\tau_n\in[\tau_n,\tau_{n+1}]$. The first subinterval is $[\tau_0,\tau_1]$ where $\tau_0=a$, and the last interval is $[\tau_{N-2},\tau_{N-1}]$

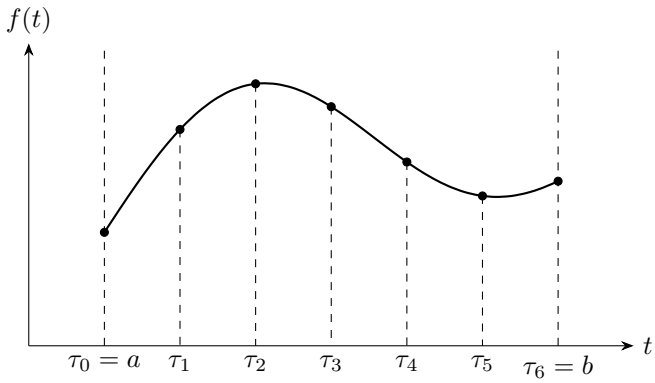


Figure 6.27: A selection of $N = 7$ regularly spaced points for an approximate computation of the function's average by (6.95).

where $\tau_{N-1} = b$. This means that we are having $N - 1$ intervals in total, corresponding to N going from 0 to $N - 2$. The last point τ_{N-1} doesn't have a corresponding interval (as one can see in Fig. 6.27). This is not a problem at all: we can discard the term at $n = N - 1$ from the sum in (6.96) without affecting the limit's value, since the contribution of each individual term approaches zero in the limit. We thereby obtain

$$\begin{aligned}\langle f\rangle_{[a,b]} &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} f(\tau_n) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-2} f(\tau_n) = \lim_{N\to\infty}\frac{1}{N-1}\sum_{n=0}^{N-2} f(\tau_n)\\ &= \lim_{N\to\infty}\sum_{n=0}^{N-2} f(\tau_n)\frac{\Delta t_n}{b-a} \qquad \text{where } \Delta t_n = \tau_{n+1} - \tau_n = \frac{b-a}{N-1}\end{aligned}$$

Or, re-notating $N - 1$ as N, which we can freely do under the limit $N \to \infty$,

$$\langle f\rangle_{[a,b]} = \lim_{N\to\infty}\sum_{n=0}^{N-1} f(\tau_n)\frac{\Delta t_n}{b-a} \qquad \text{where } \Delta t_n = \tau_{n+1} - \tau_n = \frac{b-a}{N}$$

and thus

$$\langle f\rangle_{[a,b]} = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} f(\tau_n) = \frac{1}{b-a}\lim_{N\to\infty}\sum_{n=0}^{N-1} f(\tau_n)\Delta t_n = \frac{1}{b-a}\int_a^b f(\tau)\,\mathrm{d}\tau$$

Thus we obtain

(arithmetic average of a function on an interval)

$$\langle f\rangle_{[a,b]} = \frac{1}{b-a}\int_a^b f(t)\,\mathrm{d}t \tag{6.97}$$

or, in a slightly different form

$$\int_a^b f(t)\,\mathrm{d}t = (b-a)\cdot\langle f\rangle_{[a,b]} \tag{6.98}$$

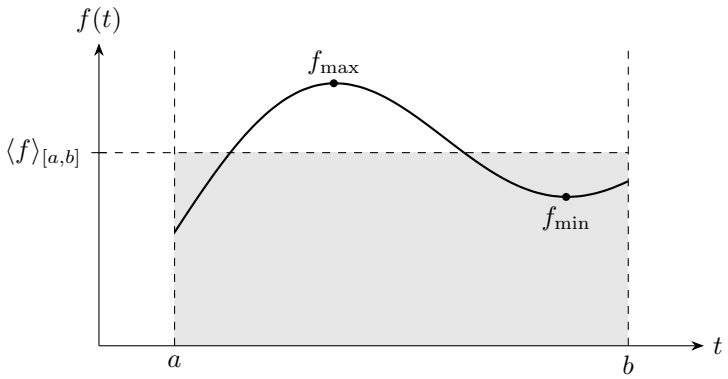


Figure 6.28: The area under a function is equal to the area of a rectangle of the height equal to the function's average. The average value will be between the function's maximum and minimum values.

Fig. 6.28 illustrates.

▷ It immediately follows that the formula (6.1) is indeed giving the arithmetic average of the flux on the interval in question. Indeed,

$$\beta(t, \Delta t) = \int_t^{t+\Delta t} I(\tau)\,\mathrm{d}\tau$$

and therefore

$$\frac{\beta(t, \Delta t)}{\Delta t} = \frac{1}{\Delta t}\int_t^{t+\Delta t} I(\tau)\,\mathrm{d}\tau$$

is the arithmetic average.

Suppose $f(t)$ is continuous on $[a, b]$. Then it should attain a minimum $f_{\min}$ and a maximum $f_{\max}$ somewhere on $[a, b]$ (possibly multiple times, but that doesn't matter). Apparently, the value of the integral $\int_a^b f(t)\,\mathrm{d}t$ must be somewhere between $(b-a)f_{\min}$ and $(b-a)f_{\max}$, which in case of $b-a>0$ can be formally written as

$$(b-a)f_{\min} \le \int_a^b f(t)\,\mathrm{d}t \le (b-a)f_{\max}$$

or, by using (6.98),

$$(b-a)f_{\min} \le (b-a)\langle f \rangle_{[a,b]} \le (b-a)f_{\max}$$

Dividing all parts of the inequality by $(b-a)$ yields

$$f_{\min} \le \langle f \rangle_{[a,b]} \le f_{\max} \tag{6.99a}$$

and we also obtain the same result if $b-a<0$. The case $b-a=0$ can be handled separately by letting $\langle f \rangle_{[a,a]} = f(a) = f_{\min} = f_{\max}$, so (6.99a) still holds.

Actually, unless the function is constant, we have $f_{\min} < f_{\max}$ and (6.99a) can be refined into

$$f_{\min} < \langle f \rangle_{[a,b]} < f_{\max} \tag{6.99b}$$

Indeed, it is clear that in this case $\langle f \rangle_{[a,b]}$ cannot be equal to $f_{\max}$, since, due to continuity of $f(t)$, any local deviation of $f(t)$ from $f_{\max}$ will occur at least on some finite subinterval of $[a, b]$ and thus the integral of $f(t)$ will be strictly less than $(b - a)f_{\max}$. The same argument goes for $f_{\min}$.

Now, since $f(t)$ is continuous, it should attain all values between $f_{\min}$ and $f_{\max}$ on (a, b), including $\langle f \rangle_{[a,b]}$. We thus obtain

(mean value theorem)

$$\Big(f \in C^0([a, b])\colon f(t) \in \mathbb{R}\Big) \implies \Big(\exists t_0 \in (a, b)\colon f(t_0) = \langle f \rangle_{[a,b]}\Big) \tag{6.100}$$

that is a real-valued function of a real argument, which is continuous on $[a, b]$, attains its "arithmetic average value" (computed over the interval $[a, b]$) somewhere inside this interval. Notice that this includes the case $f_{\min} = f_{\max}$, even though (6.99b) cannot be formally applied if $f_{\min} = f_{\max}$.

▷ The statement of the mean value theorem is also illustrated by Fig. 6.28 where there happen to be not just one, but two intersections of the function with the horizontal line at $\langle f \rangle_{[a,b]}$, that is this particular function attains its average value twice on $[a, b]$.

The requirement of the argument of $f(t)$ to be real is so that we can talk of an interval $[a, b]$ on which the function is being computed, but the theorem readily generalizes to complex- or vector-argument functions, where we are averaging over a connected area or, more generally, over a multidimensional connected hypervolume. The details will be covered in Subsection 6.9.3.

On the other hand, the requirement for $f(t)$ to be real-valued is essential. The theorem only holds for one-dimensional values of $f(t)$. In particular, it doesn't hold if the function value is an N-dimensional vector (where $N > 1$), or if the values $f(t)$ are complex. The latter can be demonstrated e.g. by considering the imaginary-argument exponential function $f(t) = e^{it}$ on $t \in [-\pi/2, \pi/2]$. Computing the mean value yields

$$\begin{aligned}\langle f \rangle_{[-\frac{\pi}{2},\frac{\pi}{2}]} &= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} e^{it}\,\mathrm{d}t = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} (\cos t + i\sin t)\,\mathrm{d}t = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \cos t\,\mathrm{d}t + 0 \\ &= \frac{1}{\pi}\sin t\Big|_{t=-\pi/2}^{\pi/2} = \frac{\sin\frac{\pi}{2} - \sin\left(-\frac{\pi}{2}\right)}{\pi} = \frac{2}{\pi}\end{aligned}$$

Fig. 6.29 illustrates. So in 2 dimensions (corresponding to the real and imaginary parts) the function doesn't have to go through the mean value, it could simply go "around" it.

▷ Notice that each of the real and imaginary parts of $f(t)$ in Fig. 6.29 does attain its mean value (so the 1D versions of the mean value theorems do hold), but they do so "at different times" (at different values of t), therefore the function $f(t)$ as a whole never attains its mean value.

Goal 6.11 completed. ■

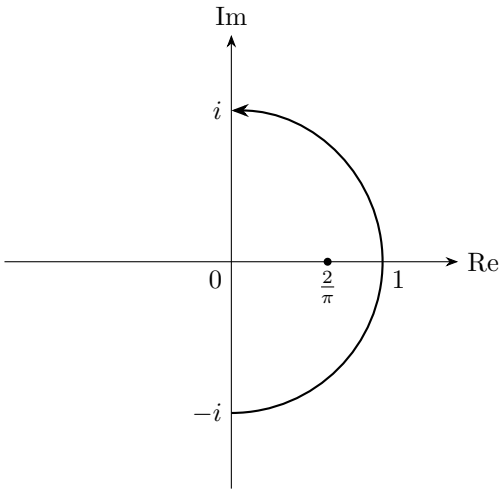


Figure 6.29: The trajectory of the complex-valued function e^{it} going from $-i$ to i for t going from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ and its average value $\frac{2}{\pi}$.

6.9.2 Differentiation-based form in 1D

The mean value theorem can be reformulated in terms of differentiation. Let $F(t)$ be an antiderivative of a continuous function $f(t)$. Then taking (6.97) and applying (6.25) we obtain another form of (6.97):

$$\langle f \rangle_{[a,b]} = \frac{F(b) - F(a)}{b - a}$$

or

$$\langle F' \rangle_{[a,b]} = \frac{F(b) - F(a)}{b - a}$$

By the mean value theorem, $\exists t_0 \colon f(t_0) = \langle f \rangle_{[a,b]}$. In other words, $\exists t_0 \colon F'(t_0) = \langle F' \rangle_{[a,b]}$. That is

$$\exists t_0 \colon F'(t_0) = \frac{F(b) - F(a)}{b - a}$$

and we obtain

(Lagrange theorem)

$$\Big(F \in C^1([a,b]) \colon F(t) \in \mathbb{R}\Big) \implies \left(\exists t_0 \in (a,b) \colon F'(t_0) = \frac{F(b) - F(a)}{b - a}\right)$$

which holds for continuously differentiable real-valued functions F of a real argument. Fig. 6.30 illustrates.

▷ Actually the theorem holds for slightly weaker conditions than that, but we won't get into those details.

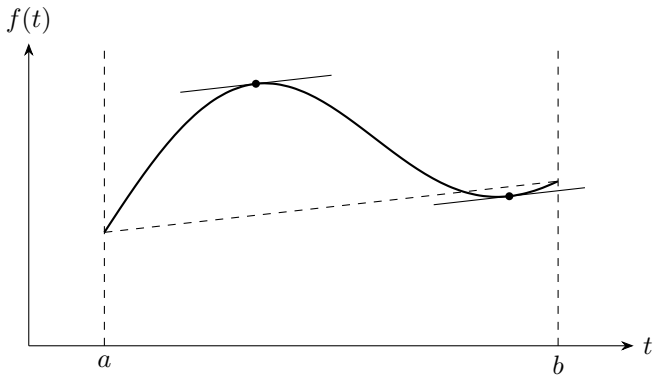


Figure 6.30: Graphical illustration of the Lagrange theorem: there is at least one (for the given function, actually, two) points where the tangent line to the function's graph is parallel to the secant line connecting $f(a)$ and $f(b)$.

6.9.3 Multiple dimensions

Goal 6.12. Generalize the mean value theorem to multidimensional case.

The arithmetic average formula (6.97) readily generalizes to multiple dimensions. The main difference is that instead of an interval of length $(b-a)$ we have e.g. a 2D shape with an area given by $S = \int_S \mathrm{d}S$, or a 3D shape with a volume given by $V = \int_V \mathrm{d}V$, etc. (where S stands both for the area of the shape and for the shape itself, and, likewise, V stands both for the volume of the shape and for the shape itself).

Respectively, for a generic shape, (6.97) takes the form

(arithmetic average of a function over the inner volume of a shape)

$$\langle f \rangle_V = \frac{1}{V} \int_V f(\mathbf{r})\,\mathrm{d}V = \frac{\displaystyle\int_V f(\mathbf{r})\,\mathrm{d}V}{\displaystyle\int_V \mathrm{d}V} \tag{6.101}$$

▷ Notice that the 1D formula (6.97) can be written in the form similar to (6.101) as

$$\langle f \rangle_{[a,b]} = \frac{1}{b-a} \int_a^b f(t)\,\mathrm{d}t = \frac{\displaystyle\int_a^b f(t)\,\mathrm{d}t}{\displaystyle\int_a^b \mathrm{d}t}$$

The mean value theorem in its integral form generalizes to a higher number of dimensions in an obvious way:

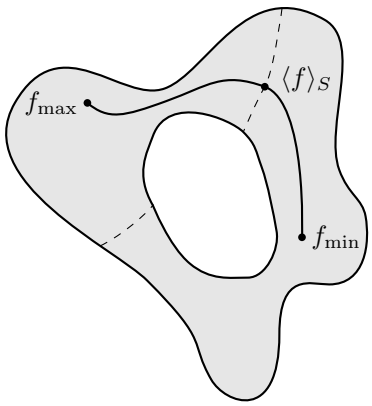


Figure 6.31: Graphical illustration of the mean value theorem for a 2D shape S. A continuous path connecting the maximum and minimum points of the function contains (at least one) point where the function attains its average value $\langle f \rangle_S$. The dashed lines are the level curves where the function value is equal to $\langle f \rangle_S$.

(mean value theorem in multiple dimensions)

$$\Big(f \in C^0(V)\colon f(\mathbf{r}) \in \mathbb{R}\Big) \implies \Big(\exists \mathbf{r}_0 \in V^\circ\colon f(\mathbf{r}_0) = \langle f \rangle_V\Big)$$

where V° denotes the interior of the shape V, and where we need to require that the shape V is compact and connected, since connected 1D intervals generalize to compact connected shapes. This will ensure that there is a continuous path lying within V connecting the maximum and minimum points of the function, and it follows, by the continuity of f, that there must be a point on this path where the function is equal $\langle f \rangle_V$. Fig. 6.31 illustrates.

▷ In 1D there was only one possible path connecting the minimum and maximum points of f on the interval of interest. In a higher number of dimensions there is generally an infinite number of paths, connecting given two points of the shape V. Each of these paths between the minimum and the maximum points will contain a point where the value of f is equal to $\langle f \rangle_V$. These mean value points should form a "barrier hypersurface" within the shape V making it impossible to go from a maximum point to a minimum point or back without crossing this barrier. In 2D this would be (at least one) barrier line consisting of mean value points, in 3D a barrier surface, etc.

Fig. 6.31 contains an example of such barrier lines (shown by dashed lines). In some cases (e.g. if the function is locally constant, staying at the value $\langle f \rangle_V$) these barrier lines or (hyper)surfaces can have a nonzero "thickness".

Goal 6.12 completed. ■

There is no straightforward differentiation-based counterpart to the multidimensional mean value theorem that we just formulated. In 1D both the function

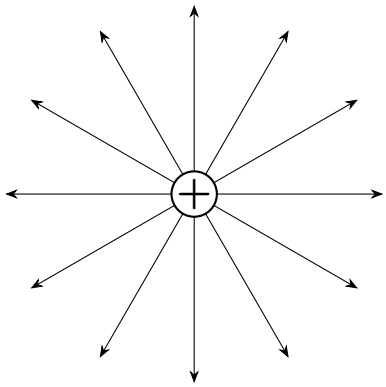


Figure 6.32: An electric charge and its field lines. In this 2D picture the 2D line density is falling off as $1/r$, in 3D the 3D line density would fall off as $1/r^2$.

and its derivative were real functions of a real argument. However if in multiple dimensions we have a real function of a vector argument, then its gradient, which is the generalization of the function's derivative to multiple dimensions, is a vector function of a vector argument, so we cannot directly apply the just obtained integral-form mean value theorem to the function's derivative, like we did in 1D to obtain the Lagrange theorem.

It would seem that, in order to apply the integral-form mean value theorem to the gradient, we'd need first to generalize the mean value theorem to vector functions of a vector argument. Except that this generalized version of the mean value theorem wouldn't hold in a general case (for the reasons similar to how the theorem doesn't hold for vector functions of a real argument). The whole situation becomes noticeably more complicated and we will leave it outside the scope of this book.

6.10 Boundary vs. interior behavior of functions

6.10.1 Gauss's theorem for electrostatic field

We now shall develop Gauss's theorem, using electrostatics to construct a reference motivation example.

A stationary electric charge generates a spherically symmetric electric field directed away from the charge (if the charge is positive), the field strength falling off as $1/r^2$, where r is the distance to the charge. It is convenient to think visually of this situation, as if there were a number of "field lines" going out of the charge equally in all directions (Fig. 6.32), the number of lines being directly proportional to the charge value q. For the purposes of the current discussion we can simply assume that the number of lines is equal to q, which is essentially just a visual convention.

▷ Obviously, the total number of lines doesn't have to be an integer, "in reality" the charge q is a real rather than integer quantity, we are just using a discrete set of lines to picture the field in a visual manner. There are no discrete field lines "in reality", the field is going through every point and not just through those through which the lines in Fig. 6.32 go.

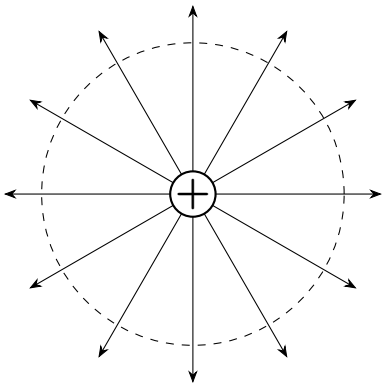


Figure 6.33: A charge surrounded by an imaginary sphere (a 2D picture of a 3D setup). The total field flux Φ through the sphere is equal to the charge q, regardless of the sphere's radius.

In such picture the density of the field lines (the number of lines per unit area orthogonal to the local direction of the lines) will fall off as $1/r^2$. Indeed, the points located at one and the same distance from the charge form a sphere (Fig. 6.33), the field lines are orthogonal to the surface of this sphere, while the area of a sphere is $4\pi r^2$. On the other hand, the total number of field lines crossing the sphere should be independent from the sphere's radius and always equal to one and the same value q. This suggests that the field line density is given by the ratio of the two numbers $q/4\pi r^2$. It would be convenient for the purposes of this discussion to associate the field line density with the field strength, where we can make both values simply equal to each other by the choice of units. We thereby obtain

$$E = \frac{q}{4\pi r^2} \tag{6.102}$$

where q is the total number of lines exiting the charge.

Multiplying the field line density E by the surface of the sphere we obtain the total number of lines going through the sphere:

$$\Phi = E \cdot 4\pi r^2 = q \tag{6.103}$$

This value Φ is referred to as the field *flux* through the sphere. Numerically it is equal to the charge q, but conceptually these are two different values.[6] Besides, flux doesn't have to be computed through a sphere, we can compute flux through any oriented surface or any fragment of an oriented surface (in other words, the surface doesn't need to be closed).

Goal 6.13. Construct a formula, which allows, given the field line density, to compute the flux of a field through a given surface.

Goal 6.13a. Assume a flat surface and a field which has identical density and direction everywhere.

[6] Normally the field flux is computed from the field strength rather than from the field line density, but since in this text both values are equal, we can define the flux in terms of the field line density.

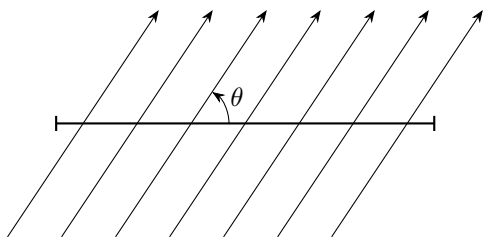


Figure 6.34: The number of field lines going through a unit area is proportional to $\sin\theta$ (for a given constant spatial density E).

The field line density is defined as the number of lines per unit area orthogonal to the local direction of the lines. So, if the surface in question is orthogonal to the field, the flux through this surface is simply equal to the product of density by the area

$$\Phi = E \cdot S$$

where S is the surface's area.

If the surface is not orthogonal to the field, the number of lines crossing the surface will become smaller (compared to the surface of the same area which is orthogonal to the field). It's not difficult to realize that it will be smaller exactly $\sin\theta$ times, where θ is the angle between the field direction and the surface (Fig. 6.34). The flux is thereby given by

$$\Phi = E \cdot S \cdot \sin\theta$$

In order to be able to define the angle between the field and the surface, we need to promote the field line density E to a vector $\vec{E}$, the direction of this vector coinciding with the direction of the field lines. We also need to describe the surface orientation, which is conveniently done by using a vector $\vec{n}$ normal to the surface. In this case the angle between $\vec{E}$ and $\vec{n}$ will be equal to $90° - \theta$ and we have

$$\Phi = E \cdot S \cdot \sin\theta = E \cdot S \cdot \cos(90° - \theta) = \vec{E} \cdot \vec{n} \cdot S$$

It is convenient to combine the product $\vec{n} \cdot S$ into a single vector $\vec{S}$, which is thereby normal to the surface and has the length equal to the surface area. We thus obtain

$$\Phi = \vec{E} \cdot \vec{S} \tag{6.104}$$

Notice that thereby flux is negative if the field $\vec{E}$ is crossing the surface in the direction opposite to the direction of the surface's normal.

Goal 6.13a completed. ■

Goal 6.13b. Generalize to surfaces of arbitrary shape.

An arbitrary-shape surface can be split into infinitely small flat surface elements $\mathrm{d}\vec{S}$. By (6.104),

$$\mathrm{d}\Phi = \vec{E} \cdot \mathrm{d}\vec{S}$$

Integrating across the entire surface yields

$$\Phi = \int_S \mathrm{d}\Phi = \int_S \vec{E} \cdot \mathrm{d}\vec{S} \tag{6.105}$$

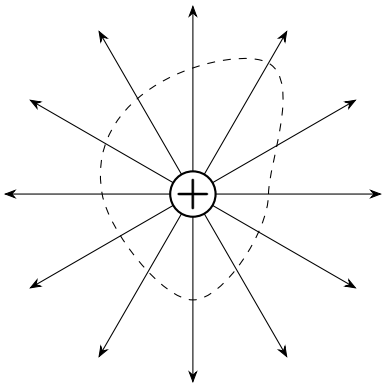


Figure 6.35: A charge surrounded by an imaginary closed surface (a 2D version of a 3D setup). The total field flux Φ through the surface is equal to the charge q, regardless of the size and the shape of the surface.

Goal 6.13b completed. ■

Goal 6.13 completed. ■

We can use (6.105) to generalize (6.103). Instead of computing the flux through a sphere centered at the charge, we could compute the flux through an arbitrary closed surface surrounding the charge (Fig. 6.35). The total number of lines exiting such surface should still be equal to q, therefore and

$$\Phi = \oint_S \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{S} = q \tag{6.106}$$

where

$$\vec{E}(\vec{r}) = \frac{q}{4\pi r^3}\vec{r} \tag{6.107}$$

The next thing we can notice is that $\mathrm{d}\Phi = \vec{E} \cdot \mathrm{d}\vec{S}$ is a signed value, which is negative if the field goes against the surface normal. So, in order to get $\Phi = q$ in (6.106), we need to let the surface normal be directed towards the outside, thereby the direction of $\mathrm{d}\vec{S}$ matching the direction of $\vec{E}$ (for $q > 0$). From that perspective, (6.106) is counting the number of field lines exiting the interior of the closed surface S. Indeed, it's the field lines exiting S, which generate positive values of $\mathrm{d}\Phi = \vec{E} \cdot \mathrm{d}\vec{S}$.

Conversely, the field lines entering S generate negative values of $\mathrm{d}\Phi = \vec{E} \cdot \mathrm{d}\vec{S}$. So, if some field lines are exiting S and some lines are entering S, the integral in (6.106) is giving the difference of the number of exiting lines and the number of entering lines. This ensures that (6.106) continues to work for more "wildly shaped" surfaces, where some field lines might be exiting the surface and then entering it again. However, ultimately each line will exit the surface, the number of exitings always being 1 greater than the number of enterings, therefore each line will be counted exactly once. Fig. 6.36 illustrates.

On the other hand, if the charge q is located outside of S, all lines that enter the surface's interior will also exit the surface's interior, the total number of exitings being equal to the total number of enterings, their difference being zero, and so being the value of the integral. Fig. 6.37 illustrates.

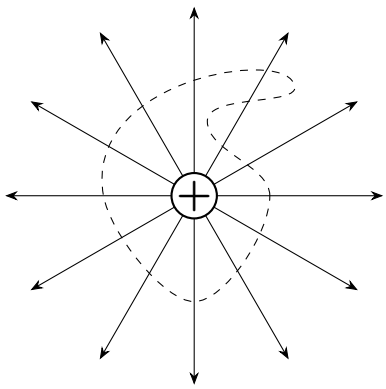


Figure 6.36: A charge inside a "wildly shaped" surface (a 2D version of a 3D setup). The total field flux through the surface is equal to the charge q, regardless of the size and shape of the surface.

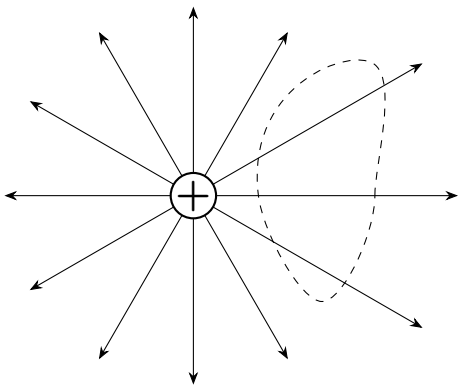


Figure 6.37: A charge outside a closed surface (a 2D version of a 3D setup). The total field flux through the surface is zero, regardless of the size and shape of the surface.

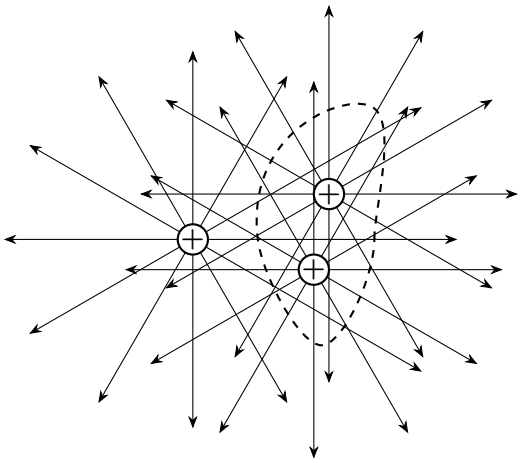


Figure 6.38: Multiple charges inside and outside the surface (a 2D version of a 3D setup). The total field flux through the surface is equal to the sum of charges q_n located inside the surface. The outside charges make zero contribution to the total flux.

The next thing that we shall bring into our model is the additivity of the electrostatic field: the (vector) strength of the field generated by several charges is equal to the sum of the strengths of the fields generated by each of these charges in isolation:

$$\vec{E}(\vec{r}) = \sum_n \frac{q_n}{4\pi \cdot |\vec{r} - \vec{r}_n|^3} (\vec{r} - \vec{r}_n) \tag{6.108}$$

▷ In (6.107) we were always assuming that the charge is located at $\vec{r} = 0$. Now that we have multiple charges located at different positions, we have to include their positions $\vec{r}_n$ explicitly into (6.108).

It immediately follows that (6.106) turns into

$$\Phi = \oint_S \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{S} = \sum_{\vec{r}_n \in S} q_n \tag{6.109}$$

where we only count the charges located inside S and ignore all the others, since the contributions of all other charges to the surface integral in (6.109) are zero. Thus, the right-hand side of (6.109) simply contains the total charge inside the surface S. Fig. 6.38 illustrates.

▷ In (6.109) we do not allow $\vec{r}_n$ to be on the boundary of S, each charge must be either fully inside or fully outside. A charge which is exactly on the boundary cannot be pointlike (unless something like Dirac delta function is employed) and must be approximated by a charge continuously spread over a small but finitely-sized volume.

As the next step, instead of assuming a number of discrete charges inside the surface, let's assume that the charge is continuously spread across the surface's interior and even beyond. The right-hand side of (6.109) thereby turns into the volume integral of the charge density ρ, yielding

(flux of electrostatic field through a closed surface)

$$\Phi = \oint_S \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{S} = \int_V \rho(\vec{r})\,\mathrm{d}V \tag{6.110}$$

where V denotes the interior volume of S.

Equation (6.110) is sometimes referred to as *Gauss's theorem*, but our derivation so far explicitly assumes an electrostatic field, or at least a field having similar properties. From the mathematical perspective, the Gauss's theorem needs to be applicable to any field. We shall address this question a bit later.

6.10.2 Divergence

The formula (6.110) can be used to test for the local presence of electric charge by measuring the field strength $\vec{E}$ across a closed surface and computing the flux. By taking progressively smaller and smaller surfaces we should be able to localize the charge better and better. In the limit of an infinitely small surface of an interior volume $\mathrm{d}V$ we shall get the flux $\mathrm{d}\Phi = \rho\,\mathrm{d}V$, respectively we should be able to compute ρ as

$$\rho = \frac{\mathrm{d}\Phi}{\mathrm{d}V} \tag{6.111}$$

Let's try to express the idea of (6.111) a bit more formally.

If the surface S is small in size, the volume integral in the right-hand side of (6.110) may be well approximated by $\rho(\vec{r}_0)\Delta V$, where ΔV is the interior volume of S and $\vec{r}_0$ is some point inside S. We therefore have

$$\Delta\Phi = \oint_S \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{S} \approx \rho(\vec{r}_0)\Delta V \tag{6.112a}$$

where $\Delta\Phi$ denotes the field flux though the boundary S of the small volume ΔV. Respectively

$$\rho(\vec{r}_0) \approx \frac{\Delta\Phi}{\Delta V} = \frac{1}{\Delta V}\oint_S \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{S} \tag{6.112b}$$

This approximation becomes exact in the limit $\Delta V \to 0$. Or, actually, not $\Delta V \to 0$, because this still would allow the surface to have finitely large linear dimension(s), but $|S| \to 0$, where $|S|$ is the diameter of the surface (the largest distance between the points of the surface):

$$\rho(\vec{r}_0) = \lim_{|S|\to 0} \frac{\Delta\Phi}{\Delta V} = \lim_{|S|\to 0} \frac{1}{\Delta V}\oint_S \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{S} \tag{6.113}$$

The right-hand side of (6.113) is referred to as *divergence* of the vector field $\vec{E}$ at the point $\vec{r}_0$:

(divergence of a vector field $\vec{F}$)

$$\operatorname{div} \vec{F}(\vec{r}_0) = \lim_{|S|\to 0} \frac{1}{\Delta V}\oint_S \vec{F}(\vec{r}) \cdot \mathrm{d}\vec{S} \tag{6.114}$$

where $|S|$ is the 3D diameter of a surface surrounding $\vec{r}_0$, and ΔV is the volume of the surface's interior.

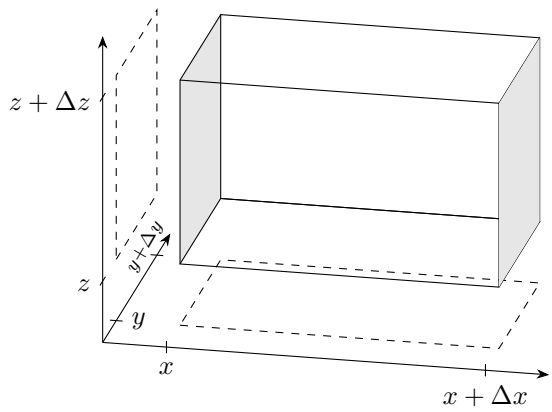


Figure 6.39: A small rectangular box-shaped volume element (6.116). The two filled sides are the ones orthogonal to the x axis.

We therefore obtain

(divergence of field strength gives the field charge density)

$$\rho(\vec{r}) = \operatorname{div} \vec{E}(\vec{r}) \tag{6.115}$$

▷ The formula (6.115) is sometimes (especially in a physical context) referred to as the differential form of the Gauss's theorem (while (6.110) is sometimes referred to as the integral form of the Gauss's theorem). This terminology however implicitly assumes that the divergence is computed by the formula (6.118), which we still have to derive.

The formula (6.115) formally establishes a relationship between $\vec{E}$ and ρ, however technically we need to use (6.114) to compute the divergence. The formula (6.114) can be computed numerically (although, with a somewhat questionable efficiency), but analytically it's pretty unwieldy. We might therefore ask, if there is a simpler analytical expression for the right-hand side of (6.114).

Goal 6.14. Derive a simple formula expressing the charge density ρ in terms of the field strength $\vec{E}$. *Hint:* apply (6.112) to an infinitely small rectangular volume.

In (6.112) we can choose the small volume ΔV to be a rectangular box (Fig. 6.39), the box's edges being parallel to the coordinate axes, so that the box itself is represented by

$$[x, x + \Delta x] \times [y, y + \Delta y] \times [z, z + \Delta z] \tag{6.116a}$$

whereas the box's volume ΔV is given by

$$\Delta V = \Delta x \Delta y \Delta z \tag{6.116b}$$

Computing the flux of the field through this box's boundary S, we have

$$\Delta\Phi = \oint_S \vec{E} \cdot \mathrm{d}\vec{S} = \oint_S \left(E_x \,\mathrm{d}S_x + E_y \,\mathrm{d}S_y + E_z \,\mathrm{d}S_z\right)$$

$$= \oint_S E_x \, \mathrm{d}S_x + \oint_S E_y \, \mathrm{d}S_y + \oint_S E_z \, \mathrm{d}S_z = \Delta\Phi_x + \Delta\Phi_y + \Delta\Phi_z$$

where we expanded the dot product $\vec{E} \cdot \mathrm{d}\vec{S}$ in terms of the coordinates of the respective two vectors, obtaining three different additive components $\Delta\Phi_x$, $\Delta\Phi_y$, $\Delta\Phi_z$ of the total flux $\Delta\Phi$.

Let's first concentrate on $\Delta\Phi_x$. The surface of the rectangular box volume (6.116) consists of 6 different sides, but 4 of the 6 sides are parallel to the x axis, respectively $\mathrm{d}S_x \equiv 0$ everywhere within these 4 sides. Thus, only the 2 sides orthogonal to the x axis (the one at x and the other at $x + \Delta x$) contribute to $\Delta\Phi_x$, where we have $|\mathrm{d}S_x| = |\mathrm{d}y \, \mathrm{d}z|$:

$$\Delta\Phi_x = \int_y^{y+\Delta y} \int_z^{z+\Delta z} E_x(x + \Delta x, \upsilon, \zeta) \, \mathrm{d}\zeta \, \mathrm{d}\upsilon - \int_y^{y+\Delta y} \int_z^{z+\Delta z} E_x(x, \upsilon, \zeta) \, \mathrm{d}\zeta \, \mathrm{d}\upsilon$$

Here $E_x = E_x(x, y, z)$ is the x component of $\vec{E}$ written as a function of the 3 coordinates, and the second integral takes the opposite sign because the box's side at x is looking into the direction opposite to the x axis (in Fig. 6.39 both of these sides are highlighted). Continuing, we obtain

$$\Delta\Phi_x = \int_y^{y+\Delta y} \int_z^{z+\Delta z} \big(E_x(x + \Delta x, \upsilon, \zeta) - E_x(x, \upsilon, \zeta)\big) \, \mathrm{d}\zeta \, \mathrm{d}\upsilon$$

In the limit at $|S| \to 0$ we obtain

$$\begin{aligned}
\lim_{|S|\to 0} \frac{\Delta\Phi_x}{\Delta V} &= \lim_{\substack{\Delta x\to 0\\ \Delta y\to 0\\ \Delta z\to 0}} \frac{\Delta\Phi_x}{\Delta V} = \\
&= \lim_{\substack{\Delta x\to 0\\ \Delta y\to 0\\ \Delta z\to 0}} \frac{1}{\Delta x \Delta y \Delta z} \int_y^{y+\Delta y} \int_z^{z+\Delta z} \big(E_x(x + \Delta x, \upsilon, \zeta) - E_x(x, \upsilon, \zeta)\big) \, \mathrm{d}\zeta \, \mathrm{d}\upsilon \\
&= \lim_{\substack{\Delta x\to 0\\ \Delta y\to 0\\ \Delta z\to 0}} \frac{1}{\Delta y \Delta z} \int_y^{y+\Delta y} \int_z^{z+\Delta z} \frac{E_x(x + \Delta x, \upsilon, \zeta) - E_x(x, \upsilon, \zeta)}{\Delta x} \, \mathrm{d}\zeta \, \mathrm{d}\upsilon \\
&= \lim_{\substack{\Delta y\to 0\\ \Delta z\to 0}} \frac{1}{\Delta y \Delta z} \int_y^{y+\Delta y} \int_z^{z+\Delta z} \left(\frac{\partial}{\partial x} E_x(x, \upsilon, \zeta)\right) \mathrm{d}\zeta \, \mathrm{d}\upsilon = \frac{\partial}{\partial x} E_x(x, y, z)
\end{aligned}$$

Obtaining similar results for $\Delta\Phi_y$ and $\Delta\Phi_z$ we have

$$\lim_{|S|\to 0} \frac{\Delta\Phi}{\Delta V} = \lim_{|S|\to 0} \frac{\Delta\Phi_x}{\Delta V} + \lim_{|S|\to 0} \frac{\Delta\Phi_y}{\Delta V} + \lim_{|S|\to 0} \frac{\Delta\Phi_z}{\Delta V} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}$$

By (6.113) we obtain

$$\rho = \lim_{|S|\to 0} \frac{\Delta\Phi}{\Delta V} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \overset{(6.115)}{=} \operatorname{div} \vec{E} \tag{6.117}$$

We thus found that $\operatorname{div} \vec{E}$ is given by the sum of partial derivatives in (6.117), which is much simpler than formula suggested by the formal definition (6.114)

of the divergence. The sum of the partial derivatives in (6.117) can be formally written as a dot product of the nabla operator (6.69) with the vector $\vec{E}$ and we obtain

(divergence formula)

$$\operatorname{div} \vec{F} = \nabla \cdot \vec{F} = \sum_{n=1}^{N} \frac{\partial F_n}{\partial r_n} \tag{6.118}$$

for the divergence of a vector field $\vec{F}(\vec{r})$ in N dimensions.

▷ Some texts may use (6.118) as the definition of divergence. In this book we prefer (6.114) as the definition, since it has a much shorter and much more intuitive motivational path compared to the formula $\nabla \cdot \vec{F}$ which needs a relatively long justification.

The readers might also notice that (6.118) is written for an N-dimensional space, while the definition (6.114) is assuming a 3-dimensional space. It is however not difficult to realize, that (6.114) is, in principle, generalizable to any number of dimensions, which we shall do shortly.

Goal 6.14 completed. ■

Substituting (6.117) into (6.110) we obtain

(Gauss's theorem)

$$\oint_S \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{S} = \int_V (\nabla \cdot \vec{E}(\vec{r}))\, \mathrm{d}V \tag{6.119}$$

that is the flux of an electrostatic field through a surface is equal to the volume integral of the field's divergence over the surface's interior. It turns out that the formula (6.119) is applicable to pretty much any continuously differentiable field (even some discontinuities can be allowed), not just electrostatic field, but we cannot be sure of that yet, and that will be the subject of Subsection 6.10.3

▷ The dot products in the left and right parts of (6.119) should be invariant under basis changes, and so must be the formula (6.119). However, one has to be careful with the "∇" operator, which (as we might remember from Section 4.14) has a different expression in nonorthonormal bases, as well as with $\mathrm{d}\vec{S}$, which is a pseudovector.

▷ Sometimes the boundary of the volume V is denoted as ∂V, in which case (6.119) may be rewritten as

$$\oint_{\partial V} \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{S} = \int_V (\nabla \cdot \vec{E}(\vec{r}))\, \mathrm{d}V$$

6.10.3 Gauss's theorem for arbitrary fields

So far in the Gauss's theorem (6.119) we are still assuming that the field in question is an electrostatic field, or a field with a similar behavior. Specifically,

we are assuming that the field is generated by the field charges, the field strength of a field generated by a charge being directed away from the charge and falling off as a squared distance. But we cannot be sure if an arbitrary vector field can be described in such terms.

▷ Actually it cannot. In order to be describable in such terms, the field must be *conservative*, that is the integral of field strength over any closed path must be zero.

On the other hand, in our derivation of the divergence formula (6.118) we didn't take any specific assumptions about the field while computing the field's flux over the boundary of an infinitely small rectangular volume $\mathrm{d}V$. So, the idea of divergence must be applicable to pretty much any differentiable field $\vec{F}$, returning something which "behaves as a field charge density" in some sense. At least $\mathrm{d}\Phi = \operatorname{div}\vec{F}\cdot\mathrm{d}V$ should give a correct value of the flux through the boundary of this infinitely small rectangular volume. It therefore makes sense to ask, whether this charge density-like quantity can be used in the "Gauss's theorem for electrostatic field" formula (6.110) and respectively in the Gauss's theorem formula (6.119) in case of an arbitrary field. Indeed, if $\mathrm{d}\Phi = \operatorname{div}\vec{F}\cdot\mathrm{d}V$, then by formally integrating $\mathrm{d}\Phi$ over the entire volume we should obtain

$$\Phi \stackrel{?}{=} \int_V \mathrm{d}\Phi = \int_V \operatorname{div}\vec{F}\cdot\mathrm{d}V \tag{6.120}$$

Except that we don't know if Φ is an additive value in the sense of the above formula, that is whether the flux through the boundary of V can be obtained by summing up the infinitely small fluxes $\mathrm{d}\Phi$ through the boundaries of the infinitely small volume elements $\mathrm{d}V$. Before we try to answer the latter question, let's however look at a few generalizations of the Gauss's theorem from 3D to other dimension counts.

Goal 6.15. Generalize Gauss's theorem for the electrostatic field to other dimension counts. Make sure to consider both the higher- and the lower-dimension counts. *Hint:* generalize the physical model of the electrostatic field to other dimension counts.

In order to generalize Gauss's theorem to other dimension counts we need to generalize the behavior of the physical 3D electrostatic field to fields in other dimensions. The key property of the electrostatic field that we used in the derivation of the Gauss's theorem was that the field lines exiting the charge go indefinitely away from the charge along straight trajectories. The field strength is thereby given by a modified version of (6.107):

$$\vec{E} = \frac{q}{k_N\cdot r^N}\vec{r} \tag{6.121}$$

where N is the number of dimensions and k_N is some proportionality coefficient, the specific value of which doesn't matter.[7] Respectively the "Gauss's theorem for the electrostatic field" formula (6.110) applies in any dimension count (for a suitable field $\vec{E}$ which can be described in the above terms), where the volume integral is replaced by a hypervolume integral and the surface integral is replaced by a hypersurface integral over the hypervolume's boundary.

[7] Apparently, k_N is the hyperarea of the $(N-1)$-dimensional unit hypersphere (embedded into the N-dimensional space). Therefore $k_3 = 4\pi$, $k_2 = 2\pi$.

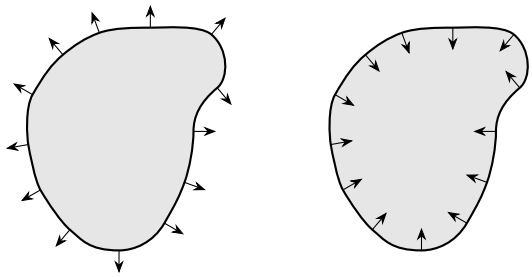


Figure 6.40: The normals to the boundary of a positively-oriented area (left) are pointing outwards. The normals to the boundary of a negatively-oriented area (right) are pointing inwards.

Apparently, the derivation of the divergence formula (6.118) holds in any number of dimensions, we just use an N-dimensional rectangular box in the derivation instead of a 3D box, the sides of this box being $(N-1)$-dimensional hypersurfaces. Respectively we still have $\rho = \operatorname{div} \vec{E}$ and therefore the Gauss's theorem formula (6.119) applies to any dimension count as well.

In particular, in 2 dimensions we obtain an area integral in the right-hand side and a closed contour integral in the left-hand side:

(Gauss's theorem in 2D)

$$\oint_C \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{C} = \int_S (\nabla \cdot \vec{E}(\vec{r}))\, \mathrm{d}S \tag{6.122a}$$

where C is the boundary curve of the area S and $\mathrm{d}\vec{C}$ is a vector normal to the curve, pointing "to the outside of S", the length of $\mathrm{d}\vec{C}$ being equal to the respective length differential of the curve C (pretty much the same like in the 3D Gauss's theorem $\mathrm{d}\vec{S}$ is normal to the surface and its length is being equal to the area differential).

▷ We kind of agreed in the context of Gauss's theorem to let the normals of the volume's boundary point outwards, however we can also recall our discussion of the area orientation in 2D. In particular, we could flip the orientation of the area S in (6.122a), which changes the sign of the integral in the right-hand side and thus the integral in the left-hand side must change its sign as well in order to preserve the equality (6.122a). A natural way to achieve this is to say that flipping the orientation of S also flips the orientation of C, thus S is an oriented 2D area and C is an oriented curve in 2D (where the curve's orientation affects the direction of the vector $\mathrm{d}\vec{C}$). Fig. 6.40 illustrates. So the normals of a positively oriented area's boundary are pointing outwards, while the normals of a negatively oriented area's boundary are pointing inwards.

Similar considerations apply in higher dimension counts, where we have an oriented volume or an oriented hypervolume with an oriented boundary surface or hypersurface. In a way, Gauss's theorem establishes the convention of the outward orientation of the normals of a positively oriented hypervolume's boundary.

In 1 dimension the vector field $\vec{E}(\vec{r})$ turns into a scalar field $E_x(x)$ and the divergence of $\vec{E}$ turns into the ordinary derivative $E'_x(x)$. The hypervolume V

becomes a 1D interval $[a, b]$, and the volume integral in the right-hand side of (6.119) becomes $\int_a^b E'_x(x)\,\mathrm{d}x$. The boundary of $[a, b]$ consists of the two points a and b and the surface integral in the left-hand side of (6.119) simply becomes $E_x(b) - E_x(a)$. The value $E_x(b)$ represents "the flux through the endpoint b of the interval" and $-E_x(a)$ represents "the flux through the endpoint a of the interval (thus the boundary point b is considered "positively oriented", and the boundary point a is considered "negatively oriented"). Thereby the 1D version of Gauss's theorem is

(Gauss's theorem in 1D)

$$E_x(b) - E_x(a) = \int_a^b \frac{\mathrm{d}}{\mathrm{d}x} E_x(x)\,\mathrm{d}x \tag{6.122b}$$

which is simply the fundamental theorem of calculus (6.25) and thereby holds for pretty much any function $E_x(x)$.

The 1D Gauss's theorem version (6.122b) takes a slightly more complicated form on a curve in 2D or in a higher-dimensional space, where we want to express the difference $\varphi(\vec{r}_1) - \varphi(\vec{r}_0)$ of the values of the function $\varphi(\vec{r})$ at the endpoints $\vec{r}_0$ and $\vec{r}_1$ of an (open) curve C via an integral of the function's derivative. The generalization of the derivative to scalar functions of vector arguments is represented by the gradient. So we want to integrate the gradient $\nabla\varphi$ with respect to $\mathrm{d}\vec{r}$. By (6.71)

$$\varphi(\vec{r}_1) - \varphi(\vec{r}_0) = \int_{\vec{r}=\vec{r}_0}^{\vec{r}_1} \mathrm{d}\varphi(\vec{r}) \overset{(6.71)}{=} \int_{\vec{r}_0}^{\vec{r}_1} \nabla\varphi \cdot \mathrm{d}\vec{r}$$

where the integration from $\vec{r}_0$ to $\vec{r}_1$ is assumed to be done along the curve C. This yields a 1D version of Gauss's theorem for a curve in a higher-dimensional space:

(generalization of Gauss's theorem for a curve)

$$\varphi(\vec{r}_1) - \varphi(\vec{r}_0) = \int_{\vec{r}_0}^{\vec{r}_1} \nabla\varphi \cdot \mathrm{d}\vec{r} \tag{6.122c}$$

where the integral in the right-hand side is computed over some curve going from $\vec{r}_0$ to $\vec{r}_1$.

Goal 6.15 completed. ■

At this point we shall discuss the applicability of (6.120) to arbitrary fields and respectively the applicability of Gauss's theorem to arbitrary fields.

Goal 6.16. Study the flux additivity (6.120) in case of arbitrary fields. *Hint:* use lower dimension counts and finite additivity as simplified cases.

If a volume V is split into a number of subvolumes V_n then the charge contributions from all these subvolumes add up to the total charge inside V. So, in order for the Gauss's theorem to hold, the fluxes though the respective boundary surfaces ∂V_n should add up to the total flux through ∂V as well. In the limit, where V_n become infinitely small, we shall obtain the situation described

by (6.120). Thus, by demonstrating the finite additivity of subvolume fluxes, we would also obtain the statement of (6.120) in the limit.

It would be easier to start in lower dimension counts. So, let's start in 1D, where the Gauss's theorem reduces to the fundamental theorem of calculus. Suppose we are given an interval $[a, b]$ and we split it into two intervals $[a, c]$ and $[c, b]$. On one hand

$$\int_a^b f(t)\,\mathrm{d}t = \int_a^c f(t)\,\mathrm{d}t + \int_c^b f(t)\,\mathrm{d}t$$

representing the right-hand side of the Gauss's theorem (the "integral of the charge density"). On the other hand,

$$\begin{aligned}\int_a^c f(t)\,\mathrm{d}t + \int_c^b f(t)\,\mathrm{d}t &= \big(F(c) - F(a)\big) + \big(F(b) - F(c)\big)\\ &= F(b) - F(a) = \int_a^b f(t)\,\mathrm{d}t\end{aligned}$$

where $f(t)$ stands for the divergence and the differences $F(x) - F(y)$ stand for the "total fluxes" through the boundaries of each respective interval. We notice that the "fluxes" $F(c)$ and $-F(c)$ through the "shared boundary" c are cancelling each other, where this boundary works once as positively oriented and once as negatively oriented.

More generally, if we introduce a number of points $a < c_1 < \ldots < c_N < b$, we would have

$$\begin{aligned}\int_a^{c_1} f(t)\,\mathrm{d}t + \ldots + \int_{c_N}^a f(t)\,\mathrm{d}t &= \big(F(c_1) - F(a)\big) + \ldots + \big(F(b) - F(c_N)\big)\\ &= F(b) - F(a) = \int_a^b f(t)\,\mathrm{d}t\end{aligned}$$

where all "shared boundary" terms $F(c_n)$ will be cancelled by the terms $-F(c_n)$.

Now we can generalize this to 2D. Suppose we are integrating the "charge density" over an area S (with its respective boundary C), which we split into S_1 and S_2 (Fig. 6.41). Adding up the fluxes through the respective boundaries C_1 and C_2, we are going to integrate over the shared boundary $C_1 \cap C_2$ twice. However the effective orientation of this shared part will be opposite, depending on whether we "arrive from the interior of S_1", or "arrive from the interior of S_2". We could loosely describe this as $C_2 \cap C_1 = -(C_1 \cap C_2)$, where the minus sign describes the orientation inversion. We therefore have

$$\begin{aligned}\int_S (\nabla\cdot\vec{F})\,\mathrm{d}\vec{S} &= \int_{S_1} (\nabla\cdot\vec{F})\,\mathrm{d}\vec{S} + \int_{S_2} (\nabla\cdot\vec{F})\,\mathrm{d}\vec{S} = \int_{C_1} \vec{F}\cdot\mathrm{d}\vec{C} + \int_{C_2} \vec{F}\cdot\mathrm{d}\vec{C}\\ &= \left(\int_{C_1\setminus C_2} \vec{F}\cdot\mathrm{d}\vec{C} + \int_{C_1\cap C_2} \vec{F}\cdot\mathrm{d}\vec{C}\right) + \left(\int_{C_2\setminus C_1} \vec{F}\cdot\mathrm{d}\vec{C} + \int_{C_2\cap C_1} \vec{F}\cdot\mathrm{d}\vec{C}\right)\\ &= \left(\int_{C_1\setminus C_2} \vec{F}\cdot\mathrm{d}\vec{C} + \int_{C_1\cap C_2} \vec{F}\cdot\mathrm{d}\vec{C}\right) + \left(\int_{C_2\setminus C_1} \vec{F}\cdot\mathrm{d}\vec{C} - \int_{C_1\cap C_2} \vec{F}\cdot\mathrm{d}\vec{C}\right)\\ &= \int_{C_1\setminus C_2} \vec{F}\cdot\mathrm{d}\vec{C} + \int_{C_2\setminus C_1} \vec{F}\cdot\mathrm{d}\vec{C}\end{aligned}$$

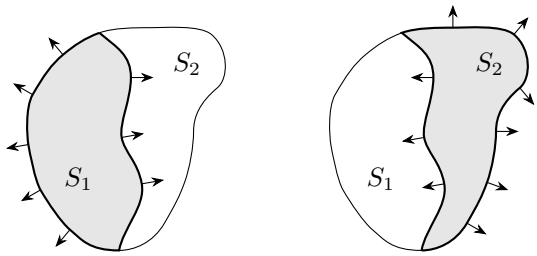


Figure 6.41: A positively-oriented 2D area split into two subareas S_1 and S_2 (which are thereby also positively oriented). The normals to the shared boundary of the two subareas are pointing in two opposite directions, depending on for which of the two subareas we want to apply the Gauss's theorem.

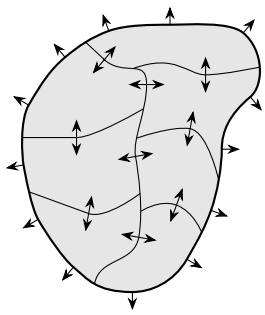


Figure 6.42: All inner fluxes through the shared boundaries cancel out and we are being left just with the flux through the outer boundary.

where the two fluxes through $C_1 \cap C_2$ cancel each other, since they are done with respect to the same surface but with opposite orientation.

Let's summarize it once again: when we compute the sum of fluxes though C_1 and C_2, the flux through the shared part $C_1 \cap C_2$ is included twice but with opposite signs and therefore cancels itself, thereby we are left with just the flux through C. Now if we split S into a large number of subareas S_n, the fluxes through the shared boundaries between S_n will cancel themselves in a similar manner and we will be left just with the flux through C (Fig. 6.42).

In 3D we observe a similar picture. Splitting the volume V into a large number of subvolumes V_n, we'll have the fluxes through the shared boundary surfaces of V_n cancelling themselves, thereby being left just with the flux through V. Higher-dimension generalizations are totally similar. Thus we obtain the finite additivity of the flux in the context of the Gauss's theorem in any dimension count. By taking the limit we obtain (6.120):

(additivity of the field flux across the surface's interior)

$$\Phi = \int_V \mathrm{d}\Phi = \int_V (\nabla \cdot \vec{F}) \cdot \mathrm{d}V \tag{6.123}$$

which now holds for pretty much any differentiable field (where the field additionally must be continuously differentiable almost everywhere in order for the

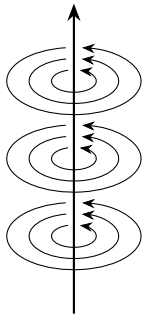


Figure 6.43: Magnetic field of an infinitely long straight wire. The vertical straight line represents the current I flowing through the wire, the circular lines are the field lines of magnetic field $\vec{B}$.

respective integrals to exist).

Goal 6.16 completed. ■

It follows from (6.123) that Gauss's theorem holds for pretty much any differentiable field (with the mentioned additional continuous differentiability requirements).

6.10.4 Green's theorem

Instead of electric field of a static charge, let's consider the magnetic field generated by a steady current I flowing through an infinitely long straight wire (Fig. 6.43).

The magnetic field will have a cylindrical symmetry with respect to rotations around the wire as an axis, rather than spherical symmetry, since this time instead of a point charge we are having an infinitely long wire. The field lines of the magnetic fields will be closed circles centered around the wire and *lying orthogonally to the wire*. The circle length $2\pi r$ grows proportionally to the radius, therefore so does the cylinder's area. Respectively, the absolute value of the field strength B falls off inversely proportionally to the radius:

$$B = \frac{I}{2\pi r} \tag{6.124}$$

where for simplicity we choose the units in such as way that we simply divide by the circle's length.

> ▷ As with electric field, we could associate the density of magnetic lines with the field strength, but, at least in this particular situation, where the field lines are orthogonal to the direction from the wire rather than directed away from the wire, it doesn't have the same amount of visual intuitivity.

So instead of a $1/r^2$ rolloff we are having a $1/r$ rolloff, which can be attributed to the picture having a cylindrical symmetry instead of a spherical symmetry. This bears a strong similarity to a "2D electrostatic field", which by (6.121) would be falling off as $1/r$. At the same time, we could notice that our picture is invariant under displacements along the wire, due to the wire being infinitely

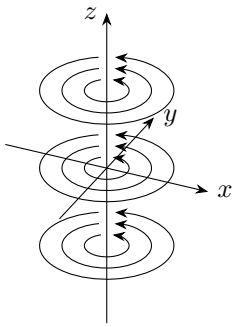


Figure 6.44: Choice of the coordinate system.

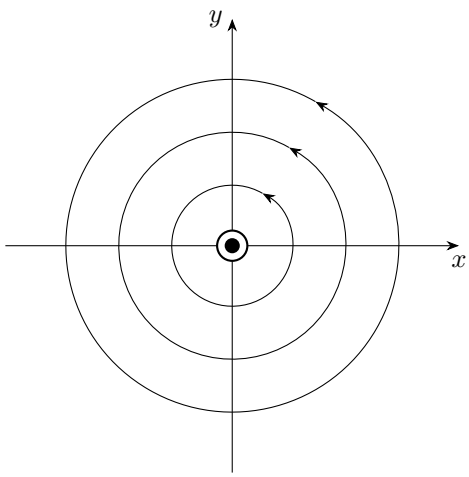


Figure 6.45: 2D version of Fig. 6.44. Thin circular lines are magnetic field lines. The dot in the center denotes the current flowing outwards from the page.

long (therefore having no preferred points on itself). This suggests that one of the dimensions can be "thrown out" and we could consider a purely 2D picture instead. In this case the similarity to the "2D electrostatic field" would be almost complete, except that the field lines are having a direction different by 90°. We could therefore expect the 2D version of Gauss's theorem to hold, provided we perform the adjustments accounting for a different field line direction.

Goal 6.17. Utilize the similarity of the setup in Fig. 6.43 to a "2D electrostatic field", in order to derive a Gauss's theorem counterpart.

Let's introduce a coordinate system where the wire coincides with the z axis, the direction of the z axis matching the direction of the current (Fig. 6.44). This picture is fully invariant under z coordinate translations, therefore we can discard the z coordinate and consider the 2D version shown in Fig. 6.45 instead.

The formula (6.124) looks exactly like a 2D version of (6.102), which would be

$$E = \frac{q}{2\pi r} \tag{6.125}$$

Comparing (6.125) and (6.124) we notice that by letting $q = I$ and $E = B$ in (6.125) we could apply (6.125) to our current magnetostatic setup.

This in turn suggests that all argument made earlier for the Gauss's theorem (in application to 2D) must be somehow applicable as well, if we account for the direction difference between $\vec{E}$ and $\vec{B}$ (which is the only remaining difference to the electrostatic setup). Let's start with (6.109), which in a 2D context takes the form

$$\Phi = \oint_C \vec{E}(\vec{r}) \cdot \mathrm{d}\vec{C} = \sum_{\vec{r}_n \in C} q_n \tag{6.126}$$

where $\mathrm{d}\vec{C}$ are normal to the contour C. Apparently instead of charges q_n we can take currents I_n. As the vector $\vec{E}$ goes, its absolute magnitude is supposed to be equal to the one of $\vec{B}$, but the directions differ by 90°, since $\vec{E}$ points away from the center, while $\vec{B}$ circulates counterclockwise around the center. We can express this as $\vec{B} = \vec{R}(\vec{E})$ and $\vec{E} = \vec{R}^{-1}(\vec{B})$, where $\vec{R}(\vec{\cdot})$ is a 90° counterclockwise rotation operator, respectively $\vec{R}^{-1}(\vec{\cdot})$ is a 90° clockwise rotation operator:

$$\oint_S \vec{R}^{-1}(\vec{B}) \cdot \mathrm{d}\vec{C} = \sum_{\vec{r}_n \in C} I_n \tag{6.127}$$

(where $\vec{r}_n$ are the positions of the currents I_n on the 2D plane). Now under the integral sign in the left-hand side of (6.127) we are having a dot product $\vec{R}^{-1}(\vec{B}) \cdot \mathrm{d}\vec{C}$. This dot product would be unchanged if we rotate both vectors 90° counterclockwise:

$$\oint_C \vec{B} \cdot \vec{R}(\mathrm{d}\vec{C}) = \sum_{\vec{r}_n \in C} I_n$$

We can notice that $\vec{R}(\mathrm{d}\vec{C}) = \mathrm{d}\vec{r}$, where $\mathrm{d}\vec{r}$ is the path differential, thereby obtaining a path integral in the left-hand side:

$$\Gamma = \oint_C \vec{B} \cdot \mathrm{d}\vec{r} = \sum_{\vec{r}_n \in C} I_n \tag{6.128}$$

The formula (6.128) is the magnetostatic counterpart of (6.109). The value Γ of the closed contour integral in (6.128) is the *circulation* of magnetic field $\vec{B}$ along the closed contour C.

Next we should adapt the 2D Gauss's theorem formula (6.122a), which, after substituting $\vec{E} = \vec{R}^{-1}(\vec{B})$ takes the form

$$\Gamma = \oint_C \vec{B} \cdot \mathrm{d}\vec{r} = \int_S (\nabla \cdot \vec{R}^{-1}(\vec{B}))\, \mathrm{d}S \tag{6.129}$$

(where we already applied the same transformations in the left-hand side that we did earlier in the derivation of (6.128)). It remains to transform the dot product $\nabla \cdot \vec{R}^{-1}(\vec{B})$ in the right-hand side.

Let's rewrite this dot product in the matrix form:

$$\nabla^{\mathrm{T}} \cdot (\mathbf{R}^{-1} \cdot \mathbf{B}) = \left((\mathbf{R}^{-1})^{\mathrm{T}} \cdot \nabla\right)^{\mathrm{T}} \cdot \mathbf{B}$$

where all factors are treated as matrices (in particular ∇ and $\mathbf{B}$ are column matrices). Assuming a positively-oriented orthonormal coordinate system, the matrix representation of the operator $\vec{R}^{-1}$ is

$$\mathbf{R}^{-1} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Respectively

$$(\mathbf{R}^{-1})^{\mathrm{T}} \cdot \nabla = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \end{bmatrix} = \begin{bmatrix} -\dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial x} \end{bmatrix}$$

and

$$\begin{aligned} \nabla \cdot \vec{R}^{-1}(\vec{B}) &= \Big((\mathbf{R}^{-1})^{\mathrm{T}} \cdot \nabla\Big)^{\mathrm{T}} \cdot \mathbf{B} = \begin{bmatrix} -\dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial x} \end{bmatrix} \cdot \begin{bmatrix} B_x \\ B_y \end{bmatrix} \\ &= \frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} = \det \begin{bmatrix} \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} \\ B_x & B_y \end{bmatrix} = \nabla \times \vec{B} \end{aligned}$$

where $\nabla \times \vec{B}$ is the "2D vector cross product". Thereby (6.129) simplifies to

(Green's theorem, applicable only in 2D)

$$\Gamma = \oint_C \vec{B} \cdot \mathrm{d}\vec{r} = \int_S (\nabla \times \vec{B})\, \mathrm{d}S \tag{6.130}$$

In 3D the expression $\nabla \times \vec{B}$ (which thereby uses a proper cross product of vectors) is referred to as *curl* or *rotor* of $\vec{B}$. Since we are using the "2D cross product" notation here, we might as well refer to $\nabla \times \vec{B}$ in 2D as "curl" or "rotor" too. We thus obtain

("2D curl"/"2D rotor" formula)

$$\operatorname{curl} \vec{F} = \operatorname{rot} \vec{F} = \nabla \times \vec{F} = \det \begin{bmatrix} \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} \\ F_x & F_y \end{bmatrix} = \frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \tag{6.131}$$

applicable only in 2D.

▷ Comparing (6.126) to (6.128) we notice that in the latter we are having the currents I_n in place of the charges q_n. In (6.130) we are having $\nabla \times \vec{B}$ in place of the charge density $\rho = \nabla \cdot \vec{E}$. This suggests that

$$j = \nabla \times \vec{B} \tag{6.132}$$

is the 2D current density.

▷ In (6.114) we have effectively defined $\operatorname{div} \vec{E} = \mathrm{d}\Phi/\mathrm{d}V$, or, in a 2D case, $\operatorname{div} \vec{E} = \mathrm{d}\Phi/\mathrm{d}S$. Respectively, 2D curl can be defined as $\operatorname{curl} \vec{B} = \mathrm{d}\Gamma/\mathrm{d}S$:

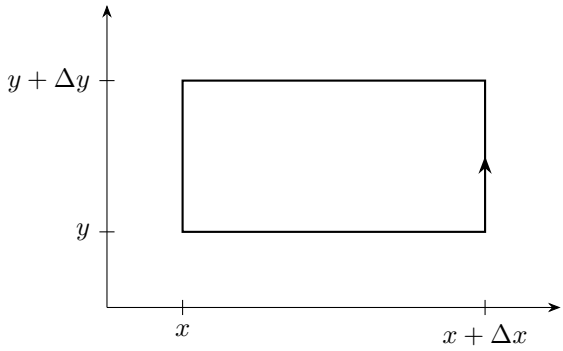


Figure 6.46: A small rectangular area.

("curl" of a vector field $\vec{F}$, 2D version)

$$\operatorname{curl} \vec{F}(\vec{r}_0) = \lim_{|C| \to 0} \frac{1}{\Delta S} \oint_C \vec{F}(\vec{r}) \cdot \mathrm{d}\vec{r} \tag{6.133}$$

(where $|C|$ is the 2D diameter of a closed curve surrounding $\vec{r}_0$, and ΔS is the area of the curve's interior).

▷ The curl formula (6.131) could have been derived from the curl definition (6.133) in a manner similar to how we derived the divergence formula (6.118) from the divergence definition (6.114) by considering the field circulation along the perimeter of an infinitely small rectangle (Fig. 6.46).

We also could have derived the entire Green's theorem by considering the circulation of magnetic field in a way similar to how we derived Gauss's theorem by considering the flux of electric field. With Gauss's theorem we had the number of field lines exiting the surface as an invariant quantity with respect to the surface's shape (this quantity being equal to the sum $\sum q_n$ of charges inside the surface). With Green's theorem the respective invariant quantity is given by the sum $\sum W_n I_n$, where W_n is the winding number of the circulation path with respect to the n-th current (therefore currents outside the path's interior are automatically discarded).

▷ Similarly to how the Gauss's theorem establishes the convention of orienting the normals to the boundary of a positively-oriented area (or volume) outwards, the Green's theorem establishes the convention of orienting the boundary of a positively-oriented area counterclockwise.

▷ Green's theorem is effectively the 2D Gauss's theorem with some readjusted vector orientation, where the field flux Φ becomes replaced with the field circulation Γ. Respectively the interior additivity of flux (6.123) becomes the additivity of circulation:

(additivity of the field circulation across the closed curve's interior)

$$\Gamma = \int_S \mathrm{d}\Gamma = \int_S (\nabla \times \vec{F}) \cdot \mathrm{d}S \tag{6.134}$$

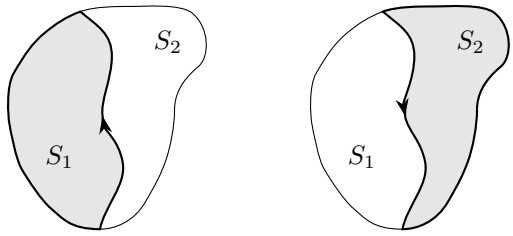


Figure 6.47: A positively-oriented 2D area split into two subareas S_1 and S_2 (which are thereby also positively oriented). The shared boundary of the two subareas has two opposite orientations, depending on for which of the two subareas we want to apply the Green's theorem.

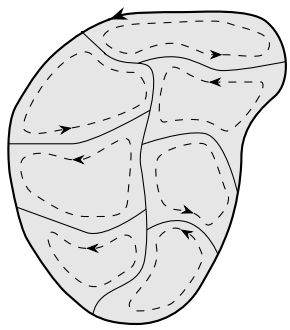


Figure 6.48: All inner circulations parts going along the shared boundaries cancel out and we are being left just with the flux through the outer boundary.

Figs. 6.47 and 6.48 illustrate.

Goal 6.17 completed. ■

Green's theorem is applicable only in 2D, because it is using the "2D curl". But it can be generalized to 3D, which is what we shall do next.

6.10.5 Stokes' theorem

We have obtained Green's theorem as a 2D formula, but the entire setup was 3D and involved a current flowing in the direction of the z axis through the xy plane. The 2D analysis was possible due to restrictions ensuring the invariance of the setup with respect to translations along the z axis. We would like to lift these restrictions, allowing arbitrary angles and shapes of the flowing currents, while the flat 2D integration area will be promoted to an arbitrary-shaped surface in 3D space.

Goal 6.18. Generalize Green's theorem to 3D. *Hint:* split the integration surface into infinitely small surface elements.

Given an arbitrary-shaped surface S in 3D, we split it into infinitely small oriented surface elements $\mathrm{d}\vec{S}$. Any such element can be considered flat, since its deviation from the flat shape is a higher-order infinitesimal, which vanishes

in the integration result (assuming the integrated function is continuous almost everywhere, which is what we need to assume anyway in order for the Riemann integral to exist). So the setup is locally 2D, except that the current doesn't have to cross the surface at a 90° angle, as we had in Green's theorem.

At this point, however, we can abandon the physical interpretation (that is we do not care if our results are physically correct or not) and concentrate on the mathematical aspects only. In that sense we can just concentrate on the field $\vec{B}$ irrespectively of its relationship to currents (or their existence whatsoever).

Let's consider $\mathrm{d}\vec{S}$ as a limiting version of a finite-size small flat elements $\Delta\vec{S}$. Given one such element $\Delta\vec{S}$, we choose the x and y axes lying in the plane of that element and orient the z axis in the direction of the normal vector $\Delta\vec{S}$. We are now going to apply Green's theorem in this coordinate system.

Computing the 2D field circulation along the boundary of the flat element ΔS in the xy plane we notice that this circulation would be the same if computed as a 3D field circulation. Indeed, the integration contour itself is flat, so $\mathrm{d}z \equiv 0$ across the entire contour. Therefore

$$\Delta\Gamma = \underbrace{\int_C \vec{B}\cdot\mathrm{d}\vec{r}}_{\text{as 3D path integral}} = \int_C (B_x\,\mathrm{d}x + B_y\,\mathrm{d}y + B_z\,\mathrm{d}z) = \int_C (B_x\,\mathrm{d}x + B_y\,\mathrm{d}y + B_z\cdot 0)$$

$$= \int_C (B_x\,\mathrm{d}x + B_y\,\mathrm{d}y) = \underbrace{\int_C \vec{B}\cdot\mathrm{d}\vec{r}}_{\text{as 2D path integral}} \tag{6.135a}$$

that is the value of B_z doesn't affect the circulation.

By Green's theorem

$$\Delta\Gamma = \int_{\Delta S}\left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}\right)\mathrm{d}S \tag{6.135b}$$

Our next goal is to try to replace the "2D cross product" in the right-hand side of (6.135b) by a proper 3D cross product.

We can notice that the value of the "2D cross product" in the right-hand side of (6.135b) is actually the z component of the proper 3D cross product $\nabla\times\vec{B}$, so we may rewrite (6.135b) as

$$\Delta\Gamma = \int_{\Delta S}(\nabla\times\vec{B})_z\,\mathrm{d}S$$

where $\nabla\times\vec{B}$ is the proper 3D cross product of ∇ and $\vec{B}$, while the notation $(\nabla\times\vec{B})_z$ refers to the z component of that cross product. Since the z axis is orthogonal to the flat surface ΔS, we have $\mathrm{d}S = \mathrm{d}S_z$ where $\mathrm{d}S_z$ is the z component of the vector $\mathrm{d}\vec{S}$, so we may replace $\mathrm{d}S$ with $\mathrm{d}S_z$:

$$\Delta\Gamma = \int_{\Delta S}(\nabla\times\vec{B})_z\,\mathrm{d}S_z$$

Noticing that the other components of $\mathrm{d}\vec{S}$ are zero ($\mathrm{d}S_x = \mathrm{d}S_y = 0$), we can replace the product under the integral sign with the full dot product without changing its value:

$$\Delta\Gamma = \int_{\Delta S}(\nabla\times\vec{B})_z\,\mathrm{d}S_z$$

$$
\begin{aligned}
&= \int_{\Delta S} \left((\nabla \times \vec{B})_x \, \mathrm{d}S_x + (\nabla \times \vec{B})_y \, \mathrm{d}S_y + (\nabla \times \vec{B})_z \, \mathrm{d}S_z \right) \\
&= \int_{\Delta S} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} \qquad (6.136)
\end{aligned}
$$

So, we transformed (6.135b) into an equal expression (6.136). Differently from (6.135b), the expression in the right-hand side of (6.135b) is in covariant form (since it's using covariant vector operations) and doesn't depend on the choice of the coordinate system. The 3D integral in (6.135a) is also in covariant form (since the dot products $\vec{B} \cdot \mathrm{d}\vec{r}$ also do not depend on the coordinate system choice). Equating the 3D integral in (6.135a) to the right-hand side of (6.136) yields

$$
\Delta\Gamma = \int_C \vec{B} \cdot \mathrm{d}\vec{r} = \int_{\Delta S} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} \qquad (6.137)
$$

where both integrals are 3D covariant-form integrals.

In the limit where the elements ΔS become infinitely small (in diameter), we would have

$$
\mathrm{d}\Gamma = (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S}
$$

On the other hand, circulation is additive across the surface's interior, which is ensured by pretty much the same mechanisms as in Green's theorem in 2D (in particular, refer to Figs. 6.47 and 6.48). It follows that the circulation along the boundary of S is

$$
\Gamma = \int_S \mathrm{d}\Gamma = \int_S (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S}
$$

and we arrive at

(Stokes' theorem, applicable only in 3D)

$$
\Gamma = \oint_C \vec{B}(\vec{r}) \, \mathrm{d}\vec{r} = \int_S \left(\nabla \times \vec{B} \right) \cdot \mathrm{d}\vec{S} \qquad (6.138)
$$

Apparently $\nabla \times \vec{B}$ is the (3D) curl of the field $\vec{B}$. Therefore Stokes' theorem says that the circulation of a vector field along a closed curve in 3D is equal to the surface integral of the field's curl over the curve's interior, where the interior can be chosen arbitrarily.

The curl (or rotor) formula is thereby

(curl/rotor formula, can be used as a definition)

$$
\operatorname{curl} \vec{F} = \operatorname{rot} \vec{F} = \nabla \times \vec{F} = \det \begin{bmatrix} \vec{e}_x & \vec{e}_y & \vec{e}_z \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ B_x & B_y & B_z \end{bmatrix} \qquad (6.139)
$$

which is is essentially the same as in 2D, except that it uses the true 3D cross product.

▷ It would be desirable to intuitively define the 3D curl along the same lines as we defined divergence and 2D curl. Loosely, $\operatorname{curl}\vec{F} = \mathrm{d}\Gamma/\mathrm{d}\vec{S}$. However here we have a problem with formalization of $\mathrm{d}\Gamma/\mathrm{d}\vec{S}$, since it's not immediately clear, what does it mean "to take a derivative with respect to a vector". On the other hand, gradient of a scalar field can be intuitively thought of as "derivative with respect to a vector" $\operatorname{grad} f = \mathrm{d}f/\mathrm{d}\vec{r}$, the direction of such derivative being defined as to ensure the equality

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\mathrm{d}f}{\mathrm{d}\vec{r}} \cdot \vec{r}\,'(t) = \frac{\mathrm{d}f}{\mathrm{d}\vec{r}} \cdot \frac{\mathrm{d}\vec{r}}{\mathrm{d}t}$$

Similarly, we could understand $\operatorname{curl}\vec{F} = \mathrm{d}\Gamma/\mathrm{d}\vec{S}$ as ensuring

$$\mathrm{d}\Gamma_1 = \frac{\mathrm{d}\Gamma}{\mathrm{d}\vec{S}} \cdot \mathrm{d}\vec{S}_1 \quad \forall \mathrm{d}\vec{S}_1 \tag{6.140}$$

where $\mathrm{d}\Gamma_1$ is the circulation across the boundary of an infinitely small surface element $\mathrm{d}\vec{S}_1$.

▷ In 2D with Green's theorem we had $j = \nabla \times \vec{B}$ as the current densities. In 3D we formally can define $\vec{j} = \nabla \times \vec{B}$ as vector current densities. Whether $\vec{j}$ obtained in this form are indeed physical current densities, is not a part of this book's materials. Readers who are interested, may check Maxwell's equations.

Goal 6.18 completed. ■

▷ Gauss's theorem (including the fundamental theorem of calculus as its particular case), Green's theorem and Stokes' theorem are all particular cases of the *generalized Stokes' theorem*, which relates an integral of a function over the boundary of a shape to an integral of some kind of function's derivative over the shape's interior.

6.10.6 Repeated application of nabla operator

In some variants of Gauss's and Stokes' theorem application we can vary the interior while keeping the boundary constant. Two particular cases of that are the following:

- in the path integral version (6.122c) of the Gauss's theorem we can vary the path while keeping the path's endpoints fixed (Fig. 6.49),
- in the Stokes' theorem (6.138) we can vary the surface while keeping the surface's boundary fixed.

Let's first concentrate on Fig. 6.49, and the path integral version of the Gauss's theorem (6.122c). In the setup implied by (6.122c), the integrals of the gradient of $\varphi(\vec{r})$ along both C_1 and C_2 are equal to $\varphi(\vec{r}_1) - \varphi(\vec{r}_2)$:

$$\int_{C_1} \nabla\varphi \cdot \mathrm{d}\vec{r} = \varphi(\vec{r}_1) - \varphi(\vec{r}_2) = \int_{C_2} \nabla\varphi \cdot \mathrm{d}\vec{r} \tag{6.141}$$

Let $\overline{C_2}$ be the same path as C_2 but with reverted direction, thereby

$$\int_{\overline{C_2}} \nabla\varphi \cdot \mathrm{d}\vec{r} = -\int_{C_2} \nabla\varphi \cdot \mathrm{d}\vec{r}$$

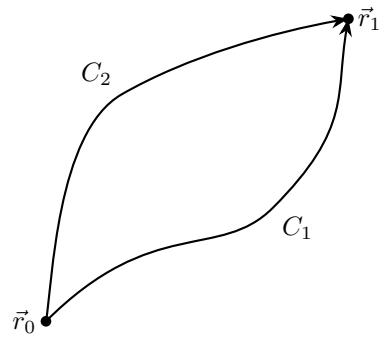


Figure 6.49: Two different oriented paths connecting the same points $\vec{r}_0$ and $\vec{r}_1$.

Combining with (6.141) yields

$$\int_{C_1} \nabla\varphi \cdot \mathrm{d}\vec{r} + \int_{\overline{C_2}} \nabla\varphi \cdot \mathrm{d}\vec{r} = \int_{C_1} \nabla\varphi \cdot \mathrm{d}\vec{r} - \int_{C_2} \nabla\varphi \cdot \mathrm{d}\vec{r} \overset{(6.141)}{=} 0$$

Thereby the integral along the closed path $C = C_1 \cup \overline{C_2}$ is also zero:

$$\oint_C \nabla\varphi \cdot \mathrm{d}\vec{r} = \int_{C_1} \nabla\varphi \cdot \mathrm{d}\vec{r} + \int_{\overline{C_2}} \nabla\varphi \cdot \mathrm{d}\vec{r} = 0$$

But there wasn't anything specific about the choice of C_1 and C_2, nor was there anything specific about the choice of the endpoints $\vec{r}_0$ and $\vec{r}_1$, all those choices could have literally been anything. Therefore the circulation of $\nabla\varphi$ along *any* closed path is zero:

(circulation of scalar field's gradient along any closed path is zero)

$$\oint_C (\nabla\varphi) \cdot \mathrm{d}\vec{r} = 0 \quad \forall C \tag{6.142}$$

Applying (6.133) to (6.142) yields

(curl of gradient is zero)

$$\nabla \times (\nabla\varphi) = \operatorname{curl}\operatorname{grad}\varphi = 0 \tag{6.143}$$

The formula (6.143) holds both for "2D curl" as well as for 3D curl. In 2D it holds by (6.133), which defined the 2D curl as the "circulation density". In 3D we might need to make a more complicated argument by applying the Stokes' theorem (6.138) to (6.142) yielding

$$\int_S \big(\nabla \times (\nabla\varphi)\big)\, \mathrm{d}\vec{S} \overset{(6.138)}{=} \oint_C (\nabla\varphi) \cdot \mathrm{d}\vec{r} = 0 \quad \forall S$$

(where S is the interior of C). The integral in the left-hand side can be zero for any S only if (6.143).

▷ Alternatively we could use the idea of "derivative with respect to a vector" (6.140) to obtain (6.143) from (6.142) in 3D.

Goal 6.19. Obtain the counterparts of (6.142) and (6.143) for the Stokes' theorem setup, where we vary the surface while keeping the fixed boundary.

Let two surfaces S_1 and S_2 have one and the same boundary C. By the Stokes' theorem (6.138),

$$\int_{S_1} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} = \oint_C \vec{B} \cdot \mathrm{d}\vec{r} = \int_{S_2} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} \tag{6.144}$$

Let $\overline{S_2}$ be the same surface as S_2 but with inverted orientation, thereby

$$\int_{\overline{S_2}} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} = -\int_{S_2} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S}$$

Combining with (6.144) yields

$$\begin{aligned}\int_{S_1} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} &+ \int_{\overline{S_2}} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} \\ &= \int_{S_1} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} - \int_{S_2} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} \overset{(6.144)}{=} 0\end{aligned}$$

Now the surfaces S_1 and $\overline{S_2}$ can be combined into a single closed surface S, since the boundary $\overline{C}$ of $\overline{S_2}$ is the inverse of the boundary C of S_1.

▷ Notice that we wouldn't be able to combine S_1 and S_2: at the "seam" (which is the curve C) the orientation of S_1 and S_2 will be mismatched with respect to the orientation of C. Indeed, suppose we combined S_1 and S_2 into a single closed surface, which could be positively or negatively oriented. For the sake of definiteness, let's assume it's positively oriented. Therefore all normals $\mathrm{d}\vec{S}$ of the surface S (which consist of all normals of S_1 and all normals of S_2) must be pointing outwards. Now, consider a small element $\Delta\vec{S}_1$ of the surface S_1, lying directly at the boundary C. In order to match the surface's orientation, the boundary has to go counterclockwise, when seen from the tip of the normal vector $\Delta\vec{S}_1$. The same should hold for a small element $\Delta\vec{S}_2$ of S_2. But both normals $\Delta\vec{S}_1$ and $\Delta\vec{S}_2$ are pointing outside of the surface's interior, while being located on opposite sides of C. It is not possible that the curve C is seen as going counterclockwise from the tips of both of these normals (Fig. 6.50 illustrates).

On the other hand, the boundary of $\overline{S_2}$ is also inverted, therefore it's not C but $\overline{C}$. The boundaries C and $\overline{C}$ are going in opposite directions, thus there is no direction conflict as with a single boundary C in Fig. 6.50.

Computing the integral across the closed surface $S = S_1 \cup \overline{S_2}$ therefore yields zero

$$\oint_S (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} = \int_{S_1} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} + \int_{\overline{S_2}} (\nabla \times \vec{B}) \cdot \mathrm{d}\vec{S} = 0$$

But there wasn't anything specific about the choice of S_1 and S_2, nor was there anything specific about the choice of the boundary C, all those choices could have literally been anything. Therefore the flux of $\nabla \times \vec{B}$ through *any* closed surface is zero:

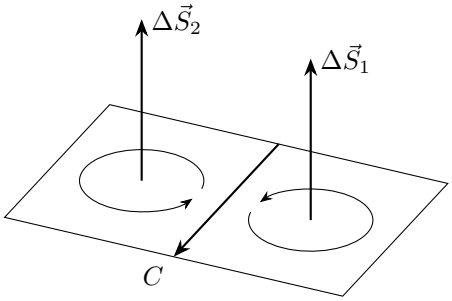


Figure 6.50: Orientation conflict of the two connected surfaces S_1 and S_2. $\Delta\vec{S}_1$ and $\Delta\vec{S}_2$ are normals to the two surface elements of S_1 and S_2, both elements directly neighboring the common boundary C (shown by the thick arrow in the middle). The two oriented circles show the expected direction of the boundary C with respect to $\Delta\vec{S}_1$ and $\Delta\vec{S}_2$. Only one of these two expectations can be satisfied at any given time. E.g. in this picture the boundary's direction matches the expectation of $\Delta\vec{S}_1$, but is opposite to the expectation of $\Delta\vec{S}_2$.

(flux of field's curl through a closed surface is zero)

$$\iint_S \left(\nabla \times \vec{F}\right) \cdot \mathrm{d}\vec{S} = 0 \tag{6.145}$$

▷ The formula (6.145) essentially states that the currents (represented by the densities $j = \nabla \times \vec{F}$) are not appearing out of nowhere and are not disappearing into nowhere: the total amounts of current entering and exiting S should be equal and cancel each other.

In physics the difference of the incoming and outgoing currents can be nonzero, but this means that the total amount of charge inside S (as well as the charge density) is not constant, in which case the magnetic field $\vec{B}$ is defined by a more complicated expression than the one we assumed in the derivation of Green's and Stokes' theorems. This however doesn't render the theorems incorrect, it's just that there is a difference between the field notated as $\vec{B}$ in our derivation of the theorems and the magnetic field predicted by the laws of physics (in simple words, in this case $\vec{B}$ doesn't stand for a physical magnetic field, but for some abstract vector function of a vector argument).

Applying (6.114) to (6.145) yields

(divergence of curl is zero)

$$\nabla \cdot \left(\nabla \times \vec{F}\right) = \operatorname{div}\operatorname{curl}\varphi = 0 \tag{6.146}$$

▷ Notice that (6.146) is written for the commonly used 3D curl. The "2D curl" we introduced earlier is a scalar function and we cannot compute its divergence.

Goal 6.19 completed. ■

▷ Another commonly encountered case of repeated application of the nabla operator is divergence of the gradient of a scalar field:

$$\nabla \cdot (\nabla \varphi) = \sum_n \frac{\partial}{\partial x_n} \left(\frac{\partial}{\partial x_n} \varphi \right) = \sum_n \frac{\partial^2}{\partial x_n^2} \varphi = \left(\sum_n \left(\frac{\partial}{\partial x_n} \right)^2 \right) \varphi = \nabla^2 \varphi$$

where ∇^2 is the *Laplace operator* or *Laplacian*:

(Laplace operator)

$$\nabla^2 = \nabla \cdot \nabla = \begin{bmatrix} \frac{\partial}{\partial x_1} & \cdots & \frac{\partial}{\partial x_N} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_N} \end{bmatrix} = \sum_n \left(\frac{\partial}{\partial x_n} \right)^2 = \operatorname{div} \operatorname{grad} \quad (6.147)$$

As shown above, Laplace operator applied to a scalar field yields the divergence of the field's gradient. In particular, if we assume that φ is an electrostatic potential field, then $\operatorname{grad} \varphi = \vec{E}$ is the electric field strength and $\nabla^2 \varphi = \operatorname{div} \vec{E} = \rho$ by (6.115) is the electric charge density. Thus $\rho = \nabla^2 \varphi$ gives the charge density of the electrostatic field.

Summary of Chapter 6

- We introduced the idea of the derivative as a way to compute the instantaneous flux of water flowing into a reservoir. We then reverted the problem and introduced Riemann integration as a way to compute the amount of flowed water given the flux as a function of time. We also extended Riemann integrals to Riemann–Stieltjes integrals in order to allow integration with respect to a function.

- We then went into multiple dimensions, introducing the concepts of partial derivatives and the nabla operator, as well as multidimensional and iterated integrals, where we also studied taking the limit, differentiation and integration with respect to a parameter under the integral sign.

- We generalized the averaging of a finite set of numbers into an average value of a function on an interval and derived the mean value theorem.

- We derived the most common particular cases of the generalized Stokes' theorem, namely the Gauss's theorem (including some of its generalizations), the Green's theorem and the Stokes' theorem, describing the relationships between the boundary and interior behaviors of functions.

Chapter 7

Exponential function

Analytic solution of differential equations as the limiting case of a finite-step scheme, and the exponential function as the limiting case of integer powers emerging in that scheme. Matrix diagonalization as a linear basis change for a differential equation system. Rotational behavior of exponentials of imaginary numbers as a limiting case of complex multiplication. Jordan normal form as a limiting case fallback of diagonalization. Discovering the spectral theorems. Linear space rotations as exponentials.

7.1 The simplest differential equation . . . 298
7.1.1 Constructive solution . . . 299
7.1.2 Negative arguments . . . 301
7.1.3 Taylor series . . . 305
7.2 Natural exponential . . . 305
7.3 Arbitrary-base exponentials . . . 308
7.4 Ordinary differential equations . . . 310
7.4.1 Solution goals . . . 310
7.4.2 Common solution approaches . . . 311
7.4.3 Constructive solution . . . 312
7.5 Matrix exponential . . . 314
7.6 Matrix diagonalization . . . 316
7.7 Second-order differential equation . . . 319
7.7.1 Common-case steps . . . 320
7.7.2 Overdamped case . . . 321
7.7.3 Motivation for complex numbers . . . 323
7.7.4 Underdamped case . . . 327
7.7.5 Critically-damped case . . . 328
7.8 Jordan normal form . . . 329
7.8.1 Rank 2 generalized eigenvectors . . . 330
7.8.2 Rank 3 generalized eigenvectors . . . 333
7.8.3 Rank 2 generalized eigenspace in 3D . . . 336
7.8.4 Jordan block . . . 338
7.8.5 Generalized eigenspaces . . . 340
7.8.6 Exponentiation of Jordan blocks . . . 343
7.9 Spectral theorems . . . 345
7.9.1 Diagonalization of quadratic forms . . . 346
7.9.2 Symmetric and Hermitian matrices . . . 353

7.9.3 Spectrum of a matrix transpose 358
7.9.4 Orthogonal and unitary matrices 363
7.9.5 Antisymmetric and anti-Hermitian matrices 366
7.9.6 Normal matrices 367
7.9.7 Non-unitary bases 369
7.9.8 Operator classes 370
7.10 Rotations as exponentials 372
Summary of Chapter 7 378

In this chapter we are going to talk about the exponential function e^x, also denoted as $\exp x$. One common approach is to define $\exp x$ by its Taylor series:

$$\exp x \triangleq 1 + x + \frac{x}{2} + \frac{x^2}{6} + \ldots + \frac{x^n}{n!} + \ldots \tag{7.1}$$

This approach creates a big motivational gap by leaving the following questions unanswered:

1. How did we come to consider the series (7.1) in the first place?
2. How did we come to the idea that the value given by the series (7.1) is equal to the power x of the number e, that is $\exp x = e^x$?

These questions are much better answered using the other, somewhat less popular definition of the $\exp x$ function:

$$\exp x \triangleq \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n \tag{7.2}$$

Compared to the elegant Taylor series-based definition (7.1), this latter definition (7.2) may look way less intuitive. However this couldn't be further away from the truth: the definition (7.2) is much more intuitively connected to what one would expect from the exponentiation operation. One of the main goals of this chapter is to promote (7.2) to the position of the "go to" definition of $\exp x$.

Before we do so, we however need to come up with some motivation to consider the function given by (7.1) and (7.2) in the first place. This motivation can come from solution of differential equations and this is what we shall look at.

Motivational starter for Chapter 7

Develop an approach for solving fixed-coefficient ordinary linear differential equations (or systems thereof).

7.1 The simplest differential equation

We begin by considering the simplest possible differential equation[1]

$$\frac{\mathrm{d}y}{\mathrm{d}t} = y \tag{7.3}$$

[1] Arguably, $\frac{\mathrm{d}y}{\mathrm{d}t} = f(t)$ is an even simpler differential equation, but that one is trivial.

where $y = y(t)$ is the unknown function. We use t rather than x for the argument of $y(\cdot)$, because, intuitively, we would like to suggest time semantics for this variable as one possible interpretation. In this case (7.3) could be describing some real physical process.

▷ Looking at the equation (7.3), we can realize its intuitive meaning: we are looking for a function equal to its own derivative. This is a well-known property of the function $\exp x$, but it's not among the primary properties of the exponentiation operation per se: it doesn't apply to exponentiation on the set of integer or even rational numbers, as we cannot differentiate there. So let's not get too distracted by this specific property or too excited about it. The reason we have introduced equation (7.3) is not that we are aiming to define the natural exponential as a function which is its own derivative. Rather, we simply take differential equations as the motivation to talk of real powers of numbers, because, besides differential equations, its a bit difficult to find another area where one would even need real powers. Of course, one could still introduce real powers for the sake of extending the exponentiation operation from integer (or maybe rational) powers to real ones, but in the absence of a proper need, such extension might be totally academic. In particular, lacking a real purpose, we wouldn't have a reference point to decide how to resolve the definition ambiguities, if any should occur, and, as we know, the exponentiation operation has quite a few of those, starting with how to define x^0 and going further into fractional and real powers, which can have multiple values.

7.1.1 Constructive solution

Goal 7.1. Find the solution of (7.3) in a constructive manner, without taking guesses how the solution might look like or using other non-obvious steps such as integrating multiplier. *Hint:* first find the solution of the discrete-step approximation of (7.3).

Intuitively, differential equations can be considered as limiting cases of difference equations. The latter approach becomes even more intuitive by specifically considering the initial value problem. Therefore let's equip (7.3) with the initial value condition $y(t_0) = y_0$:

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}t} = y \\ y(t_0) = y_0 \end{cases}$$

Without loss of generality we can assume $t_0 = 0$, since the choice of t_0 can be seen as choosing the formal origin of the t axis:

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}t} = y \\ y(0) = y_0 \end{cases} \tag{7.4}$$

Replacing the derivative by a finite difference, we obtain

$$\begin{cases} \dfrac{y(t+\Delta t) - y(t)}{\Delta t} = y(t) \\ y(0) = y_0 \end{cases} \tag{7.5}$$

Choosing a sequence of points $\{t_n\}$:

$$\{t_n\} = \{t_0,\ t_1,\ t_2,\ \dots\} \qquad \text{where } t_0 = 0$$

we define $\Delta t_n = t_n - t_{n-1}$ and $y_n = y(t_n)$ and rewrite (7.5) as

$$\frac{y_{n+1} - y_n}{\Delta t_{n+1}} = y_n \tag{7.6}$$

where the initial condition $y(t_0) = y_0$ is now specified implicitly by our choice of notation. Resolving (7.6) with respect to y_{n+1} we have

$$y_{n+1} = y_n \Delta t_{n+1} + y_n$$

or

$$y_{n+1} = (1 + \Delta t_{n+1}) y_n \tag{7.7}$$

So, our original initial value problem (7.4) can be seen as a limiting case of the difference scheme (7.7).

Goal 7.1a. Construct the solution of (7.7). *Hint:* for our purposes it suffices to assume a constant step $\Delta t_n = \text{const}$.

To simplify our further argument, we will assume that all Δt_n are identical: $\Delta t_n = \Delta t \ \forall n$. A fully rigorous argument might need to address the case of different Δt_n. We are going to skip this detail, since we already know from the existing literature that our results are going to be correct anyway. Here we are concerned with sketching an intuitive path towards an intuitively understandable definition of $\exp x$ (or, more precisely, $\exp t$, since currently we are using the variable t as the argument), while proper proofs can be found elsewhere.

With identical Δt_n, (7.7) becomes

$$y_{n+1} = (1 + \Delta t) y_n$$

and therefore

$$y_n = y_0 \cdot (1 + \Delta t)^n \tag{7.8}$$

So, our original initial value problem (7.4) can be seen as a limiting case of (7.8) and the latter is obviously an (integer-power) exponential. That is, the solution of the finite-difference initial value problem, corresponding to our original continuous initial-value problem (7.4) is an (integer-argument) exponential. *Equation (7.8) is the starting point of our intuition that the solution of the continuous initial-value problem might be an exponential as well.*

Goal 7.1a completed. ■

Goal 7.1b. Transform the discrete-step formula (7.8) into a continuous form.

Noticing that

$$\Delta t = \frac{t_n - t_0}{n} = \frac{t_n}{n}$$

and recalling that $y_n = y(t_n)$ we can rewrite (7.8) as

$$y(t_n) = y_0 \cdot \left(1 + \frac{t_n}{n}\right)^n$$

Now, considering t_n as our point of interest we can denote it as simply t, obtaining

$$y(t) = y_0 \cdot \left(1 + \frac{t}{n}\right)^n$$

Ultimately, fixing the value of t and taking the limit at $\Delta t \to 0$, that is at $n \to \infty$, we obtain

$$y(t) = y_0 \cdot \lim_{n\to\infty} \left(1 + \frac{t}{n}\right)^n \tag{7.9}$$

Goal 7.1b completed. ■

Goal 7.1 completed. ■

At this point we can notice that the limit itself in (7.9) is independent from the initial condition and depends only on t, therefore we can consider it as some function of t, which we denote as $\exp t$:

(natural exponential function)

$$\exp t = \lim_{n\to\infty} \left(1 + \frac{t}{n}\right)^n \tag{7.10a}$$

▷ Notice that we don't claim yet that $\exp t = e^t$. That is so far we consider $\exp t$ simply as some function with unknown properties, defined by (7.10a). We could have notated it as $f(t) = \lim (1 + t/n)^n$ instead.

The solution (7.9) can be thereby written as

$$y(t) = y_0 \cdot \exp t \tag{7.10b}$$

Thereby the function $\exp t$ is the solution of the initial value problem (7.4) for $y_0 = 1$:

(implicit definition of $\exp t$***)***

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \exp t = \exp t \\ \exp 0 = 1 \end{cases} \tag{7.11}$$

7.1.2 Negative arguments

Actually, for the time being, we might need to restrict t to $t \geq 0$ in the definition (7.10a). Indeed, our approach to the solution of the initial value problem (7.4) was intuitively assuming time-based interpretation, the differential equation (7.3) being intuitively understood as a process evolving in time. Since $t = 0$ is when we "started" the process, we can only consider $t \geq 0$.

We could, of course, avoid the time-based interpretation, and say that t is simply an abstract variable, without time semantics, and so we can simply apply the discretization (7.5) in the opposite direction for $t < 0$ and $\Delta t < 0$, arriving at exactly the same (7.10a) for $t < 0$ as well. This, however, would mean that we introduce two different discretization schemes for $t > 0$ and $t < 0$, going in two opposite directions along the t axis. Indeed, for $t > 0$ and $\Delta t > 0$ we had $t_{n+1} > t_n$ in the equation (7.6). So the value y_{n+1}, which is to the right of y_n on the t axis, was computed from y_n. However with $\Delta t < 0$ the value y_{n+1} is going to be to the left of y_n. Thus, one discretization scheme doesn't smoothly

transition into the other at $t = 0$, and there would be some kind of "seam" at the point $t = 0$. Intuitively we might expect that it doesn't matter, but it's not fully obvious that both schemes are equivalent in the limit.

The implicit definition (7.11) is somewhat more flexible in that sense. Instead of considering $t = 0$ as the starting time moment, we could consider the second equation $\exp 0 = 1$ simply as a constraint imposed on the solutions of the differential equation $\frac{\mathrm{d}}{\mathrm{d}t} \exp t = \exp t$, the solutions themselves running across the entire time axis from $t = -\infty$ to $t = +\infty$. Thus the definition (7.11) can be considered as defining $\exp t \ \forall t \in \mathbb{R}$ rather than only for $t \geq 0$. This is raising a natural question of whether (7.10a) can be also applied for $t < 0$, and whether the function $\exp t$ defined for $t \in \mathbb{R}$, obtained by (7.10a), is the same as the one obtained from (7.11).

Goal 7.2. Demonstrate the compatibility of (7.10a) to (7.11) for $t < 0$. *Hint:* use the time-invariance of (7.3).

The definition (7.10a) uniquely defines the function $\exp t \ \forall t$. On the other hand, our suggested interpretation of $\exp 0 = 1$ as a constraint applied to $\frac{\mathrm{d}}{\mathrm{d}t} \exp t = \exp t$ theoretically could allow multiple solutions of (7.11). Intuitively, our interpretation of (7.3) as a process evolving in time kind of excludes multiple solutions for $t \geq 0$ ("after $t = 0$"), but for $t < 0$ ("before $t = 0$") it is not so clear, if we let the time run from $-\infty$ to $+\infty$ in a uniform fashion.

> ▷ Had we suggested to interpret negative values of t in (7.11) as time running backwards, we might have been able to directly apply (7.10a), thereby reaching Goal 7.2, but ending up with the earlier mentioned "seam" at $t = 0$. This "seam", in turn, would have made uniform treatment of $\exp t$ on the entire $\mathbb{R}$ somewhat more difficult, so we'd rather not do that.

Thus we stick with the interpretation of (7.3) and (7.11) where the time runs only forward. In this case we should understand $\exp t$ at $t < 0$ in (7.11) in the following sense. Given $t_0 < 0$, the correct value of $\exp t_0$ is defined as one satisfying the following property. Treating t_0 as the starting time moment, and $\exp t_0 = y_0$ as the initial value, we let the time in (7.3) run forward from t_0 to 0. If we thereby reach the value $y(0) = 1$, the value $y_0 = \exp t_0$ is the one we have been looking for, otherwise (if $y(0) \neq 1$) we have to come up with another guess for the value of $\exp t_0$.

The just described property is (for the time being) our definition of $\exp t$ for $t < 0$, however is there only one such value satisfying this definition, or could there be sometimes more? To investigate this question, let's first study one important implication of the time-invariance of (7.3).

The time-invariance of (7.3) means that whatever starting point we choose for (7.4) and/or (7.11) instead of $t = 0$, the answer (7.9) will be essentially the same, we'll just need to replace t in the right-hand side of (7.9) with the time Δt elapsed since the starting point. That is, instead of always starting at $t = 0$, we can start at some other $t = t_0$. In this case, by (7.9), we obtain

$$\exp(t_0 + \Delta t) \overset{(7.9)}{=} y_0 \cdot \lim_{n\to\infty} \left(1 + \frac{\Delta t}{n}\right)^n = \exp t_0 \cdot \exp \Delta t \qquad \forall \Delta t \geq 0 \tag{7.12}$$

We can notice that the application of the equality $\lim (1 + \Delta t/n)^n = \exp \Delta t$ implicitly refers to (7.10a), therefore for now we have to restrict Δt to $\Delta t \geq 0$.

On the other hand, t_0 in (7.12) is not restricted to nonnegative values, since we consider $\exp t_0$ as emerging from (7.3) rather than from (7.10a).

Assuming $t_0 < 0$ and letting $\Delta t = -t_0 > 0$ in (7.12), we obtain

$$1 = \exp 0 = \exp(t_0 + \Delta t) = \exp t_0 \cdot \exp \Delta t = \exp(-\Delta t) \cdot \exp \Delta t$$

and it follows that

$$\exp(-\Delta t) = \frac{1}{\exp \Delta t} \qquad \forall \Delta t \geq 0$$

(where we switched from $\Delta t > 0$ to $\Delta t \geq 0$, since $\exp 0 = 1/\exp 0$ by direct verification). Replacing the Δt notation with t we can write

$$\exp(-t) = \frac{1}{\exp t} \qquad \forall t \geq 0 \tag{7.13}$$

where $\exp t$ in the right-hand side may be evaluated by (7.10a).

So we have obtained (7.13) from (7.11) by applying the time-invariance of (7.3). We are going to use (7.13) as a constructive definition of $\exp t$ for $t < 0$ for the time being.

Goal 7.2a. Demonstrate that (7.10a) gives the same value of $\exp t$ for $t < 0$ as (7.13). *Hint:* invert the time direction in (7.3).

It would be sufficient to demonstrate that

$$\lim_{n\to\infty} \left(1 + \frac{t}{n}\right)^n \cdot \lim_{n\to\infty} \left(1 - \frac{t}{n}\right)^n = 1 \qquad \forall t \geq 0 \tag{7.14}$$

We could try to do that by engaging in various transformations and analysis of the left-hand side of (7.14), but that would be a proof for the sake of a proof. For our purposes here, it would be more appropriate to choose a more intuitive (although possibly not fully rigorous) path.

Instead of (7.3), consider the equation:

$$\frac{\mathrm{d}\bar{y}}{\mathrm{d}t} = -\bar{y} \tag{7.15}$$

This equation can be formally obtained from (7.3) by replacing $\mathrm{d}t$ with $\mathrm{d}(-t)$, that is, by inverting time, therefore its solution should be the time-inverted solution of (7.3):

$$\bar{y}(t) = \bar{y}_0 \cdot \lim_{n\to\infty} \left(1 - \frac{t}{n}\right)^n$$

at least for $t \geq 0$. In particular, letting $y_0 = 1$, we obtain

$$\bar{y}(t) = \lim_{n\to\infty} \left(1 - \frac{t}{n}\right)^n$$

which is simply (7.10a) where t is replaced by $-t$.

▷ It's not difficult to notice that the same $\bar{y}(t)$ could have been obtained "manually" by taking (7.15) and repeating the solutions steps we did earlier to obtain (7.10a) from (7.15).

It remains to demonstrate that $y(t)\bar{y}(t) = 1\ \forall t \geq 0$ (where $y(t) = \exp t$). To do that, simply take the derivative:

$$\frac{\mathrm{d}}{\mathrm{d}t}(y\bar{y}) = \frac{\mathrm{d}y}{\mathrm{d}t}\bar{y} + y\frac{\mathrm{d}\bar{y}}{\mathrm{d}t} = y\bar{y} + y \cdot (-\bar{y}) = 0$$

and since $y(0)\bar{y}(0) = 1 \cdot 1 = 1$, it follows that $y\bar{y} = 1\ \forall t \geq 0$. Thereby we effectively demonstrated (7.14)

Goal 7.2a completed. ■

Goal 7.2 completed. ■

It follows that (7.10a) *can be freely applied for all real* t *and thereby can be used as the definition of* $\exp t\ \forall t \in \mathbb{R}$. It further follows that (7.13) can be applied for all real argument values:

("time-inversion" symmetry of exponentiation)

$$\exp t \cdot \exp(-t) = 1 \quad \forall t \in \mathbb{R} \tag{7.16}$$

Respectively we can lift the restriction $\Delta t \geq 0$ in (7.12). E.g., by formally multiplying both sides of (7.12) by $\exp(-\Delta t)$ where $\Delta t \geq 0$, we obtain

$$\exp(t_0 + \Delta t) \cdot \exp(-\Delta t) \overset{(7.12)}{=} \exp t_0 \cdot \exp \Delta t \cdot \exp(-\Delta t) \overset{(7.16)}{=} \exp t_0$$

Replacing $t_0 + \Delta t$ by t_0 yields

$$\exp(t_0) \cdot \exp(-\Delta t) = \exp(t_0 - \Delta t) \qquad \forall \Delta t \geq 0$$

which effectively covers (7.12) for $\Delta t < 0$. Thus, by additionally replacing the variable names with x and y, (7.12) becomes

("time-invariance" of exponentiation)

$$\exp(x + y) = \exp x \cdot \exp y \qquad \forall x, y \in \mathbb{R} \tag{7.17}$$

▷ We refer to (7.17) as the "time-invariance" property since it is a direct consequence of the time-invariance of (7.3).

The intuitive interpretation of (7.17) is as follows. Starting at $t = x$ we advance in time (forwards or backwards) by y time units. This causes the function value to grow by the factor $\exp y$. If $y = \Delta t > 0$, the growth factor $\exp y = \exp \Delta t$ is larger than unity. If $y < 0$ and $\Delta t = -y > 0$, the growth factor $\exp y = 1/\exp(-y) = 1/\exp \Delta t$ is smaller than unity, the function actually "grows back" to the value it had "Δt time units ago". Conversely, starting at the moment $t = x + y = x - \Delta t$, that was "Δt time units ago", and going forward by Δt time units to $t = x$ we have:

$$\exp(x + y) \cdot \exp \Delta t = \exp(x - \Delta t) \cdot \exp \Delta t = \exp x$$

7.1.3 Taylor series

Goal 7.3. Rewrite (7.10a) in the form of Taylor series. *Hint:* apply binomial expansion to the expression under the limit in (7.10a).

We have introduced the function $\exp t$ as the limit (7.10a), which gives the solution of (7.11). We still haven't demonstrated that $\exp t = e^t$ (this will be the topic of Section 7.2), but we already can use (7.10a) as the definition of the function $\exp t$. The Taylor series-based expression for $\exp t$ then shall be taken as the secondary, alternative and equivalent expression of the same definition.

We could construct this Taylor series according to the general Taylor series formula by evaluating the 1st- and higher-order derivatives of $\exp t$ at $t = 0$, which by (7.3) are all equal to $\exp 0 = 1$. However it might be more instructive to obtain this Taylor series directly by rewriting (7.10a).

Applying the binomial expansion under the limit sign in (7.10a) we obtain

$$\begin{aligned}
\exp t &= \lim_{n\to\infty}\left(1+\frac{t}{n}\right)^n \\
&= \lim_{n\to\infty}\left(1+n\frac{t}{n}+\frac{n(n-1)}{2}\left(\frac{t}{n}\right)^2+\dots\right. \\
&\qquad\qquad \left.+\frac{n(n-1)\dots(n-k+1)}{k!}\left(\frac{t}{n}\right)^k+\dots\right) \\
&= \lim_{n\to\infty}\sum_{k=0}^{n}\frac{n(n-1)\dots(n-k+1)}{n^k}\frac{t^k}{k!} = \lim_{n\to\infty}\sum_{k=0}^{n}a_k\frac{t^k}{k!}
\end{aligned}$$

where

$$a_k = \frac{n(n-1)\dots(n-k+1)}{n^k}$$

Looking at a_k and imagining n being large, we can notice that in the first few terms of the sum we have $a_k \approx 1$ (since $n-k \approx n$), while the latter terms of the sum have a_k noticeably smaller than 1, but these terms can be ignored (or at least so would we hope) due to large $k!$, yielding

(Taylor series expansion for $\exp t$***)***

$$\exp t = \lim_{n\to\infty}\left(1+t+\frac{t^2}{2}+\dots\frac{t^k}{k!}+\dots\right) = 1+t+\frac{t^2}{2}+\dots\frac{t^k}{k!}+\dots \tag{7.18}$$

A more rigorous demonstration falls outside the scope of this book.

Goal 7.3 completed. ■

7.2 Natural exponential

By its own definition (7.10a), the function $\exp t$ is a somewhat strange limit of an integer-power exponential, but why should we consider it as a real-power exponential e^t? It would be natural to consider e.g. the limit $a^b = \lim_{t\to b} a^t$ as the definition of a^b (which, in order to avoid circular argument, would need to restrict t to rational values, while a and b itself can be real), but the limit

(7.10a) has a totally different nature. We can however notice that $\exp t$ actually *does* behave as an exponential of some number, and this is due to the property (7.17). Let's begin with the following observation.

Let's again consider the initial value problem (7.11), which is an alternative definition for $\exp t$. At $t = 0$ the solution of (7.11) simply yields the initial condition $\exp 0 = 1$. Now let the differential process evolve in time for one time unit, thereby taking us to $t = 1$. By (7.17), the value of $\exp t$ should grow by $\exp 1$:

$$\exp 1 = \exp 0 \cdot \exp 1$$

Now let another time unit pass. By (7.17),

$$\exp 2 = \exp 1 \cdot \exp 1 = \exp 0 \cdot \exp 1 \cdot \exp 1$$

Letting another time unit pass we obtain

$$\exp 3 = \exp 2 \cdot \exp 1 = \exp 0 \cdot \exp 1 \cdot \exp 1 \cdot \exp 1$$

After one more we have

$$\exp 4 = \exp 3 \cdot \exp 1 = \exp 0 \cdot \exp 1 \cdot \exp 1 \cdot \exp 1 \cdot \exp 1$$

Repeating n times in total, we have

$$\exp n = \exp 0 \cdot \underbrace{\exp 1 \cdot \ldots \cdot \exp 1}_{n \text{ factors}} = 1 \cdot (\exp 1)^n = (\exp 1)^n \quad \forall n \geq 0,\, n \in \mathbb{Z} \tag{7.19}$$

According to (7.19), at least for positive integer values of t, the function $\exp t$ behaves as an exponent of the number $\exp 1$. We will denote this number as e:

$$e \triangleq \exp 1 \tag{7.20a}$$

or, applying (7.10a) on top of (7.20a),

(Euler's number)

$$e \triangleq \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n \tag{7.20b}$$

while (7.19), by (7.20a), can be rewritten as

$$\exp n = e^n \quad \forall n \geq 0,\, n \in \mathbb{Z} \tag{7.21}$$

which we shall generalize to $\exp t = e^t \ \forall t \in \mathbb{R}$ during the next steps.

Goal 7.4. Generalize (7.21) from nonnegative integer argument n to real argument $t \in \mathbb{R}$. *Hint:* extend the set of nonnegative integers to $\mathbb{R}$ step by step.

Goal 7.4a. Generalize to nonnegative rational arguments.

Instead of taking unit time interval steps in the derivation of (7.19), let's take rational time interval steps of length $\frac{1}{m}$, where $m \in \mathbb{N}$. Then, instead of (7.19), we are going to obtain

$$\exp\left(\frac{n}{m}\right) = \left(\exp\left(\frac{1}{m}\right)\right)^n \quad \forall n \geq 0,\, n \in \mathbb{Z} \tag{7.22}$$

Substituting $n = m$ into (7.22) yields

$$e = \exp 1 = \exp\left(\frac{m}{m}\right) = \left(\exp\left(\frac{1}{m}\right)\right)^m$$

and it follows that $\exp(\frac{1}{m}) = \sqrt[m]{\exp 1} = \sqrt[m]{e}$. Then, by (7.22),

$$\exp\left(\frac{n}{m}\right) = \left(\exp\left(\frac{1}{m}\right)\right)^n = (\sqrt[m]{e})^n = e^{n/m}$$

or simply

$$\exp\left(\frac{n}{m}\right) = e^{n/m} \qquad \forall \tfrac{n}{m} \geq 0,\ \tfrac{n}{m} \in \mathbb{Q} \tag{7.23}$$

which is the counterpart of (7.21) for nonnegative rational numbers.

Goal 7.4a completed. ■

Goal 7.4b. Generalize to negative rational arguments.

By (7.16)

$$\exp(-t) = \frac{1}{\exp t} = \frac{1}{e^t} = e^{-t} \quad \forall t \geq 0,\ t \in \mathbb{Q} \tag{7.24}$$

which justifies the applicability of (7.23) for $\frac{n}{m} = t < 0$, and thereby

$$\exp t = e^t \quad \forall t \in \mathbb{Q} \tag{7.25}$$

Goal 7.4b completed. ■

Goal 7.4c. Generalize to $t \in \mathbb{R}$.

Until now we have been implicitly assuming that we have defined rational powers of real bases as $a^{n/m} = \sqrt[m]{a^n} = (\sqrt[m]{a})^n$, where $\sqrt[m]{a}$ is understood as arithmetic root (returning a positive value for a positive a and any $m \in \mathbb{N}$). In Section (2.6) we used the continuity of arithmetic operations to define them for real arguments. More specifically we extrapolated (by postulating) the already established continuity of operations on the set of rational numbers to the set of real numbers. Thus, a real power t of the number e would be obtained by extending the continuity of e^t from $t \in \mathbb{Q}$ to $t \in \mathbb{R}$.

The continuity of e^t on $\mathbb{Q}$ follows from (7.25) and from the continuity of $\exp t$ on $\mathbb{R}$, the latter following directly from the fact that $\exp t$ is a solution of the differential equation (7.3) (being differentiable, $\exp t$ must be continuous). Thereby e^t can be extended to $\mathbb{R}$ in a continuous manner. Now we have two continuous functions e^t and $\exp t$, which by (7.25) are equal to each other everywhere on $\mathbb{Q}$. It follows, by their continuity, that they are also equal to each other everywhere on $\mathbb{R}$:

(the function** $\exp t$ **is the same as** e^t**)

$$\exp t = e^t \quad \forall t \in \mathbb{R} \tag{7.26}$$

Goal 7.4c completed. ■

Goal 7.4 completed. ■

7.3 Arbitrary-base exponentials

In (7.15) we modified (7.3) to switch its solution from e^t to e^{-t}. This could lead us to the idea that we could generally switch from e^t to $e^{\lambda t}$ by replacing (7.3) with

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \lambda y \tag{7.27}$$

The readers, if they wish, may repeat the steps of the derivation of the solution of (7.3) to convince themselves that the solution of (7.27) indeed $y_0 e^{\lambda t}$.

▷ The equation (7.27) can be alternatively rewritten as

$$\frac{\mathrm{d}y}{\mathrm{d}(t\lambda)} = y$$

suggesting that we're simply changing the scaling of the t axis by the factor λ and respectively the solution e^t is scaled in t by the same factor.

As a further step we may wish to rewrite $e^{\lambda t}$ as

$$e^{\lambda t} = (e^{\lambda})^t = a^t \tag{7.28}$$

where $a = e^{\lambda}$. Before we do that, we however would like to demonstrate the property $e^{xy} = (e^x)^y$.

Goal 7.5. Demonstrate $e^{xy} = (e^x)^y \ \forall x, y \in \mathbb{R}$. *Hint:* start with (7.17).

Applying (7.17) iteratively $n-1$ times, we obtain

$$\begin{aligned} e^{nx} &= \exp nx = \exp(\underbrace{x + x + \ldots + x}_{n}) \overset{(7.17)}{=} \underbrace{\exp(x) \cdot \exp(x) \cdot \ldots \cdot \exp(x)}_{n} \\ &= (\exp x)^n = (e^x)^n \quad \forall n \geq 0,\, n \in \mathbb{Z},\, x \in \mathbb{R} \end{aligned} \tag{7.29}$$

The extension of (7.29) to negative $n \in \mathbb{Z}$ occurs by (7.16):

$$e^{-nx} \overset{(7.16)}{=} \frac{1}{e^{nx}} \overset{(7.29)}{=} \frac{1}{(e^x)^n} = (e^x)^{-n}$$

thus

$$e^{nx} = (e^x)^n \quad \forall n \in \mathbb{Z},\, x \in \mathbb{R} \tag{7.30}$$

It follows that

$$\begin{aligned} e^{\frac{n}{m}x} &= \exp\left(\frac{n}{m}x\right) = \exp\left(n\frac{x}{m}\right) \overset{(7.30)}{=} \left(\exp\left(\frac{x}{m}\right)\right)^n = (e^{\frac{x}{m}})^n = \left(\sqrt[m]{e^x}\right)^n \\ &= (e^x)^{\frac{n}{m}} \qquad \forall n \in \mathbb{Z},\, m \in \mathbb{N},\, x \in \mathbb{R} \end{aligned}$$

or simply

$$e^{xy} = (e^x)^y \quad \forall x \in \mathbb{R},\, y \in \mathbb{Q} \tag{7.31}$$

The extension to $y \in \mathbb{R}$ occurs by extrapolating the continuity of e^{xy} with respect to y:

$$e^{xy} = (e^x)^y \quad \forall x, y \in \mathbb{R} \tag{7.32}$$

Goal 7.5 completed. ■

With (7.32) we can use (7.28) to compute exponentials of arbitrary positive real numbers a. Except that so far (7.28) only suggests how to convert λ to a, namely by computing $a = e^{\lambda}$. It follows that λ can be obtained from a by applying the function inverse of the natural exponential, which is

(natural logarithm)

$$y = e^x \iff x = \ln y$$

(where, since $e^x > 0\ \forall x$, the inverse function is defined for $y > 0$). Thereby the exponential a^t is understood as

(base a real exponential)

$$a^t = e^{t \ln a} \quad \forall t \in \mathbb{R},\ a \in \mathbb{R},\ a > 0 \tag{7.33}$$

▷ Substituting (7.10a) into (7.33) yields

$$a^t = e^{t \ln a} = \lim_{n \to \infty} \left(1 + \frac{t \ln a}{n}\right)^n \tag{7.34}$$

The formula (7.34) has a simple intuitive interpretation. We represent a^t as

$$a^t = (\sqrt[n]{a^t})^n = \left(a^{\frac{t}{n}}\right)^n \approx \left(1 + \frac{t \ln a}{n}\right)^n \tag{7.35}$$

If n is large (compared to $t \ln a$), the expression $\sqrt[n]{a^t}$ becomes almost linear with respect to t:

$$\sqrt[n]{a^t} = a^{\frac{t}{n}} \approx 1 + \frac{t \ln a}{n}$$

In other words, we have "zoomed in" enough into a^t to make this function look almost linear (compare to our discussion in Section 3.5). At $n \to \infty$ the equalities become exact.

Now that we have introduced (7.27) as the motivation to compute real-power exponentials a^t of arbitrary base a, we can identify two different interpretations (and two different results) of the expression a^t, depending on the context. Let $a = 2$ and let $t = \frac{1}{2}$. According to (7.27) and (7.33), we have a single value

$$a^t = 2^{\frac{1}{2}} = e^{\frac{1}{2} \ln 2} \approx 1.4142\ldots$$

However we can also interpret raising into the $\frac{1}{2}$ power in the algebraic sense, that is in the sense of solving the equation

$$\left(a^{\frac{1}{2}}\right)^2 = a^{\frac{1}{2} \cdot 2} = a^1 = a$$

which has two solutions

$$a = \pm\sqrt{2} \approx \pm 1.4142\ldots$$

So, with rational powers we can have at least these two different motivations leading to different answers to the problem of computing the exponential a^t for a

rational t. In one case (differential equation- or calculus-based) we get a single answer, in the other (algebraic) case we get multiple answers. Respectively, in the algebraic case we wouldn't even be able to take the limit $a^{n/m} \to a^t$ for $n/m \to t$, as the number of answers and the answers themselves are going to vary chaotically with n/m. This is why we insisted on coming up with a proper motivation to compute exponentials of a real argument: if we started with our standard algebraic definition of a rational power, we wouldn't be able to generalize to real powers (unless we artificially restrict the algebraic-case answers to positive results, but, lacking a proper motivation, that would be a pretty random action).

7.4 Ordinary differential equations

Having dealt with the simplest possible differential equation (7.3), we shall now concentrate on a somewhat more general class of the 1st-order fixed-coefficient linear ordinary differential equations with one variable. Generally those can be written in the form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = ay(t) + f(t) \tag{7.36}$$

7.4.1 Solution goals

Before we get to discuss the solution of (7.36), we shall agree on what do we consider as a solution of an ordinary differential equation. On one hand (and commonly), any function $y(t)$ satisfying (7.36) would be understood as a solution of (7.36), but maybe that's not necessarily what we want? Alternatively, one could argue that (7.36) should be understood as a limiting case of the difference equation

$$\frac{\Delta y}{\Delta t} = ay(t) + f(t) \tag{7.37}$$

where t takes values from a discrete sequence $\{t_n\}$ with sequence steps $\Delta t = t_{n+1} - t_n$ approaching zero. This approach gets particularly intuitive if t has time semantics. In that case the equation (7.37) itself shall be understood in a causal sense, the value $y(t_{n+1})$ being defined by the preceding values via:

$$\frac{y_{n+1} - y_n}{t_{n+1} - t_n} = ay_n + f_n$$

(where $y_n = y(t_n)$, $f_n = f(t_n)$), or, explicitly expressing y_{n+1} via the rest of the values,

$$y_{n+1} = y_n + (ay_n + f_n)(t_{n+1} - t_n) \tag{7.38}$$

At least, if the equation (7.36) is to describe some physical process, e.g. charging of a capacitor, this latter interpretation makes lots of sense. From that perspective, we should not be interested just in "some function" satisfying the equation (7.36), but only in those which emerge as limits of the solutions given by (7.38) as $\Delta t \to 0$. Technically, this usually makes no difference, thanks to the uniqueness of the solution of the initial value problem, however conceptually the distinction is rather large.

▷ One could argue that this distinction is reflected even in the terminology. Solution of differential equations is referred to as *integration of the equations*. Now, integration is exactly what the difference scheme (7.38) is doing if Δt is indefinitely small.

7.4.2 Common solution approaches

The topic of solving first-order ordinary differential equations of the form (7.36) has been well studied and there is a formal process to that with some variations. In one common approach, one usually solves the homogeneous version of (7.36):

$$\frac{\mathrm{d}y}{\mathrm{d}t} = ay(t) \tag{7.39}$$

which is just "guessed" to have the solution of the form $Ce^{\lambda t}$, which, after substitution into (7.39), yields

$$C\lambda e^{\lambda t} = aCe^{\lambda t}$$

and, after cancelling the common factor $Ce^{\lambda t}$, leaves just

$$\lambda = a$$

thereby the solution is considered to be $y(t) = Ce^{at}$. If $f(t) \equiv 0$ in (7.36) (the original equation is already homogeneous), the value of C is obtained from the initial conditions. Otherwise, the solution to the full equation (7.36) is found by the method of parameter variation, where we again "guess" that the solution has the form $y(t) = C(t)e^{\lambda t}$.

In another approach we solve the equation (7.36) by analytically integrating it. But it doesn't make too much sense to integrate the sides of (7.36) in its original form. Instead, first we are transforming (7.36) to the form

$$\frac{\mathrm{d}y}{\mathrm{d}t} - ay(t) = f(t) \tag{7.40}$$

which looks a bit more promising in respect to being integrated, but not too much yet. At this point we "guess" that we might be able to multiply both sides of (7.40) by some other function (in this case e^{-at}), so that the left-hand side of (7.40) turns into a full derivative of yet another function (in this case ye^{-at}) and therefore (7.40) can be integrated in a useful manner. This is known as the *integrating factor* approach.

There is too much "guessing" involved in either case. Of course this "guessing" is not real guessing, it's more that we are given specific instructions for some nontrivial steps. But then we might be as well given just the solution formula (and there is one). Because we still have no idea how these "guesses" could have been made initially by the person(s) who came first with this approach. Some of these "guesses" could be imagined to have been obtained by some kind of enlightenment (in particular, guessing that the solution of the homogeneous equation is $Ce^{\lambda x}$ is not fully trivial, but maybe not *that* hard), while some others, like parameter variation, look like huge gaps of omitted information. We shall now instead try to solve the equation in a "gapless" way, not involving too many (if any at all) "guesses".

7.4.3 Constructive solution

Having expressed our dissatisfaction with the traditional ways to solve the equation (7.36) in Subsection 7.4.2, we shall now describe another approach which is more in line with our intentions here.

Goal 7.6. Solve (7.36) in a constructive manner. *Hint:* follow the same approach as for (7.3).

We shall take our previous considerations from Subsection 7.4.1 into account and obtain the solution of (7.36) as a limiting case of (7.38). First, let's solve the homogeneous version (7.39) of (7.36). This is not too different from the equation (7.3) and can be even seen as a further generalization of the switch from (7.3) to (7.15). Repeating the same steps as the ones done earlier in the solution of (7.3), instead of (7.9) we obtain

$$y(t) = y_0 \cdot \lim_{n\to\infty} \left(1 + \frac{at}{n}\right)^n$$

where, compared to (7.9), we have at in the place of t, and therefore instead of being equal to e^t, the limit is equal to e^{at}:

$$\lim_{n\to\infty} \left(1 + \frac{at}{n}\right)^n = e^{at} \tag{7.41}$$

Respectively the solution to (7.39) is simply

$$y(t) = y_0 \cdot e^{at} \tag{7.42}$$

Now let's consider a more complex case of non-homogeneous (7.36). Taking all the time steps in (7.38) to be identical: $t_{n+1} - t_n = \Delta t$, we rewrite (7.38) as

$$y_{n+1} = y_n + (ay_n + f_n)\Delta t$$

or

$$y_{n+1} = (1 + a\Delta t)y_n + f_n\Delta t$$

Therefore

$$\begin{aligned}
y_1 &= (1 + a\Delta t)y_0 + f_0\Delta t \\
y_2 &= (1 + a\Delta t)y_1 + f_1\Delta t \\
&= (1 + a\Delta t)\Big((1 + a\Delta t)y_0 + f_0\Delta t\Big) + f_1\Delta t \\
&= (1 + a\Delta t)^2 y_0 + (1 + a\Delta t)f_0\Delta t + f_1\Delta t \\
y_3 &= (1 + a\Delta t)y_2 + f_2\Delta t \\
&= (1 + a\Delta t)\Big((1 + a\Delta t)^2 y_0 + (1 + a\Delta t)f_0\Delta t + f_1\Delta t\Big) + f_2\Delta t \\
&= (1 + a\Delta t)^3 y_0 + (1 + a\Delta t)^2 f_0\Delta t + (1 + a\Delta t)f_1\Delta t + f_2\Delta t
\end{aligned}$$

There begins to emerge a pattern:

$$y_n = (1 + a\Delta t)^n y_0 + \sum_{i=0}^{n-1} (1 + a\Delta t)^{n-1-i} f_i \Delta t \tag{7.43}$$

Let's now fix the value of t, that is reached after n steps (therefore $t_n = t$). Then $\Delta t = t/n$ and

$$y_n = \left(1 + \frac{at}{n}\right)^n y_0 + \sum_{i=0}^{n-1} \left(1 + \frac{at}{n}\right)^{n-1-i} f_i \frac{t}{n} \tag{7.44}$$

Now we already know that at $n \to \infty$ the first term of (7.44) turns into $y_0 e^{at}$. Let's find out what the second term turns into. The summation sign should turn into an integral, the t/n factor at the end of the expression turning into $\mathrm{d}\tau$ with the integration limits over τ going from 0 to t. The values $f_i = f(ti/n)$ should turn into $f(\tau)$ (and respectively ti/n turn into τ). It remains to do something about the factor $(1 + at/n)^{n-1-i}$. Let's rewrite it as

$$\left(1 + \frac{at}{n}\right)^{n-1-i} = \left(1 + \frac{at}{n}\right)^{n\frac{n-1-i}{n}} = \left(\left(1 + \frac{at}{n}\right)^n\right)^{\frac{n-1-i}{n}}$$

Since $(1 + at/n)^n$ turns into e^{at} we can rewrite the above as

$$\left(e^{at}\right)^{\frac{n-1-i}{n}}$$

Now at $n \to \infty$ we can replace $n - 1$ with n, obtaining

$$\left(e^{at}\right)^{\frac{n-i}{n}} = \left(e^{at}\right)^{1-\frac{i}{n}} = e^{at(1-i/n)} = e^{at}e^{-a\cdot ti/n} = e^{at}e^{-a\tau} = e^{a(t-\tau)}$$

and finally, the second term of (7.44) turns into the integral

$$\int_0^t e^{a(t-\tau)} f(\tau)\,\mathrm{d}\tau$$

while the whole expression (7.44) becomes

(solution of the 1st-order ODE (7.36))

$$y(t) = y_0 \cdot e^{at} + \int_0^t e^{a(t-\tau)} f(\tau)\,\mathrm{d}\tau \tag{7.45}$$

This is the solution of (7.36) which we obtained without any "guessing", but just by directly taking the limit of the finite-step solution (7.43).

▷ The second term of (7.45) corresponding to the nonhomogeneous term $f(t)$ of (7.36) can be seen as the sum of exponentials of the form e^{at} where the amplitude of each exponential is the infinitely small value $f(\tau)\mathrm{d}\tau$ and respectively each exponential is running for the time from τ to t (amounting to $t - \tau$ in total). The second term thereby becomes essentially identical to the first term. It's just that the first term has a finite amplitude y_0 and is running for the entire time from 0 to t (amounting to t in total), whereas the second term consists of a sum of infinitely many exponentials of infinitely small amplitudes running from different starting times to the same target time t.

A closely related idea is to see the second term as a convolution of $f(t)$ with the function

$$h(t) = \begin{cases} 0 & \text{if } t < 0 \\ e^{at} & \text{if } t \geq 0 \end{cases}$$

Goal 7.6 completed. ■

7.5 Matrix exponential

Let's generalize the previous discussion of first order ordinary differential equations to the case of multiple variables. In that case it is convenient to represent all unknown variables (functions) as components of one vector. Starting with the homogeneous equation (7.39), we therefore now let y to be a vector consisting of N components, which turns a into an $N \times N$ matrix (in the general case):

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{A}\mathbf{y} \tag{7.46}$$

Goal 7.7. Solve (7.46) in a constructive manner. *Hint:* follow the ideas used to obtain the solution of (7.36), this time using matrix arithmetic.

Repeating the same steps that we did earlier to solve (7.39), this time, instead of arriving at the equation (7.41), we arrive at the matrix version of its left-hand side

$$\lim_{n\to\infty}\left(1+\frac{\mathbf{A}t}{n}\right)^n \tag{7.47}$$

▷ In (7.47), "1" is a short notation for $\mathbf{I}$, respectively $1+\mathbf{A}t/n$ means $\mathbf{I}+\mathbf{A}t/n$. The motivation for this notation is coming from understanding matrices as operators acting on vectors and producing other vectors as the result. So, if a product $\mathbf{M}\mathbf{v}$ of a matrix $\mathbf{M}$ and a vector $\mathbf{v}$ is understood as an operator $\mathbf{M}$ acting on vector $\mathbf{v}$, then we can also understand multiplication of a vector by a scalar in the same sense: in the product $\lambda\mathbf{v}$ the operator λ acts on vector $\mathbf{v}$ and produces another vector (equal to $\lambda\mathbf{v}$) as the result. Then $\lambda + \mathbf{M}$ is simply the sum of these two operators.

Apparently the operator corresponding to the matrix $\lambda\mathbf{I}$ has the same effect on vectors as the operator corresponding to the scalar λ, therefore $\lambda\mathbf{I}$ and λ denote the same operator, and in particular "$\mathbf{I}$" and "1" denote the same operator. Practically, whenever we mix scalar and matrix terms, scalars are implicitly promoted to matrices. E.g. in $\lambda + \mathbf{M}$ the scalar λ is implicitly promoted to the matrix $\lambda\mathbf{I}$, and in $1 + \mathbf{A}t/n$ the scalar "1" is promoted to the matrix $\mathbf{I}$.

For the sake of experimenting and simplicity, let's consider the case of $t = 1$ in (7.47):

$$\lim_{n\to\infty}\left(1+\frac{\mathbf{A}}{n}\right)^n \tag{7.48}$$

The expression (7.48) is pretty much the same as the one in (7.10a), just instead of n-th power of the number $1 + t/n$ we have n-th power of the matrix $1 + \mathbf{A}/n$ under the sign of limit.

It might be not 100% clear if the limit in (7.48) should exist. However, we can follow the derivation of (7.18) to rewrite this limit as the series:

$$\lim_{n\to\infty}\left(1+\frac{\mathbf{A}}{n}\right)^n = 1 + \mathbf{A} + \frac{\mathbf{A}^2}{2} + \ldots \frac{\mathbf{A}^k}{k!} + \ldots$$

Now this series looks like it is converging.

▷ We could introduce the matrix norm $\|\mathbf{A}\|$ defined as the maximum scaling of a vector length by the matrix $\mathbf{A}$ taken across all possible nonzero vectors:

$$\|\mathbf{A}\| = \max_{\mathbf{v}\neq 0}\left\{\frac{|\mathbf{A}\mathbf{v}|}{|\mathbf{v}|}\right\}$$

Obviously, this norm is going to satisfy

$$\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$$

and

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$$

Therefore the k-th term of the series is bound by $\|\mathbf{A}\|^k/k!$ and thus the norm of the terms should quickly decrease to zero as $k \to \infty$, at least at the same speed as $x^k/k!$ would for the number $x = \|\mathbf{A}\|$.

The limit (7.48) thereby exists and, since it's a direct counterpart of (7.10a), it is natural to denote its value as $e^{\mathbf{A}}$, that is we have introduced an exponential with a matrix argument:

(matrix exponential)

$$e^{\mathbf{A}} = \lim_{n\to\infty}\left(1+\frac{\mathbf{A}}{n}\right)^n = 1 + \mathbf{A} + \frac{\mathbf{A}^2}{2} + \ldots \frac{\mathbf{A}^k}{k!} + \ldots \tag{7.49}$$

Now, returning to the equation (7.46) and the expression (7.47), which we obtained while trying to solve (7.46), we conclude that the solution of (7.46) is obtained by generalizing the solution (7.42) of the equation (7.39) to the matrix expression

$$\mathbf{y} = e^{\mathbf{A}t} \cdot \mathbf{y}_0 \tag{7.50}$$

▷ Since matrix multiplication is not commutative, we should mind the order of the factors in the product $e^{\mathbf{A}t} \cdot \mathbf{y}_0$. The correct order $e^{\mathbf{A}t} \cdot \mathbf{y}_0$ emerges from

$$\begin{aligned} \mathbf{y}_{n+1} &= \mathbf{y}_n + \mathbf{A}\mathbf{y}_n\Delta t = (1+\mathbf{A}\Delta t)\mathbf{y}_n \\ \mathbf{y}_n &= (1+\mathbf{A}\Delta t)^n\mathbf{y}_0 \end{aligned} \tag{7.51}$$

Also the opposite-order product $\mathbf{y}_0 \cdot e^{\mathbf{A}t}$ simply doesn't exist due to matrix dimension mismatch.

▷ The expression (7.50) states that the value of $\mathbf{y}$ at each time moment t is obtained by multiplying the initial state $\mathbf{y}_0$ by the matrix $e^{\mathbf{A}t}$. The matrix $e^{\mathbf{A}t}$ is therefore referred to as *propagator*, since it "propagates" the initial state further in time.

A complementary way to look at (7.50) is to consider the columns of the matrix $e^{\mathbf{A}t}$ as a set of basis vectors (where each vector is actually a vector function of t), the components of $\mathbf{y}_0$ being the coefficients of the linear combination of those basis vectors. The space spanned by those basis vectors thereby contains all possible solutions of (7.46) for all possible initial conditions. The solutions of (7.46) thereby form an N-dimensional linear space and the columns of $e^{\mathbf{A}t}$ represent one possible basis of that space.

The non-homogeneous equation (7.36) in the matrix form respectively reads

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{A}\mathbf{y}(t) + \mathbf{f}(t) \tag{7.52}$$

and the solution (7.45) is turning into

(solution of multivariable ODE (7.52))

$$\mathbf{y}(t) = e^{\mathbf{A}t} \cdot \mathbf{y}_0 + \int_0^t e^{\mathbf{A}(t-\tau)} \mathbf{f}(\tau)\, \mathrm{d}\tau \tag{7.53}$$

Goal 7.7 completed. ■

7.6 Matrix diagonalization

▷ It is recommended that the readers get familiar with the ideas discussed in Chapter 4 before they read the subsequent text, as the ability to think in basis-independent terms could have a qualitative impact on the level of understanding of the following material.

The expression (7.49) is a bit too involved to compute. When adding two matrices, we need to add N^2 components instead of just one number and when multiplying them, the number of operations is of the order of N^3. If only all matrices $\mathbf{A}$ that we have to deal with had a very simple form...

The simplest is, of course, just the identity matrix (besides the zero matrix). Indeed, in such case we'd have $\mathbf{A}^n = \mathbf{I}\ \forall n$ and respectively the Taylor series in (7.49) would simply collapse into

$$e^{\mathbf{I}} = \mathbf{I} + \mathbf{I} + \frac{\mathbf{I}}{2} + \ldots \frac{\mathbf{I}}{k!} + \ldots = \mathbf{I} \cdot \left(1 + 1 + \frac{1}{2} + \ldots + \frac{1}{k!}\right) = \mathbf{I} \cdot e$$

Also having $\mathbf{A} = \mathbf{I}$ in (7.52) leads to a simple situation. In this case the exponentials which we need to compute are having the form

$$e^{\mathbf{I}t} = \mathbf{I} + \mathbf{I}t + \frac{\mathbf{I}^2 t^2}{2} + \ldots \frac{\mathbf{I}^k t^k}{k!} + \ldots = \mathbf{I} \cdot \left(1 + t + \frac{t}{2} + \ldots + \frac{t}{k!}\right) = \mathbf{I} \cdot e^t$$

The next simplest case, which is just a little bit more complicated than the identity matrix, is the case of a diagonal matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$$

In this case we still have to do more work compared to the single-dimensional case, but both the addition and multiplication of diagonal matrices has complexity N rather than N^2 and N^3, as we only have to add or multiply the diagonal elements of two matrices 1:1. This effectively means that the computation of (7.49) falls into N separate computations, one per diagonal element:

$$e^{\lambda_i} = 1 + \lambda_i + \frac{\lambda_i^2}{2} + \ldots \frac{\lambda_i^k}{k!} + \ldots = \lim_{n \to \infty} \left(1 + \frac{\lambda_i}{n}\right)^n$$

and we obtain a diagonal matrix of exponentials:

$$e^{\mathbf{A}} = \begin{bmatrix} e^{\lambda_1} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_N} \end{bmatrix}$$

The same applies to a diagonal matrix $\mathbf{A}$ in the differential equation (7.52), since this means that we are simply having the exponentials $e^{\lambda_i t}$ as the diagonal components of the diagonal matrix $e^{\mathbf{A}t}$. In fact, the entire differential equation (7.52) falls into N independent equations, one per each y_i and λ_i:

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = \lambda_i y_i(t) + f_i(t)$$

But then, this all seems too good to be useful in practice. After all, what are the chances of us having a multivariable differential equation system with a diagonal matrix? In fact, as mentioned, having a diagonal matrix in (7.52), simply means that all equations are *already* independent. So, what are the chances of our multivariable differential equation system to consist of independent equations? There are probably quite a few of such systems, but then, why even bother considering them as multidimensional ones?

Except that... there is this thing with basis change, which we discussed at great length in Chapter 4. The equation (7.52) looks like a honest vector equation, we could even rewrite it in tensor form:

$$\frac{\mathrm{d}y^i}{\mathrm{d}t} = A^i_j y^j(t) + f^i(t) \tag{7.54}$$

This means it could be potentially written in any basis in the N-dimensional space, and in some of the bases the components of the tensor A^i_j might look simpler than in the other bases. And, with some luck, maybe there is even a basis where the components of A^i_j form a diagonal matrix.

According to (7.54), A^i_j is an operator, therefore matrix $\mathbf{A}$ is transformed by a similarity transformation under a basis change. We are therefore looking for a similarity transformation of $\mathbf{A}$, which makes $\mathbf{A}$ diagonal.

Suppose we already are in a basis where A^i_j is diagonal. The components of each basis vector $\vec{e}_i$ are all zero except for a single non-zero component in the i-th row (this component being equal to 1). After the multiplication by A^i_j only this non-zero component will "get through the operator A^i_j", effectively being multiplied by λ_i. E.g. for $\vec{e}_2$ we would have

$$\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda_2 \\ \vdots \\ 0 \end{bmatrix}$$

That is $\mathbf{A} \cdot \mathbf{e}_i = \lambda_i \mathbf{e}_i$, or, in coordinate-abstract notation, where we denote the operator A^i_j as a vector-function of a vector argument, $\vec{A}(\vec{e}_i) = \lambda_i \vec{e}_i$. However, this latter equation is written in a basis-independent form. Well, it kind of does

explicitly contain the basis vector $\vec{e}_i$, but we could notate it just as some vector $\vec{v}$, not specifically claiming it to be a basis vector:

$$\vec{A}(\vec{v}) = \lambda\vec{v} \tag{7.55}$$

(where we also dropped the index for λ since it's not too relevant at the moment).

Equality (7.55) looks like an equation which we could try to solve, thereby hopefully obtaining a basis consisting of solutions $\vec{v}$ of that equation, the operator $\vec{A}(\vec{\cdot})$ becoming diagonal in this basis. Except that we need to solve for both $\vec{v}$ and λ, since originally we don't know either of those. How do we do that? Let's rewrite (7.55) in the matrix form:

(eigenvector equation)

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{7.56}$$

Now the readers probably recognize (7.56) as the eigenvector equation. So, we are simply looking for the eigenvectors of $\mathbf{A}$. Then, if we can make a basis of these eigenvectors (referred to as the *eigenbasis* of $\mathbf{A}$), the operator $\vec{A}(\vec{\cdot})$ is going to be diagonal in this basis. This is essentially the idea of diagonalization: apply a similarity transformation to convert the operator matrix $\mathbf{A}$ to its eigenbasis.

In the application to our original purpose of solving the multivariable differential equation (7.52), however, it doesn't suffice to just diagonalize the matrix $\mathbf{A}$. We need to transform the entire equation (7.52) to the new basis, so that instead of $\mathbf{A}$, $\mathbf{f}$ and $\mathbf{y}$ we get some other matrix and vectors $\mathbf{A}'$, $\mathbf{f}'$, $\mathbf{y}'$:

$$\frac{\mathrm{d}\mathbf{y}'}{\mathrm{d}t} = \mathbf{A}'\mathbf{y}'(t) + \mathbf{f}'(t) \tag{7.57}$$

▷ Apparently in (7.57) the primes denote the representations in the new basis rather than differentiation.

Notice that according to (7.54), both $\mathbf{f}$ and $\mathbf{y}$ are "honest vectors" (they have superscript indices), therefore they both transform contravariantly. Equation (7.57) looks the same as (7.52), however, this time the matrix $\mathbf{A}'$ is diagonal and thereby the equation (7.57) is easy to solve.

This is one of the main practical applications of diagonalization process. As the readers can see, it involves a bit more than just a formal diagonalization of the matrix.

Invariance of matrix exponentiation

The solution formula (7.53) applied to (7.57) looks like

$$\mathbf{y}'(t) = e^{\mathbf{A}'t} \cdot \mathbf{y}'_0 + \int_0^t e^{\mathbf{A}'(t-\tau)}\mathbf{f}'(\tau)\,\mathrm{d}\tau$$

On the other hand, by simply multiplying both sides of (7.53) by $\mathbf{T}^{-1}$, we have

$$\mathbf{y}'(t) = \mathbf{T}^{-1}\mathbf{y}(t) = \mathbf{T}^{-1}e^{\mathbf{A}t} \cdot \mathbf{y}_0 + \int_0^t \mathbf{T}^{-1}e^{\mathbf{A}(t-\tau)}\mathbf{f}(\tau)\,\mathrm{d}\tau$$

$$= \mathbf{T}^{-1}e^{\mathbf{A}t}\mathbf{T}\cdot\mathbf{T}^{-1}\mathbf{y}_0 + \int_0^t \mathbf{T}^{-1}e^{\mathbf{A}(t-\tau)}\mathbf{T}\mathbf{T}^{-1}\mathbf{f}(\tau)\,\mathrm{d}\tau$$

where $\mathbf{T}$ is the basis transformation matrix. This suggests that $e^{\mathbf{A}'t} = \mathbf{T}^{-1}e^{\mathbf{A}t}\mathbf{T}$, or, by letting $t = 1$, $e^{\mathbf{A}'} = \mathbf{T}^{-1}e^{\mathbf{A}}\mathbf{T}$. That is $e^{\mathbf{A}}$ *is transformed by the same similarity law as* $\mathbf{A}$:

(matrix exponential is invariant under similarity transformation)

$$e^{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}} = \mathbf{T}^{-1}e^{\mathbf{A}}\mathbf{T} \tag{7.58}$$

or, in other words, taking an exponential of an operator matrix is a covariant-form operation.

▷ To demonstrate (7.58) explicitly, we can notice that, by (7.49),

$$\begin{aligned} e^{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}} &= \mathbf{T}^{-1}\cdot 1\cdot\mathbf{T} + \mathbf{T}^{-1}\mathbf{A}\mathbf{T} + \frac{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\cdot\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}{2} + \dots \\ &= \mathbf{T}^{-1}\cdot\left(1 + \mathbf{A} + \frac{\mathbf{A}\cdot\mathbf{A}}{2} + \dots\right)\cdot\mathbf{T} = \mathbf{T}^{-1}e^{\mathbf{A}}\mathbf{T} \end{aligned}$$

Or, using the limit form,

$$e^{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}} = \lim_{n\to\infty}\left(1 + \frac{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}{n}\right)^n = \lim_{n\to\infty}\mathbf{T}^{-1}\left(1 + \frac{\mathbf{A}}{n}\right)^n\mathbf{T} = \mathbf{T}^{-1}e^{\mathbf{A}}\mathbf{T}$$

where we noticed that

$$\begin{aligned} \left(1 + \frac{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}{n}\right)^n &= \prod_{\nu=1}^{n}\left(1 + \frac{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}{n}\right) = \prod_{\nu=1}^{n}\mathbf{T}^{-1}\left(1 + \frac{\mathbf{A}}{n}\right)\mathbf{T} \\ &= \mathbf{T}^{-1}\cdot\prod_{\nu=1}^{n}\left(1 + \frac{\mathbf{A}}{n}\right)\cdot\mathbf{T} \end{aligned}$$

7.7 Second-order differential equation

We have come to the idea of diagonalizing a matrix as a way to simplify the computation of the matrix exponential in the differential equation solution formula (7.53). Let's try to apply this approach to solve a generic homogeneous second-order fixed-coefficient linear ordinary differential equation

$$y''(t) - 2py'(t) + qy(t) = 0 \tag{7.59}$$

Goal 7.8. Solve (7.59) by matrix diagonalization. *Hint:* first convert (7.59) to a matrix-form multivariable equation.

▷ We notated the coefficient at $y'(t)$ as $-2p$ just in order to get slightly simpler expressions in the transformations that follow. Exactly the same steps could have been done with this coefficient notated as p and respectively the equation notated as $y''(t) + py'(t) + qy(t) = 0$.

7.7.1 Common-case steps

Goal 7.8a. Find the eigenvalues and the eigenvectors of the matrix form of the equation (7.59).

Introducing the vector

$$\mathbf{y}(t) = \begin{bmatrix} y'(t) \\ y(t) \end{bmatrix} \tag{7.60}$$

we can rewrite (7.59) in the matrix form as

$$\begin{bmatrix} y''(t) \\ y'(t) \end{bmatrix} = \begin{bmatrix} 2p & -q \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} y'(t) \\ y(t) \end{bmatrix} \tag{7.61}$$

or simply

$$\mathbf{y}' = \mathbf{A}\mathbf{y} \tag{7.62a}$$

where

$$\mathbf{A} = \begin{bmatrix} 2p & -q \\ 1 & 0 \end{bmatrix} \tag{7.62b}$$

Now we would like to try to diagonalize $\mathbf{A}$. In order to do so we need to solve the eigenvector equation (7.56), thereby obtaining the eigenbasis of $\mathbf{A}$.

We solve (7.56) using the standard approach of solving the eigenvector equation. First we rewrite it as

$$(\mathbf{A} - \lambda)\mathbf{v} = 0 \tag{7.63}$$

Looking at (7.63) we notice that the matrix $\mathbf{A} - \lambda$ must be singular in order for (7.63) to be satisfied by a non-zero vector $\mathbf{v}$. Conversely, if $\mathbf{A} - \lambda$ is singular, then there is definitely some $\mathbf{v} \neq 0$ satisfying (7.63). Therefore we're looking to solve the equation

$$\det(\mathbf{A} - \lambda) = 0$$

which we can also equivalently rewrite as

$$\det(\lambda - \mathbf{A}) = 0 \tag{7.64}$$

in order to get the positive unity leading coefficient in the polynomial that the left-hand side of (7.64) represents.

The expression $\det(\lambda - \mathbf{A})$ is an N-th order polynomial in λ (where N is the number of space dimensions), referred to as the *characteristic polynomial* of the matrix $\mathbf{A}$. Therefore (7.64) is a polynomial equation of N-th order in λ. By the fundamental theorem of algebra, (7.64) has N solutions (some of which may be repeated and/or complex). Having found the solutions λ_i of the equation (7.64), we construct matrices $\mathbf{A} - \lambda_i$ and look for vectors $\mathbf{v}_i$ satisfying $(\mathbf{A} - \lambda_i)\mathbf{v}_i = 0$.

With (7.62b), equation (7.64) becomes

$$\det \begin{bmatrix} \lambda - 2p & q \\ -1 & \lambda \end{bmatrix} = 0$$

that is

$$(\lambda - 2p)\lambda + q = 0$$

or simply

$$\lambda^2 - 2p\lambda + q = 0 \tag{7.65}$$

▷ Notice that (7.65) is having the same form as (7.59), except that the derivatives of $y(t)$ have been replaced by powers of λ. Equation (7.65) is referred to as *characteristic equation* of (7.59).

It follows that

$$\lambda = p \pm \sqrt{p^2 - q} \tag{7.66}$$

Respectively (7.63), which we choose to rewrite as $(\lambda - \mathbf{A})\mathbf{v} = 0$, becomes

$$\begin{bmatrix} \lambda - 2p & q \\ -1 & \lambda \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -p \pm \sqrt{p^2 - q} & q \\ -1 & p \pm \sqrt{p^2 - q} \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

Now the rows of the matrix $\lambda - \mathbf{A}$ in the above must be linearly dependent, since $\det(\lambda - \mathbf{A}) = 0$, and we should be able to discard any of the two equations corresponding to the two rows of $\lambda - \mathbf{A}$.

▷ We can verify the linear dependence of the rows of $\lambda - \mathbf{A}$ by multiplying the second row of the matrix by $p \mp \sqrt{p^2 - q}$, yielding

$$\begin{bmatrix} -p \pm \sqrt{p^2 - q} & q \\ -p \pm \sqrt{p^2 - q} & q \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

In fact we should discard the equation corresponding to the first row of $\lambda - \mathbf{A}$ because at $q = 0$ this row may become fully equal to zero. We are thereby left with the equation corresponding to the second row of $\lambda - \mathbf{A}$:

$$-v_1 + \lambda v_2 = 0$$

or simply

$$v_1 = \lambda v_2 \tag{7.67}$$

which by letting $v_2 = 1$ yields

$$v_1 = \lambda = p \pm \sqrt{p^2 - q}$$

thus the components of the vector $\mathbf{v}$ are

$$\mathbf{v} = \begin{bmatrix} \lambda \\ 1 \end{bmatrix} = \begin{bmatrix} p \pm \sqrt{p^2 - q} \\ 1 \end{bmatrix} \tag{7.68}$$

Goal 7.8a completed. ■

At this point we get three different situations depending on the sign of $p^2 - q$.

7.7.2 Overdamped case

Goal 7.8b. Complete the solution of (7.59) for the case of two different real eigenvalues.

If $p^2 - q > 0$ (referred to as the *overdamped* case, because the *damping coefficient* p is "large"), we get two different *real* eigenvalues from (7.66) and two different respective eigenvectors from (7.68). We therefore can apply a

similarity transformation of $\mathbf{A}$ to its eigenbasis, converting $\mathbf{A}$ to the diagonal form:

$$\mathbf{A}' = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

where $\lambda_{1,2}$ are by (7.66).

Respectively, the solution formula (7.50) in the new basis becomes

$$\mathbf{y}'(t) = \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} \cdot \mathbf{y}'_0$$

(where, as usual in such cases, the primes on vectors denote a different basis and not differentiation), and thus

$$\mathbf{y}(t) = \mathbf{T} \cdot \mathbf{y}'(t) = \mathbf{T} \cdot \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} \cdot \mathbf{y}'_0$$

where $\mathbf{T}$ is the basis transformation matrix. This means that the components of the vector $\mathbf{y}(t)$ are simply linear combinations of the exponential functions $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$. In particular, this holds for the second component of $\mathbf{y}(t)$, which by (7.60) is simply the unknown function $y(t)$ that we're solving (7.59) for, therefore

$$y(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} \tag{7.69}$$

where C_1 and C_2 are the coefficients of the linear combination.

> ▷ There is a well-known solution *Ansatz* for (7.59) (or any homogeneous linear ODE with fixed coefficients, for that matter) by assuming that the solutions of the equation are taking the form $e^{\lambda t}$ or a linear combination thereof, effectively yielding the formula (7.69). However here we have obtained (7.69) explicitly, without any assumptions or "guesses" of that kind.

Knowing that the solution has the form (7.69), we do not actually need to explicitly convert to the eigenbasis of $\mathbf{A}$ and back, we could just use the initial condition $\mathbf{y}(0) = \mathbf{y}_0$ which in terms of $y(t)$ reads

$$\begin{cases} y'(0) = y'_0 \\ y(0) = y_0 \end{cases} \tag{7.70}$$

> ▷ Differently from primes on vectors and matrices, which we used to denote a different basis, the primes on scalars in (7.70) denote differentiation.

Substituting (7.69) into (7.70) yields

$$\begin{cases} \lambda_1 C_1 + \lambda_2 C_2 = y'_0 \\ C_1 + C_2 = y_0 \end{cases}$$

which is solved by

$$\begin{cases} C_1 = \dfrac{y'_0 - \lambda_2 y_0}{\lambda_1 - \lambda_2} \\ C_2 = \dfrac{y'_0 - \lambda_1 y_0}{\lambda_2 - \lambda_1} \end{cases} \tag{7.71}$$

Substituting (7.71) into (7.69) yields

(solution of the 2nd order differential equation (7.59) for the overdamped case)

$$y(t) = \frac{y_0' - \lambda_2 y_0}{\lambda_1 - \lambda_2} e^{\lambda_1 t} + \frac{y_0' - \lambda_1 y_0}{\lambda_2 - \lambda_1} e^{\lambda_2 t} \tag{7.72}$$

which is the standard way to write down the general solution of (7.59) for the case $p^2 - q > 0$.

Goal 7.8b completed. ■

7.7.3 Motivation for complex numbers

Goal 7.8c. Complete the solution of (7.59) for the case when there are no real eigenvalues.

In the *underdamped case* $p^2 - q < 0$, the equation (7.65) doesn't have real solutions. On the other hand, the solution of the original differential equation (7.59) still should exist, simply as a limit of the respective finite-difference scheme. Indeed, the finite-difference scheme (7.51) always has a solution, which is explicitly given by the scheme itself. Its limiting case must also exist, since it's simply given by the respective matrix exponential (7.50). And this limit must be real, since there are only real numbers involved in (7.51).

The lack of real solutions of (7.65) suggests that the equation's matrix $\mathbf{A}$ cannot be diagonalized by a *real* similarity transformation. And this can be seen as one of the main motivating cases for the introduction of complex numbers. We shall now discuss this motivation.

Generally we would like our set of numbers in use to be closed under the operations that we apply to these numbers. A simple example of that is the introduction of negative numbers: it makes the set of numbers closed under subtraction (in simple words, we should be able to subtract any number from any number and obtain a meaningful result). Whether negative numbers exist "for real" is a somewhat philosophical question, however they allow us to noticeably increase the uniformity of the argument that we have to make every once in a while.

E.g. consider the equation

$$ax + 10 = 10x + b$$

The solution of this equation would be typically given by the formula

$$x = \frac{b - 10}{a - 10}$$

However notice that for $a < 10$, $b < 10$ the intermediate values in this formula become negative, even though the final answer is still positive. Had we refused to deal with negative numbers, we would have at least had to derive and offer a second formula

$$x = \frac{10 - a}{10 - b}$$

which is to be used if $a < 10$, $b < 10$. The readers can imagine how complicated this could grow in certain situations, even if we are only interested in positive final results.

Another example is division by zero. The sets of real and rational numbers are closed under division, except for division by zero. However there are situations where, despite zero division in the intermediate computations, the final computation result is finite. This also requires taking extra steps, such as considering the cases of zero divisor separately and/or resolving uncertainties (such as the ones arising from dividing zero by zero) by taking the limit. Differently from subtraction where we managed to make the set of numbers fully closed under subtraction by introducing negative numbers, zero division doesn't have such a simple and elegant solution. Respectively the cases of division by zero are more cumbersome and involved to handle.

Complex numbers are addressing a similar situation: even though we might be interested only in purely real results (and very often we are), they are highly useful in representing the intermediate values. Our idea would be to make our set of numbers closed under the square root operation, so that we can use one and the same formula (7.66) regardless of the sign of $p^2 - q$. This is achieved by introducing a new number i such that $i^2 = 1$.

▷ Notice that it follows that $(-i)^2 = -1$ as well. In fact, we cannot distinguish between i and $-i$ at all. It's just that we agree to denote one of the solutions of the equation $x^2 = 1$ as i and the other solution is automatically denoted as $-i$, but we don't know, which is which. Which one of i and $-i$ is located in the upper and which one in the lower complex semiplane is also just a matter of convention. Not that it really matters, since the numbers are anyway imaginary (pun intended).

Notice that it's different from the distinction between negative and positive real numbers, which is a real distinction: squaring a positive number keeps it positive, while squaring a negative number flips its sign. On the other hand $i^2 = (-i)^2 = -1$, $e^{\pm i\pi/2} = \pm i$ etc.

▷ Historically, complex numbers were introduced to represent intermediate values emerging in Cardano's method of solving cubic equations. Here we are developing a similar motivation for complex numbers arising from solution of differential equations.

Complex exponentials

We won't go through the details of construction of complex number algebra here, it is pretty straightforward and this information can be easily found elsewhere. However we shall take a detailed look about computing exponentials of a complex argument.

It is well known that, by the Euler's formula,

$$e^{i\varphi} = \cos\varphi + i\sin\varphi \tag{7.73}$$

The formula (7.73) is customarily derived using the Taylor series-based definition (7.1) by formally substituting an imaginary x into the formula. This may be seen as overly formal and even arbitrary, also failing to provide any intuitively convincing argument as to why imaginary power implies rotation, other that the original Taylor series for $\exp x$ formally falling apart into the cosine and sine series. Instead, to be more convincing, let's apply the definition (7.2), which explicitly describes $\exp x$ in terms of powers.

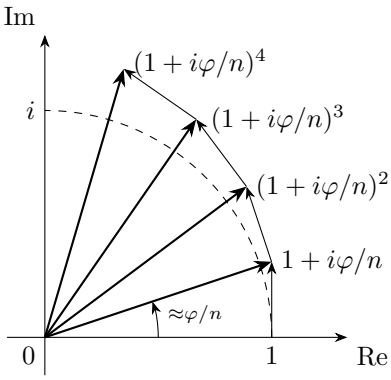


Figure 7.1: Approximated exponential of an imaginary argument.

Recall the idea of (7.35). For $a = e$ it simply reads

$$e^t \approx \left(1 + \frac{t}{n}\right)^n \tag{7.74}$$

Now let $t = i\varphi$ be a purely imaginary value. Then (7.74) becomes

(exponential of an imaginary argument approximated as a sequence of approximate rotations)

$$e^{i\varphi} \approx \left(1 + i\frac{\varphi}{n}\right)^n \tag{7.75}$$

Since φ/n is small, each of the factors $1 + i\varphi/n$ approximately represents a rotation by an angle equal to φ/n radian (Fig. 7.1 illustrates). This rotation is combined with a minor increase in the distance from the origin, but this increase has a smaller order of magnitude compared to the rotation and can be ignored as $n \to \infty$. On the other hand, the rotations sum up to a nonzero angle $n \cdot (\varphi/n) = \varphi$ and this is exactly the effect of computing $e^{i\varphi}$, which algebraically is described by the Euler's formula (7.73).

▷ More precisely, the rotations are by $\arctan \frac{\varphi}{n}$ radian, but, since φ/n is small, we have

$$\arctan \frac{\varphi}{n} = \frac{\varphi}{n} + o(1/n)$$

therefore the combined effect of n rotations is

$$n \arctan \frac{\varphi}{n} = n\frac{\varphi}{n} + n \cdot o(1/n) = \varphi + o(1)$$

Similarly, each change of the distance to the origin is by the factor

$$\sqrt{1 + \left(\frac{\varphi}{n}\right)^2} = 1 + \frac{1}{2}\left(\frac{\varphi}{n}\right)^2 + o(1/n^2)$$

(where we performed the Taylor expansion of $\sqrt{x}$ at $x = 1$) and the combined effect of n such changes is

$$\left(1 + \frac{1}{2}\left(\frac{\varphi}{n}\right)^2 + o(1/n^2)\right)^n = 1 + n\frac{1}{2}\left(\frac{\varphi}{n}\right)^2 + o(1/n) = 1 + o(1)$$

If the exponent is not purely imaginary, but essentially complex: $x = u + i\varphi$, then (7.74) becomes

$$e^{u+i\varphi} \approx \left(\left(1 + \frac{u}{n}\right) + i\frac{\varphi}{n}\right)^n$$

This time the change of the distance to the origin is not negligible and accumulates in the usual way into

$$\left(1 + \frac{u}{n}\right)^n \approx e^u$$

▷ More precisely

$$\begin{aligned}
\arctan \frac{\varphi/n}{1 + u/n} &= \arctan\left(\frac{\varphi}{n}\left(1 + \frac{1}{1 + u/n} - 1\right)\right) \\
&= \arctan\left(\frac{\varphi}{n}\left(1 - \frac{u/n}{1 + u/n}\right)\right) = \arctan\left(\frac{\varphi}{n}\left(1 + o(1)\right)\right) \\
&= \frac{\varphi}{n} + o(1/n)
\end{aligned}$$

so, we're effectively having rotations by the same angle φ/n. As to the distance to the origin,

$$\sqrt{\left(1 + \frac{u}{n}\right)^2 + \left(\frac{\varphi}{n}\right)^2} = \sqrt{1 + 2\frac{u}{n} + O(1/n^2)} = 1 + \frac{u}{n} + O(1/n^2)$$

where we already have seen that $O(1/n^2)$ doesn't affect the accumulated length.

Trigonometric and hyperbolic functions

From (7.73) we can express cosine and sine via an exponential of an imaginary argument. Substituting $-\varphi$ for φ in (7.73) yields

$$e^{-i\varphi} = \cos\varphi - i\sin\varphi \tag{7.76}$$

(since cosine is even and sine is odd function). Then, adding (7.73) and (7.76) and dividing by 2 yields

$$\cos\varphi = \frac{e^{i\varphi} + e^{-i\varphi}}{2} \tag{7.77a}$$

while subtracting (7.76) from (7.73) and dividing by $2i$ yields

$$\sin\varphi = \frac{e^{i\varphi} - e^{-i\varphi}}{2i} \tag{7.77b}$$

Removing i from the exponentials in (7.77) and from the denominator of (7.77b) (so that the result is real for $\varphi \in \mathbb{R}$) yields *hyperbolic cosine and sine* functions

$$\cosh\varphi = \frac{e^{\varphi} + e^{-\varphi}}{2} \tag{7.78a}$$

$$\sinh\varphi = \frac{e^{\varphi} - e^{-\varphi}}{2} \tag{7.78b}$$

so that

$$\cosh i\varphi = \cos\varphi \tag{7.79a}$$

$$\sinh i\varphi = i\sin\varphi \tag{7.79b}$$

7.7.4 Underdamped case

Now that we introduced complex numbers we can simply try to reuse the already obtained solution formula (7.72) for the underdamped case $p^2 - q < 0$. In this case (7.66) delivers two complex-conjugate values. However it would be suspicious if the result obtained from (7.72) was complex, after all our differential equation (7.59) describes a purely real process (assuming real p, q, y_0 and y_0'). Also a finite-difference scheme approximating this equation would be purely real as well.

We have said earlier that our intention with complex numbers in the solution process of (7.59) was to represent intermediate results. Maybe the final result is still purely real. In order to try to exploit the complex-conjugate symmetry of the results of (7.66), let's rewrite (7.72) in terms of hyperbolic functions:

$$\begin{aligned}
y(t) &= \frac{y_0' - \lambda_2 y_0}{\lambda_1 - \lambda_2} e^{\lambda_1 t} + \frac{y_0' - \lambda_1 y_0}{\lambda_2 - \lambda_1} e^{\lambda_2 t} \\
&= \frac{y_0' - (p - \sqrt{p^2 - q}) y_0}{2\sqrt{p^2 - q}} e^{(p + \sqrt{p^2 - q})t} + \frac{y_0' - (p + \sqrt{p^2 - q}) y_0}{-2\sqrt{p^2 - q}} e^{(p - \sqrt{p^2 - q})t} \\
&= \frac{e^{pt}}{2\sqrt{p^2 - q}} \cdot \Big(\big(y_0' - (p - \sqrt{p^2 - q}) y_0\big) e^{t\sqrt{p^2 - q}} \\
&\qquad\qquad - \big(y_0' - (p + \sqrt{p^2 - q}) y_0\big) e^{-t\sqrt{p^2 - q}} \Big) \\
&= \frac{e^{pt}}{\sqrt{p^2 - q}} \cdot \left((y_0' - p y_0) \sinh\Big(t\sqrt{p^2 - q}\Big) + y_0 \sqrt{p^2 - q} \cdot \cosh\Big(t\sqrt{p^2 - q}\Big) \right) \\
&= e^{pt} \cdot \left(y_0 \cosh\Big(t\sqrt{p^2 - q}\Big) + (y_0' - p y_0) \frac{\sinh\Big(t\sqrt{p^2 - q}\Big)}{\sqrt{p^2 - q}} \right) \qquad (7.80)
\end{aligned}$$

Next, using (7.79), we replace hyperbolic cosine and sine in (7.80) by their trigonometric versions:

$$\cosh\Big(t\sqrt{p^2 - q}\Big) = \cos\Big(t\sqrt{q - p^2}\Big) \qquad (7.81a)$$

$$\frac{\sinh\Big(t\sqrt{p^2 - q}\Big)}{\sqrt{p^2 - q}} = \frac{\sin\Big(t\sqrt{q - p^2}\Big)}{\sqrt{q - p^2}} \qquad (7.81b)$$

Thereby (7.80) turns into

(solution of the 2nd order differential equation (7.59) for the underdamped case)

$$y(t) = e^{pt} \cdot \left(y_0 \cos\Big(t\sqrt{q - p^2}\Big) + (y_0' - p y_0) \frac{\sin\Big(t\sqrt{q - p^2}\Big)}{\sqrt{q - p^2}} \right) \qquad (7.82)$$

It is easy to see that the expression (7.82) is purely real if $p^2 < q$.

▷ Same as the solution (7.72) is formally applicable for $p^2 < q$, the solution (7.82) is formally applicable for $p^2 > q$, it's just that there would be some imaginary intermediate values.

Goal 7.8c completed. ■

7.7.5 Critically-damped case

Goal 7.8d. Complete the solution of (7.59) for the boundary case between two real eigenvalues and no real eigenvalues.

In the *critically-damped case* $p^2 - q = 0$, the eigenvectors (7.68) are identical and we cannot diagonalize the matrix $\mathbf{A}$, as the (effectively) single obtained eigenvector cannot make a basis. We could however notice that the solution of (7.59) must be continuous with respect to the parameters p and q (small changes of p and/or q should result in small changes in the solution $y(t)$). This should follow from the continuity of the formula (7.50) with respect to small changes of matrix $\mathbf{A}$. We could therefore take the solution (7.72) and let $(p^2 - q) \to 0$, thereby obtaining the solution for $p^2 - q = 0$ as the limiting case of (7.72). Actually, instead of taking the limit (7.72), it's easier to take the limit of its equivalent form (7.80) or (7.82).

Taking (7.82) and letting $(p^2 - q) \to 0$ we notice the "zero divided by zero" uncertainty in the second term, since both the sine and the denominator approach zero. This uncertainty can be resolved using the *cardinal sine* function $\operatorname{sinc} t$ as follows:

$$\begin{aligned} y(t) &= e^{pt} \cdot \left(y_0 \cos\left(t\sqrt{q-p^2}\right) + (y_0' - py_0) \frac{\sin\left(t\sqrt{q-p^2}\right)}{\sqrt{q-p^2}} \right) \\ &= e^{pt} \cdot \left(y_0 \cos\left(t\sqrt{q-p^2}\right) + (y_0' - py_0) t \frac{\sin\left(t\sqrt{q-p^2}\right)}{t\sqrt{q-p^2}} \right) \\ &= e^{pt} \cdot \left(y_0 \cos\left(t\sqrt{q-p^2}\right) + (y_0' - py_0) t \operatorname{sinc}\left(t\sqrt{q-p^2}\right) \right) \end{aligned} \tag{7.83}$$

where

$$\operatorname{sinc} t = \lim_{t\to 0} \frac{\sin t}{t} = \begin{cases} \frac{\sin t}{t} & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

is the *cardinal sine* function. Taking the limit of (7.83) at $(p^2 - q) \to 0$ and noticing that $\cos 0 = 1$ and $\operatorname{sinc} 0 = 1$ we obtain

$$y(t) = e^{pt} \cdot (y_0 + (y_0' - py_0)t) = y_0 e^{pt} + (y_0' - py_0) \cdot te^{pt} \tag{7.84}$$

Formula (7.84) is a particular case of the generic form $y(t) = C_1 e^{pt} + C_2 te^{pt}$ which is a well-known solution *Ansatz* for the equation (7.59) in case $p^2 = q$ (that is $\lambda_1 = \lambda_2$). However here we have obtained (7.84) by explicit derivation rather than guessing.

▷ We could have obtained (7.84) using the hyperbolic solution formula (7.80) instead of the trigonometric one (7.82). Taking (7.80) we rewrite it along the lines of (7.83):

$$\begin{aligned} y(t) &= e^{pt} \cdot \left(y_0 \cosh\left(t\sqrt{p^2-q}\right) + (y_0' - py_0)\frac{\sinh\left(t\sqrt{p^2-q}\right)}{\sqrt{p^2-q}} \right) \\ &= e^{pt} \cdot \left(y_0 \cosh\left(t\sqrt{p^2-q}\right) + (y_0' - py_0)t\frac{\sinh\left(t\sqrt{p^2-q}\right)}{t\sqrt{p^2-q}} \right) \\ &= e^{pt} \cdot \left(y_0 \cosh\left(t\sqrt{p^2-q}\right) + (y_0' - py_0)t\,\mathrm{sinhc}\left(t\sqrt{p^2-q}\right) \right) \end{aligned} \tag{7.85}$$

where

$$\mathrm{sinhc}\, t = \lim_{t\to 0} \frac{\sinh t}{t} = \begin{cases} \frac{\sinh t}{t} & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

is the "cardinal hyperbolic sine" function. Taking the limit of (7.85) at $(p^2 - q) \to 0$ and noticing that $\cosh 0 = 1$ and $\mathrm{sinhc}\, 0 = 1$ we obtain (7.84). Obtaining one and the same result as the limiting cases of (7.80) and (7.82) corresponds to the solution continuity from both sides $(p^2 - q) \to +0$ and $(p^2 - q) \to -0$.

▷ We could have arrived at (7.84) in a different fashion. At $q = p^2$ the equation (7.59) turns into

$$y'' - 2py' + p^2 y = 0 \tag{7.86}$$

Introducing the function $u(t)$ satisfying $u' = pu$ we could replace (7.86) by

$$y' = py + u \tag{7.87}$$

Indeed, if y satisfies (7.87) then

$$y'' = (py + u)' = py' + u' = py' + pu = py' + p(y' - py) = 2py' - p^2 y$$

which is the same as (7.86).

However, in order to come up with this idea, one might need to utilize the concept of *Jordan chains*, since this is what the functions $y(t)$ and $u(t)$ represent.

Goal 7.8d completed. ■

Goal 7.8 completed. ■

7.8 Jordan normal form

In Subsection 7.7.5 we have utilized the idea of continuity of the solution of a differential equation with respect to its parameters in order to apply the diagonalization technique to a matrix which is not diagonalizable, by treating it as a limiting case of a diagonalizable case. However we have done so just for one specific matrix in a 2-dimensional space. We would like therefore to generalize the previous observations to arbitrary differential equations of the form (7.46). This generalization appears as the so called *Jordan normal form*.

▷ In this text we shall concentrate on how Jordan normal form emerges, rather than on the details of its definition. Readers who are not familiar with the idea of Jordan normal form yet, are advised to at least briefly consult other sources before reading on.

Goal 7.9. Construct Jordan normal form as a limiting case fallback of matrix diagonalization.

7.8.1 Rank 2 generalized eigenvectors

Goal 7.9a. Find how the limiting behavior discussed in Subsection 7.7.5 can be expressed in eigenvector terms. *Hint:* the linear space of solutions of the equation (7.59) is always two-dimensional. If $\lambda_1 \neq \lambda_2$, then one possible basis for this space is provided by the functions $e^{\lambda_1 t}$, $e^{\lambda_2 t}$ (acting as basis vectors in this solution space), the expansion of an arbitrary solution $y(t)$ in this basis being given by (7.69). If $\lambda_1 = \lambda_2$, these functions become identical and cannot serve as a basis, so we need a different basis choice.

We begin by trying to make some further observations about the situation discussed in Section 7.7 and, more specifically, in Subsection 7.7.5, this time concentrating on the behavior of the matrix's eigenvectors. Let's introduce the notation $\lambda = p$ and $\Delta\lambda = \sqrt{p^2 - q}$ where p and q are by (7.59). The eigenvalues of $\mathbf{A}$ given by (7.66) are thereby $\lambda_{1,2} = \lambda \pm \Delta\lambda$ and the eigenvectors of $\mathbf{A}$ given by (7.68) are

$$\mathbf{v}_{1,2} = \begin{bmatrix} \lambda \pm \Delta\lambda \\ 1 \end{bmatrix}$$

We thus have

$$\lambda_1 = \lambda + \Delta\lambda \qquad \mathbf{v}_1 = \begin{bmatrix} \lambda + \Delta\lambda \\ 1 \end{bmatrix}$$

$$\lambda_2 = \lambda - \Delta\lambda \qquad \mathbf{v}_2 = \begin{bmatrix} \lambda - \Delta\lambda \\ 1 \end{bmatrix}$$

At $(p^2 - q) \to 0$ we are having $\Delta\lambda \to 0$, $\lambda_{1,2} \to \lambda$ and $\mathbf{v}_{1,2} \to \mathbf{v}$ where

$$\mathbf{v} = \begin{bmatrix} \lambda \\ 1 \end{bmatrix}$$

Thus the eigenvectors become identical, we effectively have only the eigenvector $\mathbf{v}$ (with the eigenvalue λ) left, which cannot serve as a basis of the 2D space, we need one more vector. It is natural and simplest to pick up the vector $\mathbf{v}' = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ as the second basis vector in this case, since it's guaranteed to be different from $\mathbf{v}$ regardless of λ. Also $\mathbf{v}'$ can be expressed via $\mathbf{v}_{1,2}$:

$$\mathbf{v}' = \frac{\mathbf{v}_1 - \mathbf{v}_2}{2 \cdot \Delta\lambda} \tag{7.88a}$$

At the same time we can also express $\mathbf{v}$ via $\mathbf{v}_{1,2}$:

$$\mathbf{v} = \frac{\mathbf{v}_1 + \mathbf{v}_2}{2} \tag{7.88b}$$

Fig. 7.2 illustrates.

Equations 7.88 effectively provide a conversion from the eigenbasis $\mathbf{v}_1$, $\mathbf{v}_2$ to the basis $\mathbf{v}$, $\mathbf{v}'$. We can make the result of this conversion formally independent from $\Delta\lambda$ by taking the limits:

$$\mathbf{v} = \lim_{\Delta\lambda \to 0} \frac{\mathbf{v}_1 + \mathbf{v}_2}{2} \qquad \mathbf{v}' = \lim_{\Delta\lambda \to 0} \frac{\mathbf{v}_1 - \mathbf{v}_2}{2 \cdot \Delta\lambda} \tag{7.89}$$

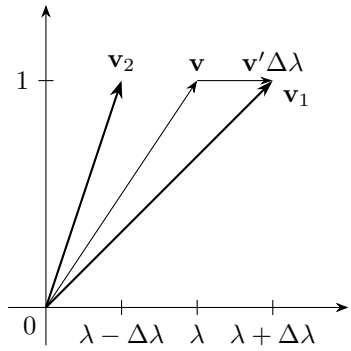


Figure 7.2: The eigenbasis and the limiting generalized eigenbasis.

Since $\mathbf{v}$ is the eigenvector of $\mathbf{A}$ at $\Delta\lambda = 0$, equation (7.56) applies: $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Even though $\mathbf{v}'$ is not an eigenvector, it's obtained from two eigenvectors via (7.89). Let's try to apply $\mathbf{A}$ to $\mathbf{v}'$ at $\Delta\lambda \to 0$. Let

$$\Delta\mathbf{v} = \frac{\mathbf{v}_1 - \mathbf{v}_2}{2} = \begin{bmatrix} \Delta\lambda \\ 0 \end{bmatrix}$$

therefore

$$\mathbf{v}_1 = \mathbf{v} + \Delta\mathbf{v}$$

and

$$\mathbf{v}' = \lim_{\Delta\lambda\to 0} \frac{\mathbf{v}_1 - \mathbf{v}}{\Delta\lambda} = \lim_{\Delta\lambda\to 0} \frac{\Delta\mathbf{v}}{\Delta\lambda}$$

Applying $\mathbf{A}$ to $\mathbf{v}'$ yields

$$\begin{aligned}
\mathbf{A}\mathbf{v}' &= \mathbf{A} \cdot \lim_{\Delta\lambda\to 0} \frac{\mathbf{v}_1 - \mathbf{v}}{\Delta\lambda} = \lim_{\Delta\lambda\to 0} \frac{\mathbf{A}\mathbf{v}_1 - \mathbf{A}\mathbf{v}}{\Delta\lambda} = \lim_{\Delta\lambda\to 0} \frac{\lambda_1\mathbf{v}_1 - \lambda\mathbf{v}}{\Delta\lambda} \\
&= \lim_{\Delta\lambda\to 0} \frac{(\lambda + \Delta\lambda)\mathbf{v}_1 - \lambda\mathbf{v}}{\Delta\lambda} = \lim_{\Delta\lambda\to 0} \left(\frac{\lambda(\mathbf{v}_1 - \mathbf{v})}{\Delta\lambda} + \mathbf{v}_1 \right) \\
&= \lambda \cdot \lim_{\Delta\lambda\to 0} \frac{\mathbf{v}_1 - \mathbf{v}}{\Delta\lambda} + \lim_{\Delta\lambda\to 0} \mathbf{v}_1 = \lambda\mathbf{v}' + \mathbf{v}
\end{aligned}$$

▷ In the above transformations we have been a bit nonchalant about the dependency of $\mathbf{A}$ on $\Delta\lambda$. More formally, we could have written this dependency explicitly as $\mathbf{A}(\Delta\lambda)$. In this case $\mathbf{A}$ under the limit sign turns into $\mathbf{A}(\Delta\lambda)$, while $\mathbf{A}$ outside of the limit sign turns into $\mathbf{A}(0)$. On the other hand, by the continuity of $\mathbf{A}(\Delta\lambda)$, with some care we are also free to write $\mathbf{A}(0)$ under the limit sign, if we need to.

That is

$$\mathbf{A}\mathbf{v}' = \lambda\mathbf{v}' + \mathbf{v} \tag{7.90a}$$

Notice that the equation (7.90a), that the vector $\mathbf{v}'$ satisfies, looks very similar to the eigenvector equation (7.56). Let's see how many different solutions (for the unknown $\mathbf{v}'$) does (7.90a) have. Rewriting (7.90a) as

$$(\mathbf{A} - \lambda)\mathbf{v}' = \mathbf{v} \tag{7.90b}$$

we obtain an equation similar to (7.63).

Now, having a single eigenvector means that (7.63) has a 1-dimensional space of solutions, thus rank $(\mathbf{A} - \lambda) = 2 - 1 = 1$. Therefore (7.90) also has a 1-dimensional space of solutions. It is easily noticed that this dimension stretches along the direction of the vector $\mathbf{v}$, since, given $\mathbf{v}'$ which satisfies (7.90), any vector of the form $\mathbf{v}' + \alpha\mathbf{v}$ also satisfies (7.90):

$$(\mathbf{A} - \lambda)(\mathbf{v}' + \alpha\mathbf{v}) = (\mathbf{A} - \lambda)\mathbf{v}' + (\mathbf{A} - \lambda)\alpha\mathbf{v} = (\mathbf{A} - \lambda)\mathbf{v}' + 0 = \mathbf{v}$$

All solutions of (7.90) are different from (in the sense that they are not collinear to) the eigenvector $\mathbf{v}$, since otherwise we would have had $(\mathbf{A} - \lambda)\mathbf{v}' = 0$ rather than $(\mathbf{A} - \lambda)\mathbf{v}' = \mathbf{v} \neq 0$. Therefore any solution of (7.90) can be taken as a second basis vector to complement the eigenvector $\mathbf{v}$ as the first basis vector.

Furthermore, given $\mathbf{v}'$ satisfying (7.90), we can change the length of this $\mathbf{v}'$ by some factor $\lambda' \neq 0$ and the resulting vector will obviously still usable as a basis vector. The equation (7.90b) rewritten to accommodate this vector with the changed length becomes

$$(\mathbf{A} - \lambda)\mathbf{v}' = \lambda'\mathbf{v} \tag{7.91a}$$

and (7.90a) respectively becomes

$$\mathbf{A}\mathbf{v}' = \lambda\mathbf{v}' + \lambda'\mathbf{v} \tag{7.91b}$$

(where λ' is an arbitrary nonzero value: $\lambda' \neq 0$).

▷ Apparently, another way to look at (7.91) is to notice that the eigenvector $\mathbf{v}$ can be replaced by $\lambda\mathbf{v}$, the latter also being an eigenvector.

The solutions $\mathbf{v}'$ of (7.91) thereby fill the entire 2D space with the exception of the line where the eigenvectors of $\mathbf{A}$ lie. Thereby any of these solutions can be used to provide a second basis vector to complement the eigenvector $\mathbf{v}$ as the first basis vector. The vectors $\mathbf{v}'$ satisfying (7.91) are referred to as *rank 2 generalized eigenvectors* of the matrix $\mathbf{A}$. Therefore, in the case of $\mathbf{v}_1 = \mathbf{v}_2$, a basis can be constructed from an eigenvector $\mathbf{v}$ and a generalized eigenvector $\mathbf{v}'$.

▷ Commonly the equation (7.90) is used instead of (7.91) as the definition of rank 2 generalized eigenvectors, however that has a sole purpose of ensuring a unique form of $\mathbf{A}$ in the basis consisting of $\mathbf{v}$ and $\mathbf{v}'$. Equation (7.90) will still allow the generalized eigenvector to go in any direction (except in the direction of $\mathbf{v}$), but will put a constraint on the length of $\mathbf{v}'$, uniquely specifying the length for each given direction. As long as we're not concerned with $\mathbf{A}$ attaining a unique form, we can take (7.91) as the definition of rank 2 generalized eigenvectors, thereby those vectors filling almost the entire 2D space.

We can map the just obtained results to the differential equation solution (7.84) obtained earlier. Indeed, in (7.84) we have switched from the basis functions $e^{\lambda_1 t}$, $e^{\lambda_2 t}$ to the basis functions e^{pt}, te^{pt}. Now, recalling that we introduced the notation $\lambda = p$, we can rewrite the latter basis as $e^{\lambda t}$, $te^{\lambda t}$. We can also notice that by (7.62a) the multiplication by $\mathbf{A}$ should correspond to differentiation. In particular, $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ corresponds to $\frac{\mathrm{d}}{\mathrm{d}t}e^{\lambda t} = \lambda e^{\lambda t}$, so that $e^{\lambda t}$ corresponds to $\mathbf{v}$. Now, if we differentiate $te^{\lambda t}$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(te^{\lambda t}) = \lambda \cdot te^{\lambda t} + e^{\lambda t}$$

Notice how the latter equation exactly matches (7.90a).

Further, in (7.85) the basis function $te^{\lambda t}$ is obtained as the limit

$$\begin{aligned}\lim_{\Delta\lambda\to 0}\frac{e^{\lambda_1 t}-e^{\lambda_2 t}}{2\cdot\Delta\lambda} &= \lim_{\Delta\lambda\to 0}\frac{e^{(\lambda+\Delta\lambda)t}-e^{(\lambda-\Delta\lambda)t}}{2\cdot\Delta\lambda} = e^{\lambda t}\cdot\lim_{\Delta\lambda\to 0}\frac{e^{\Delta\lambda t}-e^{-\Delta\lambda t}}{2\cdot\Delta\lambda}\\ &= e^{\lambda t}\cdot\lim_{\Delta\lambda\to 0}\frac{e^{\Delta\lambda t}-e^{-\Delta\lambda t}}{2\cdot\Delta\lambda t}t = te^{\lambda t}\cdot\lim_{\Delta\lambda\to 0}\frac{e^{\Delta\lambda t}-e^{-\Delta\lambda t}}{2\cdot\Delta\lambda t}\\ &= te^{\lambda t}\cdot\lim_{\Delta\lambda\to 0}\operatorname{sinhc}(\Delta\lambda t) = te^{\lambda t}\cdot 1 = te^{\lambda t}\end{aligned}$$

This exactly corresponds to $\mathbf{v}'$ being obtained as the limit of $(\mathbf{v}_1-\mathbf{v}_2)/2\Delta\lambda$. Generally, in some way there is a correspondence between basis functions and vector expressions

$$\begin{aligned}
&e^{\lambda_1 t} = e^{(\lambda+\Delta\lambda)t} && \longleftrightarrow && \mathbf{v}_1\\
&e^{\lambda_2 t} = e^{(\lambda-\Delta\lambda)t} && \longleftrightarrow && \mathbf{v}_2\\
&e^{\lambda t} && \longleftrightarrow && \mathbf{v}\\
&te^{\lambda t} && \longleftrightarrow && \mathbf{v}'\\
&\frac{e^{\lambda_1 t}+e^{\lambda_2 t}}{2} = e^{\lambda t}\cosh(\Delta\lambda t) && \longleftrightarrow && \frac{\mathbf{v}_1+\mathbf{v}_2}{2}\\
&\frac{e^{\lambda_1 t}-e^{\lambda_2 t}}{2} = e^{\lambda t}\sinh(\Delta\lambda t) && \longleftrightarrow && \frac{\mathbf{v}_1-\mathbf{v}_2}{2}\\
&\frac{e^{\lambda_1 t}+e^{\lambda_2 t}}{2\Delta\lambda} = te^{\lambda t}\operatorname{sinhc}(\Delta\lambda t) && \longleftrightarrow && \frac{\mathbf{v}_1-\mathbf{v}_2}{2\Delta\lambda}
\end{aligned}$$

Goal 7.9a completed. ■

7.8.2 Rank 3 generalized eigenvectors

Goal 7.9b. Extend the results just obtained for a 2×2 matrix to 3×3 matrices, where all 3 eigenvectors will coincide in the limit.

In Subsection 7.8.1 we provided an analysis of the limiting behavior of the eigenvectors arising in the solution of the 2nd order differential equation (7.59). Let's try to see how does the situation of coinciding eigenvectors appear in case of the 3rd order differential equation:

$$y''' + py'' + qy' + ry = 0 \tag{7.92}$$

It is easy to generalize the observations obtained in Section 7.7: for a 3rd order differential equation, the diagonalized matrix will contain the eigenvalues λ_1, λ_2 and λ_3 on the main diagonal, yielding the solution of the initial value problem in the form $C_1e^{\lambda_1 t}+C_2e^{\lambda_2 t}+C_3e^{\lambda_3 t}$. By substituting the basis solution functions $e^{\lambda_1 t}$, $e^{\lambda_2 t}$ and $e^{\lambda_3 t}$ into (7.92) we are going to obtain the characteristic equation

$$\lambda^3 + p\lambda^2 + q\lambda + r = 0 \tag{7.93}$$

which should be satisfied by the eigenvalues λ_1, λ_2, λ_3.

We can also obtain the characteristic equation (7.93) explicitly by converting (7.92) to the matrix form, yielding

$$\begin{bmatrix} y''' \\ y'' \\ y' \end{bmatrix} = \begin{bmatrix} -p & -q & -r \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} y'' \\ y' \\ y \end{bmatrix} \tag{7.94}$$

thus

$$\mathbf{A} = \begin{bmatrix} -p & -q & -r \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{7.95}$$

and the eigenvalue equation is

$$\begin{aligned} \det(\lambda - \mathbf{A}) = \det \begin{bmatrix} \lambda + p & q & r \\ -1 & \lambda & 0 \\ 0 & -1 & \lambda \end{bmatrix} &= 1 \cdot \det \begin{bmatrix} \lambda + p & r \\ -1 & 0 \end{bmatrix} + \lambda \cdot \det \begin{bmatrix} \lambda + p & q \\ -1 & \lambda \end{bmatrix} \\ &= r + \lambda \cdot \big((\lambda + p)\lambda + q\big) = r + \lambda^3 + \lambda^2 p + \lambda q = 0 \end{aligned}$$

which matches (7.93).

To obtain the case $\lambda_1 = \lambda_2 = \lambda_3$ as a limiting case of 3 different eigenvalues we will let λ denote the common limiting value: $\lambda_i \to \lambda \ \forall i$. We will also specifically choose the way how λ_i approach λ by specifying

$$\begin{aligned} \lambda_1 &\equiv \lambda \\ \lambda_2 &= \lambda + \Delta\lambda \\ \lambda_3 &= \lambda - \Delta\lambda \end{aligned}$$

Then by letting $\Delta\lambda \to 0$ we obtain $\lambda_i \to \lambda \ \forall i$.

A characteristic equation, which yields the specified above λ_i as its solutions, has the form

$$\begin{aligned} (\tilde{\lambda} - \lambda_1)(\tilde{\lambda} - \lambda_2)(\tilde{\lambda} - \lambda_3) &= (\tilde{\lambda} - \lambda) \cdot (\tilde{\lambda} - (\lambda + \Delta\lambda)) \cdot (\tilde{\lambda} - (\lambda - \Delta\lambda)) \\ &= (\tilde{\lambda} - \lambda) \cdot ((\tilde{\lambda} - \lambda) - \Delta\lambda) \cdot ((\tilde{\lambda} - \lambda) + \Delta\lambda) = (\tilde{\lambda} - \lambda) \cdot ((\tilde{\lambda} - \lambda)^2 - \Delta\lambda^2) \\ &= (\tilde{\lambda} - \lambda)^3 - \Delta\lambda^2 \cdot (\tilde{\lambda} - \lambda) = \tilde{\lambda}^3 - 3\lambda\tilde{\lambda}^2 + 3\lambda^2\tilde{\lambda} - \lambda^3 - \Delta\lambda^2 \cdot \tilde{\lambda} + \lambda\Delta\lambda^2 \\ &= \tilde{\lambda}^3 - 3\lambda\tilde{\lambda}^2 + (3\lambda^2 - \Delta\lambda^2)\tilde{\lambda} + \lambda(\Delta\lambda^2 - \lambda^2) = 0 \end{aligned}$$

where $\tilde{\lambda}$ is the unknown. Therefore the coefficients of (7.93) are

$$p = -3\lambda \qquad q = 3\lambda^2 - \Delta\lambda^2 \qquad r = \lambda(\Delta\lambda^2 - \lambda^2)$$

and the matrix (7.95) is

$$\mathbf{A} = \begin{bmatrix} 3\lambda & \Delta\lambda^2 - 3\lambda^2 & \lambda(\lambda^2 - \Delta\lambda^2) \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{7.96}$$

The eigenvectors $\mathbf{v}_i$ are obtained as solutions of $(\mathbf{A} - \lambda_i)\mathbf{v}_i = 0$. For $\lambda_1 = \lambda$ we have

$$\begin{bmatrix} 2\lambda & \Delta\lambda^2 - 3\lambda^2 & \lambda(\lambda^2 - \Delta\lambda^2) \\ 1 & -\lambda & 0 \\ 0 & 1 & -\lambda \end{bmatrix} \cdot \mathbf{v}_1 = 0 \quad \text{yielding} \quad \mathbf{v}_1 = \begin{bmatrix} \lambda^2 \\ \lambda \\ 1 \end{bmatrix}$$

Notice that since the matrix $\mathbf{A} - \lambda_1$ is singular and since its 2nd and 3rd rows are linearly independent, its 1st row must be a linear combination of the 2nd and the 3rd row. Therefore it suffices to obtain $\mathbf{v}_1$ satisfying the equations generated by the 2nd and the 3rd row while the 1st row can be ignored.

For $\lambda_2 = \lambda + \Delta\lambda$ we obtain

$$\begin{bmatrix} 2\lambda - \Delta\lambda & \Delta\lambda^2 - 3\lambda^2 & \lambda(\lambda^2 - \Delta\lambda^2) \\ 1 & -(\lambda + \Delta\lambda) & 0 \\ 0 & 1 & -(\lambda + \Delta\lambda) \end{bmatrix} \cdot \mathbf{v}_2 = 0 \quad \text{yielding} \quad \mathbf{v}_2 = \begin{bmatrix} (\lambda + \Delta\lambda)^2 \\ \lambda + \Delta\lambda \\ 1 \end{bmatrix}$$

where we again used only the 2nd and the 3rd rows of $\mathbf{A} - \lambda_2$, and for $\lambda_3 = \lambda - \Delta\lambda$ we obtain

$$\begin{bmatrix} 2\lambda + \Delta\lambda & \Delta\lambda^2 - 3\lambda^2 & \lambda(\lambda^2 - \Delta\lambda^2) \\ 1 & -(\lambda - \Delta\lambda) & 0 \\ 0 & 1 & -(\lambda - \Delta\lambda) \end{bmatrix} \cdot \mathbf{v}_3 = 0 \quad \text{yielding} \quad \mathbf{v}_3 = \begin{bmatrix} (\lambda - \Delta\lambda)^2 \\ \lambda - \Delta\lambda \\ 1 \end{bmatrix}$$

At $\Delta\lambda \to 0$ we have $\mathbf{v}_i \to \mathbf{v}$ where $\mathbf{v} = \mathbf{v}_1 = [\lambda^2 \ \lambda \ 1]^{\mathrm{T}}$. We thereby have $\mathbf{v} = \mathbf{v}_1$ as our first (and so far the only) basis vector in case $\Delta\lambda = 0$.

We are then going to take the same approach as in Subsection 7.8.1, taking a limit of the scaled difference of two eigenvectors to obtain the rank 2 generalized eigenvector

$$\mathbf{v}' = \lim_{\Delta\lambda \to 0} \frac{\mathbf{v}_2 - \mathbf{v}}{\Delta\lambda} = -\lim_{\Delta\lambda \to 0} \frac{\mathbf{v}_3 - \mathbf{v}}{\Delta\lambda} = \lim_{\Delta\lambda \to 0} \begin{bmatrix} (\lambda + \Delta\lambda)^2 - \lambda^2 \\ (\lambda + \Delta\lambda) - \lambda \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

It is easily verified that $\mathbf{v}'$ satisfies the usual rank 2 generalized eigenvector equation (7.90). Indeed

$$\begin{aligned} \mathbf{A}\mathbf{v}' &= \mathbf{A} \cdot \lim_{\Delta\lambda \to 0} \frac{\mathbf{v}_2 - \mathbf{v}}{\Delta\lambda} = \lim_{\Delta\lambda \to 0} \frac{\mathbf{A}\mathbf{v}_2 - \mathbf{A}\mathbf{v}}{\Delta\lambda} = \lim_{\Delta\lambda \to 0} \frac{(\lambda + \Delta\lambda)\mathbf{v}_2 - \lambda\mathbf{v}}{\Delta\lambda} \\ &= \lim_{\Delta\lambda \to 0} \frac{\lambda(\mathbf{v}_2 - \mathbf{v}) - \Delta\lambda \cdot \mathbf{v}_2}{\Delta\lambda} = \lambda \cdot \lim_{\Delta\lambda \to 0} \frac{\mathbf{v}_2 - \mathbf{v}}{\Delta\lambda} + \lim_{\Delta\lambda \to 0} \mathbf{v}_2 = \lambda\mathbf{v}' + \mathbf{v} \end{aligned}$$

which is the same equation (7.90) that we had in the 2D case. Also, same as in the 2D case, (7.90) implies that the length of $\mathbf{v}'$ can be arbitrary, that is we can use (7.91) instead of (7.90), and that the eigenvector $\mathbf{v}$ (scaled to an arbitrary length) can be added to $\mathbf{v}'$ without affecting (7.90) or (7.91), also not affecting the linear independence of $\mathbf{v}'$ and $\mathbf{v}$. Thereby all possible vectors $\mathbf{v}'$ form a 2D plane (spanned by the vectors $\mathbf{v}$ and $\mathbf{v}'$) in the 3D space as well, with the exception of the 1D eigenspace going along the vector $\mathbf{v}$, which is subtracted from that plane.

Taking $\mathbf{v}'$ as the second basis vector, we are going to obtain the third basis vector $\mathbf{v}'' = [1 \ 0 \ 0]^{\mathrm{T}}$ as

$$\mathbf{v}'' = \lim_{\Delta\lambda \to 0} \frac{\mathbf{v}_2 - \Delta\lambda \cdot \mathbf{v}' - \mathbf{v}}{\Delta\lambda^2} = \lim_{\Delta\lambda \to 0} \frac{1}{\Delta\lambda^2} \begin{bmatrix} (\lambda + \Delta\lambda)^2 - 2\Delta\lambda - \lambda^2 \\ (\lambda + \Delta\lambda) - \Delta\lambda - \lambda \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Similarly to the eigenvector $\mathbf{v}$ and the rank 2 generalized eigenvector $\mathbf{v}'$, we would be interested in the effect of the matrix $\mathbf{A}$ on the vector $\mathbf{v}''$. In particular

we would like to know if we can factor the multiplier λ out of the expression, similarly to how it appears in the equations (7.56), (7.90a) and (7.91b). We get

$$\begin{aligned}
\mathbf{A}\mathbf{v}'' &= \mathbf{A}\cdot\lim_{\Delta\lambda\to 0}\frac{\mathbf{v}_2-\Delta\lambda\cdot\mathbf{v}'-\mathbf{v}}{\Delta\lambda^2} = \lim_{\Delta\lambda\to 0}\frac{\mathbf{A}\mathbf{v}_2-\Delta\lambda\cdot\mathbf{A}\mathbf{v}'-\mathbf{A}\mathbf{v}}{\Delta\lambda^2}\\
&= \lim_{\Delta\lambda\to 0}\frac{(\lambda+\Delta\lambda)\mathbf{v}_2-\Delta\lambda(\lambda\mathbf{v}'+\mathbf{v})-\lambda\mathbf{v}}{\Delta\lambda^2}\\
&= \lim_{\Delta\lambda\to 0}\frac{\lambda(\mathbf{v}_2-\Delta\lambda\cdot\mathbf{v}'-\mathbf{v})+\Delta\lambda(\mathbf{v}_2-\mathbf{v})}{\Delta\lambda^2}\\
&= \lambda\cdot\lim_{\Delta\lambda\to 0}\frac{\mathbf{v}_2-\Delta\lambda\cdot\mathbf{v}'-\mathbf{v}}{\Delta\lambda^2}+\lim_{\Delta\lambda\to 0}\frac{\mathbf{v}_2-\mathbf{v}}{\Delta\lambda} = \lambda\mathbf{v}''+\mathbf{v}'
\end{aligned}$$

That is

$$\mathbf{A}\mathbf{v}'' = \lambda\mathbf{v}''+\mathbf{v}' \tag{7.97a}$$

or

$$(\mathbf{A}-\lambda)\mathbf{v}'' = \mathbf{v}' \tag{7.97b}$$

That is $\mathbf{v}''$ and $\mathbf{v}'$ are in the same relationship as $\mathbf{v}'$ and $\mathbf{v}$.

Apparently, we can change the length of $\mathbf{v}'$ arbitrarily, which by (7.97a) changes the length of $\mathbf{v}''$ by the same factor. Therefore we can also rewrite (7.97b) as

$$(\mathbf{A}-\lambda)\mathbf{v}'' = \lambda''\mathbf{v}' \tag{7.98a}$$

which turns (7.97a) into

$$\mathbf{A}\mathbf{v}'' = \lambda\mathbf{v}''+\lambda''\mathbf{v}' \tag{7.98b}$$

where $\lambda''\neq 0$.

According to (7.98a), multiplication of $\mathbf{v}''$ by $\mathbf{A}-\lambda$ yields a rank 2 generalized eigenvector, but not an eigenvector as in (7.91a) or a zero as in (7.63). Therefore $\mathbf{v}''$ is different from all possible rank 2 generalized eigenvectors $\mathbf{v}'$ and from all possible eigenvectors $\mathbf{v}$. Therefore any vector $\mathbf{v}''$ satisfying (7.98a) is not lying in the 2D plane spanned by the vectors $\mathbf{v}'$ and $\mathbf{v}$, which automatically implies that the vectors $\mathbf{v}''$, $\mathbf{v}'$ and $\mathbf{v}$ are linearly independent and span the entire 3D space. Vector $\mathbf{v}''$ is referred to as *rank 3 generalized eigenvector*.

Goal 7.9b completed. ■

7.8.3 Rank 2 generalized eigenspace in 3D

In Subsections 7.8.1 and 7.8.2 the generalized eigenspaces (corresponding to the eigenvalue λ) were occupying the entire linear space, due to algebraic multiplicity of the eigenvalue being equal to the dimension of the entire space. It might be worth it to also consider an example where a generalized eigenspace is only a subspace of the entire linear space.

Goal 7.9c. Study one example case (with some choice of the specific limiting values of the eigenvalues) where only 2 of the 3 eigenvectors of a 3×3 matrix coincide in the limit.

To do that let's again consider a 3rd order differential equation, but this time let only two of the 3 eigenvalues coincide. Specifically, let $\lambda_1 = 1$, $\lambda_2 = \Delta\lambda$, $\lambda_3 = -\Delta\lambda$ where $\Delta\lambda\to 0$. The characteristic equation (7.93) becomes

$$(\lambda-1)(\lambda-\Delta\lambda)(\lambda+\Delta\lambda) = (\lambda-1)(\lambda^2-\Delta\lambda^2) = \lambda^3-\lambda^2-\Delta\lambda^2\lambda+\Delta\lambda^2 = 0$$

that is

$$\lambda^3 - \lambda^2 - \Delta\lambda^2\lambda + \Delta\lambda^2 = 0$$

The equation (7.92) is thereby

$$y''' - y'' - \Delta\lambda^2 y' + \Delta\lambda^2 y = 0$$

and the matrix (7.95) is

$$\mathbf{A} = \begin{bmatrix} 1 & \Delta\lambda^2 & -\Delta\lambda^2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

The eigenvectors $\mathbf{v}_i$ are obtained as solutions of $\mathbf{A} - \lambda_i$ where for $\lambda_1 = 1$ we have

$$\begin{bmatrix} 0 & \Delta\lambda^2 & -\Delta\lambda^2 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \cdot \mathbf{v}_1 = 0 \quad \text{yielding} \quad \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

where, as in Subsection 7.8.2 it suffices to solve the equations corresponding to the 2nd and the 3rd rows of matrix $\mathbf{A} - \lambda_1$. For $\lambda_2 = \Delta\lambda$ we have

$$\begin{bmatrix} 1 - \Delta\lambda & \Delta\lambda^2 & -\Delta\lambda^2 \\ 1 & -\Delta\lambda & 0 \\ 0 & 1 & -\Delta\lambda \end{bmatrix} \cdot \mathbf{v}_2 = 0 \quad \text{yielding} \quad \mathbf{v}_2 = \begin{bmatrix} \Delta\lambda^2 \\ \Delta\lambda \\ 1 \end{bmatrix}$$

and for $\lambda_3 = -\Delta\lambda$ we have

$$\begin{bmatrix} 1 + \Delta\lambda & \Delta\lambda^2 & -\Delta\lambda^2 \\ 1 & \Delta\lambda & 0 \\ 0 & 1 & \Delta\lambda \end{bmatrix} \cdot \mathbf{v}_3 = 0 \quad \text{yielding} \quad \mathbf{v}_3 = \begin{bmatrix} \Delta\lambda^2 \\ -\Delta\lambda \\ 1 \end{bmatrix}$$

At $\Delta\lambda \to 0$ we respectively have a 2D plane spanned by the eigenvectors $\mathbf{v}_2$ and $\mathbf{v}_3$ turning into a generalized eigenplane which is orthogonal to the vector $[1\ 0\ 0]^{\mathrm{T}}$ (and corresponds to the eigenvalue $\lambda = 0$ of multiplicity 2). Indeed, the 2D plane spanned by $\mathbf{v}_2$ and $\mathbf{v}_3$ is also spanned by the vectors

$$\frac{\mathbf{v}_1 + \mathbf{v}_2}{2} = \begin{bmatrix} \Delta\lambda^2 \\ 0 \\ 1 \end{bmatrix} \to \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{7.99a}$$

and

$$\frac{\mathbf{v}_1 - \mathbf{v}_2}{2\Delta\lambda} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{7.99b}$$

Now, the eigenvector obtained as the limit of (7.99a) satisfies the equation $(\mathbf{A} - \lambda)\mathbf{v} = 0$ (where $\lambda = 0$) while the generalized eigenvector obtained from (7.99b) satisfies the equation $(\mathbf{A} - \lambda)\mathbf{v}' = \mathbf{v}$. Multiplying the latter equation by $\mathbf{A} - \lambda$ from the left, we obtain that the generalized eigenvector must satisfy

$$(\mathbf{A} - \lambda) \cdot (\mathbf{A} - \lambda)\mathbf{v}' = (\mathbf{A} - \lambda)\mathbf{v} = 0$$

that is

$$(\mathbf{A}-\lambda)^2\mathbf{v}' = 0 \tag{7.100}$$

There are however too many vectors satisfying (7.100) since it is also satisfied by ordinary eigenvectors:

$$(\mathbf{A}-\lambda)^2\mathbf{v} = (\mathbf{A}-\lambda)\cdot(\mathbf{A}-\lambda)\mathbf{v} = (\mathbf{A}-\lambda)\cdot 0 = 0$$

But this means that (7.100) is actually the equation for the entire generalized eigenplane, containing both the eigenvectors and the rank 2 generalized eigenvectors corresponding to the eigenvalue $\lambda = 0$.

Goal 7.9c completed. ■

7.8.4 Jordan block

Goal 7.9d. Derive the idea of Jordan block from the previous results. *Hint:* switch to the basis consisting of generalized eigenvectors.

Previously discussed examples of generalized eigenvectors can be generalized to an arbitrary rank n as:

$$\mathbf{A}\mathbf{v}^{(n)} = \lambda\mathbf{v}^{(n)} + \lambda^{(n)}\mathbf{v}^{(n-1)} \tag{7.101a}$$

$$(\mathbf{A}-\lambda)\mathbf{v}^{(n)} = \lambda^{(n)}\mathbf{v}^{(n-1)} \tag{7.101b}$$

where $\mathbf{v}^{(0)} = 0$, $\mathbf{v}^{(1)} = \mathbf{v}$, $\mathbf{v}^{(2)} = \mathbf{v}'$, $\mathbf{v}^{(3)} = \mathbf{v}''$, etc., and $\lambda^{(2)} = \lambda' \neq 0$, $\lambda^{(3)} = \lambda'' \neq 0$, etc., while the value of $\lambda^{(0)}$ doesn't matter (notice that n is larger by 1 than the number of primes in the notation). Also notice that thereby the eigenvector $\mathbf{v}$ can be referred to as *rank 1 generalized eigenvector*.

> ▷ While ordinary eigenvectors can be referred to as rank 1 generalized eigenvectors, and thereby ordinary eigenvectors are one particular case of generalized eigenvectors, sometimes by saying "generalized eigenvectors" we will mean "generalized eigenvectors of rank 2 or higher", thereby excluding ordinary eigenvectors. Whether ordinary eigenvectors are included or not in such cases will need to be inferred from the context. E.g. when we will be talking of whether a particular matrix has generalized eigenvectors or not, we will obviously mean generalized eigenvectors of rank 2 or higher, since any matrix has at least one generalized eigenvector of rank 1.

On the other hand, the idea of (7.100) can be generalized into a slightly different approach to define an n-th rank generalized eigenvector by

$$\begin{cases} (\mathbf{A}-\lambda)^n\mathbf{v}^{(n)} = 0 \\ (\mathbf{A}-\lambda)^{n-1}\mathbf{v}^{(n)} \neq 0 \end{cases} \tag{7.102}$$

where the first of the equations makes sure that $\mathbf{v}^{(n)}$ is a generalized eigenvector of a rank not exceeding n, while the second equation makes sure that the rank of $\mathbf{v}^{(n)}$ cannot be less than n. The main difference of (7.102) from (7.101) is that (7.102) defines *some* generalized eigenvector of rank n corresponding to the eigenvalue λ, while (7.101) defines a rank n generalized eigenvector, which is specifically attached to the given $(n-1)$-th rank generalized eigenvector.

At $n = \mu$ (where μ is the algebraic multiplicity of λ) the first equation of (7.102) turns into the generalized eigenspace equation:

$$(\mathbf{A} - \lambda)^{\mu}\mathbf{u} = 0 \tag{7.103}$$

Vector $\mathbf{u}$ belongs to the generalized eigenspace corresponding to the eigenvalue λ of algebraic multiplicity μ iff it satisfies (7.103).

Switching to a *generalized eigenbasis* consisting of the generalized eigenvectors $\mathbf{v}^{(n)}$ satisfying (7.101), we find the matrix $\mathbf{A}$ taking the form (as per (7.101a))

$$\mathbf{A} = \begin{bmatrix} \lambda & \lambda^{(2)} & 0 & \cdots & 0 & 0 \\ 0 & \lambda & \lambda^{(3)} & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & \lambda^{(n)} \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

For the sake of definiteness we can let $\lambda^{(n)} = 1$, obtaining the matrix $\mathbf{A}$ in the form of a *Jordan block*

(Jordan block)

$$\mathbf{A} = \mathbf{J} = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} = \lambda + \mathbf{N} \tag{7.104}$$

where

$$\mathbf{N} = \mathbf{A} - \lambda = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{7.105}$$

Notice that by letting $\lambda^{(n)} = 1$ the equations (7.101a) and (7.101b) turn into

(Jordan chain equations)

$$\mathbf{A}\mathbf{v}^{(n)} = \lambda\mathbf{v}^{(n)} + \mathbf{v}^{(n-1)} \tag{7.106a}$$

$$(\mathbf{A} - \lambda)\mathbf{v}^{(n)} = \mathbf{v}^{(n-1)} \tag{7.106b}$$

Since (in the new basis) $\mathbf{A} - \lambda = \mathbf{N}$, the equation (7.106b) can be rewritten as

$$\mathbf{N}\mathbf{v}^{(n)} = \mathbf{v}^{(n-1)}$$

while (7.102) becomes

$$\begin{cases} \mathbf{N}^n \mathbf{v}^{(n)} = 0 \\ \mathbf{N}^{n-1} \mathbf{v}^{(n)} \neq 0 \end{cases}$$

which also implies $\mathbf{N}^N \mathbf{u} = 0 \ \forall \mathbf{u}$ and therefore $\mathbf{N}^N = 0$.

The equations (7.106) connect eigenvectors of consecutive ranks into some kind of chain, referred to as *Jordan chain*.

Goal 7.9d completed. ■

7.8.5 Generalized eigenspaces

We have observed how in a number of particular cases, as all or some of the eigenvalues approach the same limit, a generalized eigenplane or generalized eigenspace emerges as a limit of the space spanned by the respective eigenvectors. We however cannot be sure yet, whether this is the only possible outcome or some other situations are also possible.

Goal 7.9e. Study which possible situations can emerge in case one or more of the eigenvalues of a matrix are repeated. Do we always obtain enough eigenvectors and/or generalized eigenvectors to make a basis? *Hint:* recall from Chapter 5 that a determinant can describe a volume or a volume deformation.

Let

$$\det(\lambda - \mathbf{A}) = \prod_{i=1}^{M} (\lambda - \lambda_i)^{\mu_i} \tag{7.107}$$

be the characteristic polynomial of $\mathbf{A}$, where $\lambda_i \neq \lambda_j \ \forall i \neq j$, thus λ_i are the eigenvalues of the respective algebraic multiplicities μ_i, their total number being M. First we should notice that the respective eigenvectors (1st rank generalized eigenvectors) $\mathbf{v}_i$ are all linearly independent.

▷ Indeed, it is easy to realize that if $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for some particular i, then $\mathbf{A}\mathbf{v}_i$ cannot simultaneously be equal to $\lambda_j \mathbf{v}_i$ for $j \neq i$. So, two eigenvectors corresponding to two different eigenvalues must have different directions. This, however, doesn't mean linear independence of a larger set of different eigenvectors yet. The latter can be argued by assuming $\mathbf{v}_i = \sum_{j \neq i} \alpha_j \mathbf{v}_j$, where $\{\mathbf{v}_j\}$ is any linearly independent subset of eigenvectors of $\mathbf{A}$, and considering

$$\mathbf{A}\mathbf{v}_i = \mathbf{A} \sum_{j \neq i} \alpha_j \mathbf{v}_j = \sum_{j \neq i} \alpha_j \mathbf{A}\mathbf{v}_j = \sum_{j \neq i} \alpha_j \lambda_j \mathbf{v}_j \tag{7.108a}$$

$$\lambda_i \mathbf{v}_i = \lambda_i \sum_{j \neq i} \alpha_j \mathbf{v}_j = \sum_{j \neq i} \alpha_j \lambda_i \mathbf{v}_j \tag{7.108b}$$

Now, if $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, that is left-hand sides of (7.108) are equal, then the right-hand sides are equal, which would mean that there are two different linear expansions of one and the same vector in terms of linearly independent vectors $\mathbf{v}_j$, which is impossible. This shows linear independence of any set of eigenvectors corresponding to different eigenvalues of a matrix.

However, if the total number of different eigenvalues M is smaller than the matrix dimension N, so far we can be certain only to get M different linearly independent eigenvectors $\mathbf{v}_i$, which thereby do not constitute a basis of an N-dimensional space. One possibility to fill up the remaining dimensions would be

by the respective generalized eigenvectors, but we cannot be sure if there always exists "a sufficient number of those". Let's find out.

For the sake of definiteness, let's concentrate on λ_1 and $\mathbf{v}_1$, where we generally assume $\mu_1 > 1$, since the case $\mu_1 = 1$ is trivial. We'd like to make some intuitive observations corresponding to the fact that λ_1 is a root of multiplicity μ_1 of the characteristic polynomial (7.107) of $\mathbf{A}$. We may recall from Chapter 5, that the determinant in the left-hand side of (7.107) describes the (hyper)volume deformation caused by the matrix $\lambda - \mathbf{A}$. The value λ_1 being a root of multiplicity μ_1 of (7.107) therefore means that the volume decays to zero as $(\lambda - \lambda_1)^{\mu_1}$ at $\lambda \to \lambda_1$. Let's think, how such volume deformation can emerge in terms of vector transformations caused by multiplication by $\lambda - \mathbf{A}$.

Let $\mathbf{v}_{11}, \mathbf{v}_{12}, \ldots, \mathbf{v}_{1\mu_1}$ be a set of *linearly independent* vectors spanning a hyperparallelepiped in a μ_1-dimensional subspace. We are interested in the image of this hyperparallelepiped under the transformation by $\lambda - \mathbf{A}$. The (hyper)volume of this image is apparently a function of λ:

$$V(\lambda) = \left|(\lambda - \mathbf{A})\mathbf{v}_{11} \wedge (\lambda - \mathbf{A})\mathbf{v}_{12} \wedge \ldots \wedge (\lambda - \mathbf{A})\mathbf{v}_{1\mu_1}\right| \tag{7.109}$$

We would like to know, under which circumstances $V(\lambda)$ decays as $(\lambda - \lambda_1)^{\mu_1}$ at $\lambda \to \lambda_1$.

One obvious possibility is if each of the factors $(\lambda - \mathbf{A})\mathbf{v}_{1i}$ decays as $(\lambda - \lambda_1)$. Thereby $(\lambda_1 - \mathbf{A})\mathbf{v}_{1i} = 0\ \forall i$, which means that all of $\mathbf{v}_{1i}$ are eigenvectors of $\mathbf{A}$ associated with the eigenvalue λ_1. Thus, we are having μ_1 linearly independent eigenvectors corresponding to λ_1, therefore there is a μ_1-dimensional eigenspace spanned by these vectors. Thus we have "used up" all μ_1 dimensions which we would like to associate with λ_1, thereby obtaining the necessary number vectors for the eigenbasis. Notice that in this case any vector inside the subspace spanned by $\mathbf{v}_{1i}$ is an eigenvector corresponding to λ_1.

We already have seen that this is not the only possibility, as sometimes there are not enough eigenvectors. Let's see how else we can have $V(\lambda)$ decay as $(\lambda - \lambda_1)^{\mu_1}$. It might be helpful to rewrite the factors in (7.109) in terms of $\lambda - \lambda_1$ rather than in terms of λ, to make the behavior at $\lambda \to \lambda_1$ more obvious:

$$(\lambda - \mathbf{A})\mathbf{v}_{1i} = \big((\lambda_1 - \mathbf{A}) + (\lambda - \lambda_1)\big)\mathbf{v}_{1i} = (\lambda_1 - \mathbf{A})\mathbf{v}_{1i} + (\lambda - \lambda_1)\mathbf{v}_{1i} \tag{7.110}$$

Also to make the subsequent analysis simpler, let's temporarily assume $\mu_1 = 2$ and then generalize to higher values of μ_1.

So suppose $\mu_1 = 2$. We already know that there must be at least one eigenvector of $\mathbf{A}$ corresponding to λ_1 and that one will already contribute a factor of $\lambda - \lambda_1$ to the volume decay, therefore it should "belong to the volume $V(\lambda)$". So let $\mathbf{v}_{11} = \mathbf{v}_1$ be this eigenvector. We obtain

$$\begin{aligned} V(\lambda) &= \left|(\lambda - \mathbf{A})\mathbf{v}_{11} \wedge (\lambda - \mathbf{A})\mathbf{v}_{12}\right| = \left|(\lambda - \mathbf{A})\mathbf{v}_1 \wedge (\lambda - \mathbf{A})\mathbf{v}_{12}\right| \\ &= \left|(\lambda - \lambda_1)\mathbf{v}_1 \wedge (\lambda - \mathbf{A})\mathbf{v}_{12}\right| \\ &\overset{(7.110)}{=} \left|(\lambda - \lambda_1)\mathbf{v}_1 \wedge \big((\lambda_1 - \mathbf{A})\mathbf{v}_{12} + (\lambda - \lambda_1)\mathbf{v}_{12}\big)\right| \end{aligned} \tag{7.111a}$$

As mentioned, the first factor $(\lambda - \lambda_1)\mathbf{v}_1$ decays as $\lambda - \lambda_1$. We therefore need the second factor to contribute another decay factor of $\lambda - \lambda_1$, so that $V(\lambda)$ decays as $(\lambda - \lambda_1)^2$.

The term $(\lambda_1 - \mathbf{A})\mathbf{v}_{12}$ in the second factor of (7.111a) is constant, so the only way the second factor of (7.111a) can contribute a $\lambda - \lambda_1$ decay factor is if $(\lambda_1 - \mathbf{A})\mathbf{v}_{12}$ makes a zero contribution to $V(\lambda)$, that is

$$\mathbf{v}_1 \wedge (\lambda_1 - \mathbf{A})\mathbf{v}_{12} = 0 \tag{7.111b}$$

That is $(\lambda_1 - \mathbf{A})\mathbf{v}_{12}$ must be collinear to $\mathbf{v}_1$ (where $(\lambda_1 - \mathbf{A})\mathbf{v}_{12} = 0$ counts as being collinear to any vector). It follows that $(\lambda_1 - \mathbf{A})^2 \cdot \mathbf{v}_{12} = 0$, therefore $\mathbf{v}_{12}$ is either an eigenvector or a rank 2 generalized eigenvector corresponding to the eigenvalue λ_1.

So, to sum up the case $\mu_1 = 2$, we established that there is always a choice of $\mathbf{v}_{11}$, $\mathbf{v}_{12}$ spanning the volume $V(\lambda)$, such that

$$(\lambda_1 - \mathbf{A}) \cdot \mathbf{v}_{11} = 0 \tag{7.112a}$$

$$(\lambda_1 - \mathbf{A})^2 \cdot \mathbf{v}_{12} = 0 \tag{7.112b}$$

which implies that $\mathbf{v}_{11}$ is an eigenvector and $\mathbf{v}_{12}$ is an eigenvector or a rank 2 generalized eigenvector, respectively amounting to 2 Jordan chains of length 1, or a single Jordan chain of length 2.

Now suppose $\mu_1 = 3$. The situation with $\mathbf{v}_{11}$ and $\mathbf{v}_{12}$ can be handled similarly to the case $\mu_1 = 2$, so that both of these vectors together contribute a factor of $(\lambda - \lambda_1)^2$ to $V(\lambda)$. Respectively, (7.112) hold and we have

$$V(\lambda) = \Big|(\lambda - \lambda_1)\mathbf{v}_{11} \wedge (\lambda - \lambda_1)\mathbf{v}_{12} \wedge \big((\lambda_1 - \mathbf{A})\mathbf{v}_{13} + (\lambda - \lambda_1)\mathbf{v}_{13}\big)\Big|$$

where the term $(\lambda_1 - \mathbf{A})\mathbf{v}_{12}$ in the second factor was either zero from the very beginning or got cancelled by the first factor. By the argument similar to $\mu_1 = 2$, we need $(\lambda_1 - \mathbf{A})\mathbf{v}_{13}$ to make zero contribution to $V(\lambda)$, which is possible only if $(\lambda_1 - \mathbf{A})\mathbf{v}_{13}$ is lying in the plane spanned by $\mathbf{v}_{11}$ and $\mathbf{v}_{12}$ (as with collinearity, this includes the case $(\lambda_1 - \mathbf{A})\mathbf{v}_{13} = 0$). We thus obtain $(\lambda_1 - \mathbf{A})^3 \cdot \mathbf{v}_{13} = 0$, which together with (7.112) gives

$$(\lambda_1 - \mathbf{A}) \cdot \mathbf{v}_{11} = 0 \tag{7.113a}$$

$$(\lambda_1 - \mathbf{A})^2 \cdot \mathbf{v}_{12} = 0 \tag{7.113b}$$

$$(\lambda_1 - \mathbf{A})^3 \cdot \mathbf{v}_{13} = 0 \tag{7.113c}$$

As with (7.112), the meaning of (7.113) is that the image of a volume spanning the 3 linearly independent vectors $\mathbf{v}_{11}$, $\mathbf{v}_{12}$, $\mathbf{v}_{13}$ decays as $(\lambda - \lambda_1)^3$ at $\lambda \to \lambda_1$ iff $\mathbf{v}_{11}$, $\mathbf{v}_{12}$, $\mathbf{v}_{13}$ satisfy (7.113). Thus, given $V(\lambda)$ decaying as $(\lambda - \lambda_1)^3$ at $\lambda \to \lambda_1$, there also must be a respective choice of $\mathbf{v}_{11}$, $\mathbf{v}_{12}$, $\mathbf{v}_{13}$.

There were two different ways, how (7.112) could have been achieved. The vector $\mathbf{v}_{11}$ was fixed in its eigenvector role, while $\mathbf{v}_{12}$ could have been another eigenvector or a rank 2 generalized eigenvector. Since (7.112) is a subset of (7.113), the same holds for (7.113), while $\mathbf{v}_{13}$ could be yet another eigenvector, or a generalized eigenvector of rank 2 or 3. Apparently, there are different ways how $\mathbf{v}_{11}$, $\mathbf{v}_{12}$ and $\mathbf{v}_{13}$ can be organized into Jordan chains, with the extreme cases being a single Jordan chain of length 3 or three Jordan chains of length 1.

▷ It must be noted, that the equations (7.113) ensure that there are choices of $\mathbf{v}_{1i}$ such that $\mathbf{v}_{1i}$ span the entire generalized eigenspace associated with λ_1, but they actually do not ensure that the vectors $\mathbf{v}_{1i}$ will automatically connect into Jordan chains in the sense of (7.101) or (7.106). E.g. we could have a situation where $\mathbf{v}_{11}$ and $\mathbf{v}_{12}$ are two different eigenvectors, while $(\lambda_1 - \mathbf{A})\mathbf{v}_{13} = \mathbf{v}_{11} + \mathbf{v}_{12}$. That is, $\mathbf{v}_{13}$ is a rank 2 generalized eigenvector, but it is associated with the sum of $\mathbf{v}_{11}$ and $\mathbf{v}_{12}$, rather than with one of these vectors. In this case we'll need to e.g. replace $\mathbf{v}_{12}$ by $\mathbf{v}_{11} + \mathbf{v}_{12}$ in its role as an eigenvector, which will allow $\mathbf{v}_{13}$ to "connect" to this eigenvector. The readers are welcome to figure out further details by themselves.

Higher values of μ_1 can be addressed by continuing in a similar manner. It follows that for any matrix $\mathbf{A}$ there is always a generalized eigenbasis consisting of its generalized eigenvectors organized into Jordan chains. Respectively any matrix $\mathbf{A}$ can be converted to the Jordan normal form by a similarity transformation.

Goal 7.9e completed. ■

7.8.6 Exponentiation of Jordan blocks

In case of a diagonal matrix $\mathbf{A}$ the exponential $e^{\mathbf{A}t}$ is a diagonal matrix consisting of the exponentials $e^{\lambda_n t}$ on the main diagonal. Respectively, the components of the solution of the homogeneous differential equation $\mathbf{y}' = \mathbf{A}\mathbf{y}$ (where $\mathbf{A}$ is not necessarily in diagonal form, but is diagonalizable) are linear combinations of the exponentials $e^{\lambda_n t}$.

If the matrix $\mathbf{A}$ is not diagonalizable, we can switch to the basis of its generalized eigenvectors instead, thereby bringing $\mathbf{A}$ to the Jordan normal form. If we attempt to compute $e^{\mathbf{A}t}$, apparently different Jordan blocks are not going to interfere with each other and will be exponentiated separately. It suffices therefore to find out the components of $e^{\mathbf{J}t}$ where $\mathbf{J}$ is a general Jordan block matrix.

Goal 7.9f. Construct a way to compute the matrix components of $e^{\mathbf{J}t}$.

Let λ be the eigenvalue corresponding to $\mathbf{J}$ and let $\mathbf{u}(t)$ be the solution of the equation $\mathbf{u}' = \mathbf{J}\mathbf{u}$, so that $\mathbf{u} = e^{\mathbf{J}t} \cdot \mathbf{u}_0$. Using (7.104) we can rewrite the equation $\mathbf{u}' = \mathbf{J}\mathbf{u}$ component by component as

$$\begin{aligned} u_1' &= \lambda u_1 + u_2 \\ u_2' &= \lambda u_2 + u_3 \\ &\vdots \\ u_{N-1}' &= \lambda u_{N-1} + u_N \\ u_N' &= \lambda u_N \end{aligned}$$

Therefore

$$\begin{aligned} u_N(t) &= u_N(0) \cdot e^{\lambda t} \\ u_{N-1}(t) &= u_{N-1}(0) \cdot e^{\lambda t} + \int_0^t e^{\lambda(t-\tau)} u_N(\tau)\,\mathrm{d}\tau \\ &= u_{N-1}(0) \cdot e^{\lambda t} + u_N(0) \int_0^t e^{\lambda(t-\tau)} e^{\lambda\tau}\,\mathrm{d}\tau \end{aligned}$$

$$
\begin{aligned}
&= u_{N-1}(0)\cdot e^{\lambda t} + u_N(0)\int_0^t e^{\lambda t}\,\mathrm{d}\tau \\
&= u_{N-1}(0)\cdot e^{\lambda t} + u_N(0)\cdot te^{\lambda t}
\end{aligned}
$$

(where we have used (7.45) to solve the equations)

$$
\begin{aligned}
u_{N-2}(t) &= u_{N-2}(0)\cdot e^{\lambda t} + \int_0^t e^{\lambda(t-\tau)}u_{N-1}(\tau)\,\mathrm{d}\tau \\
&= u_{N-2}(0)\cdot e^{\lambda t} + u_{N-1}(0)\cdot te^{\lambda t} + u_{N-2}(0)\cdot\frac{t^2}{2}e^{\lambda t}
\end{aligned}
$$

etc.

This means that

$$
e^{\mathbf{J}t} = \begin{bmatrix} e^{\lambda t} & \cdots & \frac{t^{N-3}}{(N-3)!}e^{\lambda t} & \frac{t^{N-2}}{(N-2)!}e^{\lambda t} & \frac{t^{N-1}}{(N-1)!}e^{\lambda t} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2}e^{\lambda t} \\ 0 & \cdots & 0 & e^{\lambda t} & te^{\lambda t} \\ 0 & \cdots & 0 & 0 & e^{\lambda t} \end{bmatrix} \tag{7.114}
$$

Formula (7.114) can be also obtained directly. First we use the distributive law:

$$
e^{\mathbf{J}t} = e^{(\lambda\mathbf{I}+\mathbf{N})t} = e^{\lambda t\mathbf{I}}\cdot e^{\mathbf{N}t} \tag{7.115}
$$

▷ The matrix exponential distributive law $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}\cdot e^{\mathbf{B}}$ applies if the matrices $\mathbf{A}$ and $\mathbf{B}$ commute, that is $\mathbf{AB} = \mathbf{BA}$:

$$
\begin{aligned}
e^{\mathbf{A}+\mathbf{B}} &= \lim_{N\to\infty}\prod_{n=1}^{N}\left(1+\frac{\mathbf{A}+\mathbf{B}}{n}\right)^n = \lim_{N\to\infty}\prod_{n=1}^{N}\left(\left(1+\frac{\mathbf{A}}{n}\right)\left(1+\frac{\mathbf{B}}{n}\right)-O(1/n^2)\right)^n \\
&= \lim_{N\to\infty}\prod_{n=1}^{N}\left(\left(1+\frac{\mathbf{A}}{n}\right)\left(1+\frac{\mathbf{B}}{n}\right)\right)^n = \lim_{N\to\infty}\prod_{n=1}^{N}\left(1+\frac{\mathbf{A}}{n}\right)^n\left(1+\frac{\mathbf{B}}{n}\right)^n \\
&= \lim_{N\to\infty}\prod_{n=1}^{N}\left(1+\frac{\mathbf{A}}{n}\right)^n\cdot\lim_{N\to\infty}\prod_{n=1}^{N}\left(1+\frac{\mathbf{B}}{n}\right)^n = e^{\mathbf{A}}\cdot e^{\mathbf{B}}
\end{aligned}
$$

where the commutativity was applied in replacing $((1+\mathbf{A}/n)(1+\mathbf{B}/n))^n$ by the product $(1+\mathbf{A}/n)^n(1+\mathbf{B}/n)^n$. It follows that if $\mathbf{A}$ and $\mathbf{B}$ commute, then so do $e^{\mathbf{A}}$ and $e^{\mathbf{B}}$. Indeed: $e^{\mathbf{A}}\cdot e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{B}+\mathbf{A}} = e^{\mathbf{B}}\cdot e^{\mathbf{A}}$.

If the matrices do not commute then the formula $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}\cdot e^{\mathbf{B}}$ attains a more complicated form. This is one of the central topics in *Lie theory*.

The identity matrix $\mathbf{I}$ commutes with any matrix, therefore $\mathbf{IN} = \mathbf{NI}$ and $\lambda t\mathbf{I}\cdot\mathbf{N}t = \mathbf{N}t\cdot\lambda t\mathbf{I}$, justifying the transformation (7.115).

In the right-hand side of (7.115) we already know that the first factor is a diagonal matrix with all diagonal components equal to $e^{\lambda t}$. The second factor gives a finite sum

$$
e^{\mathbf{N}t} = \mathbf{I} + \mathbf{N}t + \mathbf{N}^2\frac{t^2}{2} + \ldots + \mathbf{N}^{N-1}\frac{t^{N-1}}{(N-1)!} \tag{7.116a}
$$

since $\mathbf{N}^n = 0\ \forall n > N$. Computing the powers $\mathbf{N}^n$ explicitly:

$$\mathbf{N}^2 = \begin{bmatrix} 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \quad \text{etc.}$$

yields

$$e^{\mathbf{N}t} = \begin{bmatrix} 1 & \cdots & \frac{t^{N-3}}{(N-3)!} & \frac{t^{N-2}}{(N-2)!} & \frac{t^{N-1}}{(N-1)!} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & t & \frac{t^2}{2} \\ 0 & \cdots & 0 & 1 & t \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \tag{7.116b}$$

and (7.114) follows.

Goal 7.9f completed. ■

It follows from (7.114) that

(determinant of a matrix exponential is the exponential of the matrix trace)

$$\det(e^{\mathbf{A}}) = e^{\operatorname{tr}\mathbf{A}} \tag{7.117}$$

Indeed, by (7.114), equation (7.117) holds trivially in a basis where $\mathbf{A}$ is in Jordan normal form. Now all operations in (7.117) (matrix exponential, determinant and trace) are invariant under similarity transformations by (7.58), (5.66) and (4.38), therefore (7.117) also holds in any other basis.

Goal 7.9 completed. ■

7.9 Spectral theorems

In this section we are going to find out some further details about eigenvalues and eigenvectors for certain particular matrix classes, summarized as *spectral theorems*. These details will allow us to quickly predict (in some cases) whether a matrix is diagonalizable, and what kind of eigenvectors and eigenvalues it has.

▷ The set of the eigenvalues of a matrix is referred to as the *spectrum of the matrix*, hence the name "spectral theorems". The reason that the set of the eigenvalues is referred to as spectrum has possibly to do with the fact that in case the matrix describes a differential equation system of the form (7.46), the imaginary parts of the eigenvalues determine the frequencies of the sinusoidal components of the solution.

In this section, for the sake of simplicity of discussion, we are going to assume an *orthonormal* or a *unitary* (that is, "complex orthonormal") basis. Most prominently, we are going to write dot products of vectors as $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{x}^*\mathbf{y}$. Dropping this assumption is not going to fundamentally change the results of the discussion and will be done at the end of this section, in Subsection 7.9.7.

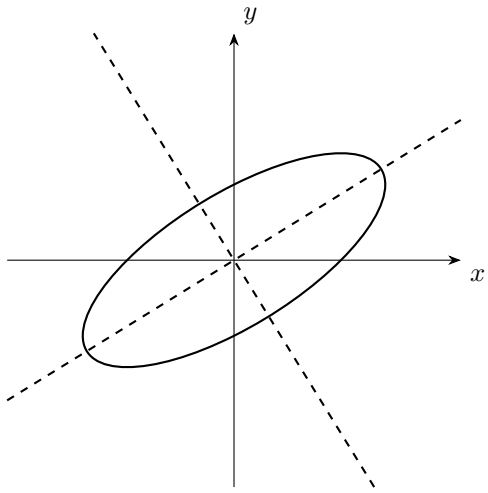


Figure 7.3: The ellipse $x^2 - 2xy + 2y^2 = 2$ and its principal axes.

7.9.1 Diagonalization of quadratic forms

We begin the discovery of spectral theorems by making some observations about quadratic forms.

Given a matrix $\mathbf{Q}$, consider the quadratic form

$$Q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} \tag{7.118}$$

The level sets of (7.118) are given by the equation

$$Q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = C \tag{7.119}$$

where C is some constant. Graphically, the solutions of (7.119) are represented by (N-dimensional) ellipsoids or hyperboloids. E.g. in 2 dimensions the graph of the implicit equation

$$Q(\mathbf{x}) = Q\left(\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]\right) = Q(x, y) = q_{11}x^2 + (q_{12} + q_{21})xy + q_{22}y^2 = C \tag{7.120}$$

may have the form of an ellipse (Fig. 7.3) or a hyperbola (Fig. 7.4). In higher dimension counts we have the respective ellipsoid and hyperboloid surfaces or hypersurfaces.

Both the ellipse and the hyperbola have two principal axes each (shown in Figs. 7.3 and 7.4 by dashed lines) which are orthogonal to each other. Let's try to find out why this is the case.

Goal 7.10. Demonstrate that an N-dimensional quadratic hypersurface defined by (7.119) has N mutually orthogonal principal axes.

Goal 7.10a. Construct a formal definition of a principal axis of a quadratic hypersurface, that is simultaneously intuitive. *Hint:* explore the symmetries of the hypersurface.

In Figs. 7.3 we can notice that the ellipse exhibits a mirror-type symmetry with respect to each of the principal axis and there are no other mirror-type

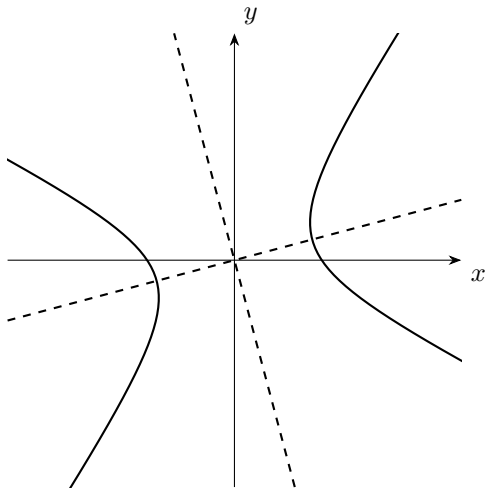


Figure 7.4: The hyperbola $3x^2 + 4xy - 4y^2 = 4$ and its principal axes.

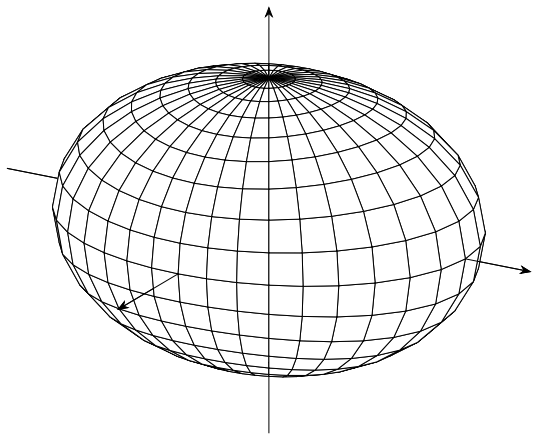


Figure 7.5: The 3D ellipsoid $25x^2 + 16y^2 + 36z^2 = 36$.

symmetries. Our first idea might be therefore to define principal axis as an axis of a mirror symmetry, but we should immediately realize that this can work only in 2D: in a larger dimension count a mirror symmetry needs a reference (hyper)plane rather than a line. We could conjecture that maybe the 2D mirror symmetry with respect to an axis in higher dimension count generalizes to a rotational symmetry with respect to a line (the reference line serving as a rotation axis), but we can immediately realize that already in 3D the ellipsoids generally do not have such rotation symmetry (Fig. 7.5).

The ellipsoids however still have mirror-type symmetries, where mirror planes are orthogonal to principal axes. We could therefore define a principal axis as a line going through the center of an ellipsoid in the direction orthogonal to a mirror plane (or hyperplane in more than 3D).

It is left as an exercise for the readers to extend the above reasoning to hyperboloid hypersurfaces, arriving at the same definition of a principal axis,

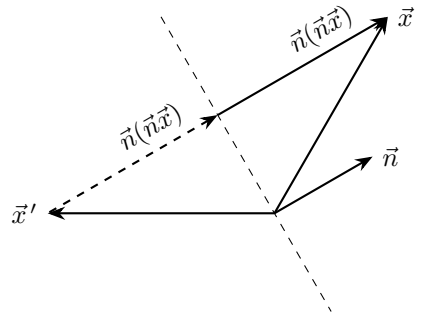


Figure 7.6: Mirror reflection of a vector by (7.121).

which is thereby universally applicable to all quadratic hypersurfaces.

Goal 7.10a completed. ■

Goal 7.10b. Given a quadratic form's matrix $\mathbf{Q}$, find the principal axes of the respective hypersurface. *Hint:* use the hypersurface's symmetry defined by a principal axis.

Since a principal axis is normal to a mirror hyperplane, the respective mirror reflection expression is

$$\mathbf{x}' = \mathbf{x} - 2\mathbf{n} \cdot (\mathbf{n}^{\mathrm{T}}\mathbf{x}) = \mathbf{x} - 2\mathbf{N}\mathbf{x} \tag{7.121}$$

where $|\mathbf{n}| = 1$, that is $\mathbf{n}$ is a normalized principal axis vector. Respectively, $\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})$, alternatively written as $\mathbf{N}\mathbf{x}$ (where $\mathbf{N} = \mathbf{n}\mathbf{n}^{\mathrm{T}}$ is the projection matrix), is projection of $\mathbf{x}$ on $\mathbf{n}$ (Fig. 7.6).

The mirror symmetry of the graph of the equation (7.119) means that the equation (7.119) must be invariant under (7.121):

$$Q(\mathbf{x}) = C \iff Q(\mathbf{x}') = C$$

that is

$$Q(\mathbf{x}) = C \iff Q\big(\mathbf{x} - 2\mathbf{n} \cdot (\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) = C$$

that is

$$\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = C \iff \big(\mathbf{x} - 2\mathbf{n} \cdot (\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(\mathbf{x} - 2\mathbf{n} \cdot (\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) = C$$

that is

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = C \iff \Big(&\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) \\ &+ \big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)\Big) = C\end{aligned}$$

that is

$$-\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) + \big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) = 0$$

Since the first term is a scalar, it's unchanged by transposition, so we can further transform the left-hand side as

$$-\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) + \big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)$$

$$
\begin{aligned}
&= -\Big(\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\mathbf{x}\Big)^{\mathrm{T}} - \mathbf{x}^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) + \big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) \\
&= -\mathbf{x}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) - \mathbf{x}^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) + \big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) \\
&= -\mathbf{x}^{\mathrm{T}}(\mathbf{Q}^{\mathrm{T}} + \mathbf{Q})\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) + \big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(2\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) \\
&= -4\mathbf{x}^{\mathrm{T}}\frac{\mathbf{Q}^{\mathrm{T}} + \mathbf{Q}}{2}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) + 4\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) = 0 \qquad (7.122)
\end{aligned}
$$

Now we can notice that we can also replace $\mathbf{Q}$ in the second term with $\frac{\mathbf{Q}^{\mathrm{T}}+\mathbf{Q}}{2}$ without changing the term's value. Generally, we can do such replacement in any quadratic form $Q(\mathbf{x})$, without changing quadratic the form's value:

$$
\begin{aligned}
\mathbf{x}^{\mathrm{T}}\frac{\mathbf{Q}^{\mathrm{T}} + \mathbf{Q}}{2}\mathbf{x} &= \frac{1}{2}\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}\right) = \frac{1}{2}\left((\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x})^{\mathrm{T}} + \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}\right) \\
&= \frac{1}{2}\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}\right) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = Q(\mathbf{x})
\end{aligned}
$$

Thus

(quadratic form is unchanged by symmetrization of its matrix)

$$
Q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\frac{\mathbf{Q}^{\mathrm{T}} + \mathbf{Q}}{2}\mathbf{x} \qquad (7.123)
$$

▷ Effectively, in the sum $Q(x) = \sum q_{ij}x_ix_j$, for each index pair $i \neq j$ we have a sum of two terms $q_{ij}x_ix_j + q_{ji}x_jx_i = (q_{ij} + q_{ji})x_ix_j$. Replacing $\mathbf{Q}$ with $\frac{\mathbf{Q}^{\mathrm{T}}+\mathbf{Q}}{2}$ simply redistributes the sum $q_{ij} + q_{ji}$ equally between both terms q_{ij} and q_{ji}, without changing the value $q_{ij} + q_{ji}$.

Continuing the transformations of (7.122) with (7.123) in mind we have

$$
\begin{aligned}
&-4\mathbf{x}^{\mathrm{T}}\frac{\mathbf{Q}^{\mathrm{T}} + \mathbf{Q}}{2}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) + 4\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) \\
&\quad = -4\mathbf{x}^{\mathrm{T}}\frac{\mathbf{Q}^{\mathrm{T}} + \mathbf{Q}}{2}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) + 4\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\frac{\mathbf{Q}^{\mathrm{T}} + \mathbf{Q}}{2}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) \\
&\quad = 4\big(-\mathbf{x} + \mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\frac{\mathbf{Q}^{\mathrm{T}} + \mathbf{Q}}{2}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) = 0
\end{aligned}
$$

To simplify the expression further we shall assume that $\mathbf{Q}$ is already symmetric (by (7.123) we can do this without loss of generality, since, if $\mathbf{Q}$ is not symmetric, we can always symmetrize it without changing the quadratic form $Q(\mathbf{x})$). We thereby obtain

$$
4\big(-\mathbf{x} + \mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) = 0
$$

or

$$
\big(\mathbf{x} - \mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) = 0
$$

which must hold $\forall\mathbf{x}$:

$$
\big(\mathbf{x} - \mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\big(\mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big) = 0 \quad \forall\mathbf{x}
$$

Noticing that $(\mathbf{n}^{\mathrm{T}}\mathbf{x})$ is a scalar, we can factor it out:

$$
\big(\mathbf{x} - \mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\mathbf{n}\cdot(\mathbf{n}^{\mathrm{T}}\mathbf{x}) = 0 \quad \forall\mathbf{x}
$$

that is

$$\big(\mathbf{x} - \mathbf{n}(\mathbf{n}^{\mathrm{T}}\mathbf{x})\big)^{\mathrm{T}}\mathbf{Q}\mathbf{n} = 0 \ \vee \ \mathbf{n}^{\mathrm{T}}\mathbf{x} = 0 \quad \forall \mathbf{x} \tag{7.124}$$

Let's concentrate on the first equality in (7.124). Since $(\mathbf{n}^{\mathrm{T}}\mathbf{x})$ is a scalar, we can rewrite it as

$$\big(\mathbf{x} - (\mathbf{n}^{\mathrm{T}}\mathbf{x})\mathbf{n}\big)^{\mathrm{T}}\mathbf{Q}\mathbf{n} = 0$$

We can see the left-hand side as a dot product of two vectors:

$$\big(\mathbf{x} - (\mathbf{n}^{\mathrm{T}}\mathbf{x})\mathbf{n}\big)^{\mathrm{T}} \cdot \mathbf{Q}\mathbf{n} = 0 \tag{7.125}$$

Now looking at the first vector $\mathbf{x} - (\mathbf{n}^{\mathrm{T}}\mathbf{x})\mathbf{n}$ we notice that $(\mathbf{n}^{\mathrm{T}}\mathbf{x})\mathbf{n}$ is the projection of $\mathbf{x}$ on $\mathbf{n}$, therefore $\mathbf{x} - (\mathbf{n}^{\mathrm{T}}\mathbf{x})\mathbf{n}$ is the "projection remainder", which is thereby lying in the plane orthogonal to $\mathbf{n}$ (taking any value within that plane, since $\mathbf{x}$ is arbitrary). It follows that, in order to ensure (7.125) for all $\mathbf{x}$, vector $\mathbf{Q}\mathbf{n}$ must be collinear to $\mathbf{n}$, that is $\mathbf{n}$ must be an eigenvector of the *symmetric* matrix $\mathbf{Q}$.

We can also notice that the second condition $\mathbf{n}^{\mathrm{T}}\mathbf{x} = 0$ in (7.124) doesn't change anything at the requirement of $\mathbf{n}$ to be an eigenvector of $\mathbf{Q}$, since we still need (7.125) to hold in all cases when $\mathbf{n}^{\mathrm{T}}\mathbf{x} \neq 0$.

Principal axes of a quadratic hypersurface: in an orthonormal basis, principal axes of a quadratic curve, surface or hypersurface are eigenvectors of its *symmetrized* matrix.

Goal 7.10b completed. ■

Goal 7.10c. Demonstrate mutual orthogonality of principal axes of a quadratic hypersurface. *Hint:* start with a 2D curve and generalize to higher dimension counts.

Suppose we are given a 2D quadratic curve (an ellipse or a hyperbola). Since principal axes are the eigenvectors of the respective matrix, we are guaranteed to have at least one principal axis. Let's rotate the 2D coordinate system (x, y) in such a way that this principal axis coincides with the x axis. The curve is thereby symmetric with respect to the inversion of the x axis:

$$Q(x, y) = C \iff Q(-x, y) = C$$

Now, given

$$Q(x, y) = Q\left(\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]\right) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = q_{11}x^2 + (q_{12} + q_{21})xy + q_{22}y^2$$

we have

$$Q(-x, y) = q_{11}x^2 - (q_{12} + q_{21})xy + q_{22}y^2$$

and

$$\Big(Q(x, y) = C \iff Q(-x, y) = C\Big) \iff \Big((q_{12} + q_{21})xy = 0 \quad \forall x, y\Big)$$

which is possible only if $q_{12} + q_{21} = 0$ and respectively

$$Q(x, y) = q_{11}x^2 + q_{22}y^2$$

It immediately follows that $Q(x,y) = C \iff Q(x,-y) = C$ and therefore the y axis is the other principal axis of the curve (there cannot be more than 2 eigenvectors and respectively more than 2 principal axes). Since y is orthogonal to x, so are the principal axes of the curve.

Let's now try to generalize to an arbitrary N-dimensional hypersurface. Its matrix must have at least one eigenvector, therefore the hypersurface must have at least one principal axis. Let's rotate the N-dimensional coordinate system $(x_1, \ldots, x_N)$ in such a way that this principal axis coincides with the x_1 axis. The curve is thereby symmetric with respect to the inversion of the x_1 axis:

$$Q(x_1, x_2, \ldots, x_N) = C \iff Q(-x_1, x_2, \ldots, x_N) = C$$

Now, given

$$Q(x_1, x_2, \ldots, x_N) = \sum_{i,j=1}^{N} q_{ij} x_i x_j = q_{11} x_1^2 + \sum_{i=2}^{N} (q_{1i} + q_{i1}) x_1 x_i + \sum_{i,j=2}^{N} q_{ij} x_i x_j$$

we have

$$Q(-x_1, x_2, \ldots, x_N) = q_{11} x_1^2 - \sum_{i=2}^{N} (q_{1i} + q_{i1}) x_1 x_i + \sum_{i,j=2}^{N} q_{ij} x_i x_j$$

and

$$\Big(Q(x_1, x_2, \ldots, x_N) = C \iff Q(-x_1, x_2, \ldots, x_N) = C\Big) \iff \left(\sum_{i=2}^{N} (q_{1i} + q_{i1}) x_1 x_i = 0 \quad \forall x_1, \ldots, x_N\right)$$

The right-hand side is only possible if $q_{1i} + q_{i1} = 0 \ \forall i$. Now, recall that without loss of generality we can assume that the matrix $\mathbf{Q}$ is symmetric. In that case the condition $q_{1i} + q_{i1} = 0 \ \forall i$ is equivalent to the condition $q_{1i} = q_{i1} = 0 \ \forall i$ and

$$Q(x_1, x_2, \ldots, x_N) = q_{11} x_1^2 + \sum_{i,j=2}^{N} q_{ij} x_i x_j = q_{11} x_1^2 + Q_1(x_2, \ldots, x_N)$$

where

$$Q_1(x_2, \ldots, x_N) = \sum_{i,j=2}^{N} q_{ij} x_i x_j$$

is an $(N-1)$-dimensional quadratic form in the subspace spanned by the coordinate axes $x_2, \ldots x_N$.

We can repeat the same argument in this $(N-1)$-dimensional subspace for the form Q_1, where we would first obtain an eigenvector of the respective submatrix $\mathbf{Q}_1$ of the matrix $\mathbf{Q}$ (which is guaranteed to exist, since the submatrix $\mathbf{Q}_1$ must have at least one eigenvector). Then we would rotate the coordinate system of the subspace, aligning this eigenvector with the x_2 axis, yielding

$$Q(x_1, x_2, \ldots, x_N) = q_{11} x_1^2 + q_{22} x_2^2 + \sum_{i,j=3}^{N} q_{ij} x_i x_j = q_{11} x_1^2 + q_{22} x_2^2 + Q_2(x_2, \ldots, x_N)$$

Repeating until we exhaust all space dimensions, we obtain a *diagonal* quadratic form

$$Q(x_1, x_2, \ldots, x_N) = q_{11}x_1^2 + q_{22}x_2^2 + \ldots + q_{NN}x_N^2 = \sum_{i=1}^{N} q_{ii}x_i^2 \tag{7.126}$$

Thereby we demonstrated

The "spectral theorem" for quadratic forms: in an orthonormal basis, any quadratic form is diagonalizable by basis rotations (plus an additional symmetrization of the form, if its not originally symmetric).

▷ We refer to the above as the "spectral theorem", because its statement is pretty much the same as the statement of the spectral theorem for symmetric matrices, which we shall obtain in Subsection 7.9.2.

The diagonal quadratic form (7.126) exhibits mirror symmetries with respect to all coordinate variables $x_1, \ldots x_N$, and thereby the coordinate axes are the principal axes of the hypersurface. Since our coordinate axes are orthogonal by choice, so are the principal axes.

Goal 7.10c completed. ■

It also automatically follows from our derivation of (7.126) that not only principal axes are mutually orthogonal, but also that there are always exactly N principal axes, with N is the space dimension.

Goal 7.10 completed. ■

We can make one more observation about the principal axes. In Figs. 7.3, 7.4 and 7.5 we can notice that the curve or the surface is always orthogonal to the principal axes at the points of its intersection with the principal axes. Let's see whether this property holds in general.

Goal 7.11. Demonstrate that quadratic hypersurface is orthogonal to its principal axes at the points of intersection. *Hint:* draw a parametric line lying within the hypersurface and going through the intersection point.

Let $\mathbf{x}(t)$ be a parametric line lying within the quadratic surface, that is $\mathbf{x}(t)$ satisfies (7.119) $\forall t$:

$$\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = C \quad \forall t$$

where we implicitly understand $\mathbf{x}$ as $\mathbf{x}(t)$. Differentiating with respect to t yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}\right) = \frac{\mathrm{d}}{\mathrm{d}t}C = 0$$

Transforming the left-hand side we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}\right) &= \dot{\mathbf{x}}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{Q}\dot{\mathbf{x}} = \dot{\mathbf{x}}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \left(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\dot{\mathbf{x}}\right)^{\mathrm{T}} = \dot{\mathbf{x}}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \dot{\mathbf{x}}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{x} \\ &= \dot{\mathbf{x}}^{\mathrm{T}}(\mathbf{Q} + \mathbf{Q}^{\mathrm{T}})\mathbf{x} = 2 \cdot \dot{\mathbf{x}}^{\mathrm{T}}\frac{\mathbf{Q} + \mathbf{Q}^{\mathrm{T}}}{2}\mathbf{x} = 0\end{aligned}$$

Without loss of generality, we can assume that $\mathbf{Q}$ is already symmetric and replace the latest equation with

$$\dot{\mathbf{x}}^{\mathrm{T}}\mathbf{Q}\mathbf{x} = 0$$

Now at the points of intersection of the hypersurface with its principal axis, $\mathbf{x}(t)$ becomes an eigenvector of $\mathbf{Q}$, and the above equation transforms into

$$\dot{\mathbf{x}}^{\mathrm{T}} \cdot \lambda\mathbf{x} = \lambda \cdot \dot{\mathbf{x}}^{\mathrm{T}} \cdot \mathbf{x} = 0$$

where λ is the respective eigenvalue. Unless $\lambda = 0$, which signifies a singular case, it follows that

$$\dot{\mathbf{x}}^{\mathrm{T}} \cdot \mathbf{x} = 0$$

that is the curve $\mathbf{x}(t)$ is going orthogonally to the vector $\mathbf{x}(t)$ at the point of intersection. Since the curve $\mathbf{x}(t)$ is arbitrary, it follows that the entire hypersurface is orthogonal to $\mathbf{x}(t)$ at this point.

Goal 7.11 completed. ■

7.9.2 Symmetric and Hermitian matrices

Symmetric matrices

In Subsection 7.9.1 we found out that the principal axes of quadratic curves, surfaces and hypersurfaces correspond to the eigenvectors of the matrices of their respective quadratic forms, and we also found that these principal axes are mutually orthogonal and there are always exactly N of those. It follows that the (symmetric) matrices of quadratic forms admit an orthogonal eigenbasis (in which these quadratic forms become diagonal). However, given an arbitrary symmetric matrix, we can always build a quadratic form around that matrix. It follows that any symmetric matrix admits an orthogonal eigenbasis.

From the perspective of the properties of a generic symmetric matrix (not semantically related to a quadratic form) such reference to a quadratic form is, however, pretty artificial: given a symmetric matrix, we construct a symmetric quadratic form, which is diagonalizable by rotations and so is our original matrix, which thereby admits an orthogonal eigenbasis. This might qualify for a formal proof, but offers a limited amount of intuitive insights in regards to the properties of symmetric matrices per se. We would therefore like to demonstrate the same statement for the symmetric matrices directly.

Goal 7.12. Demonstrate that symmetric matrices admit an orthogonal eigenbasis. *Hint:* multiplication of an eigenvector by the matrix doesn't change the direction of the eigenvector.

Let $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$ be a symmetric $N \times N$ matrix. Given two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ of this matrix $\mathbf{A}$, we need to deduce

$$\mathbf{v}_1^{\mathrm{T}} \cdot \mathbf{v}_2 = 0 \tag{7.127}$$

We need to make use of the fact that $\mathbf{A}$ is symmetric and somehow connect this fact to (7.127). Recall that multiplication of an eigenvector by $\mathbf{A}$ doesn't change the eigenvector's direction, more precisely it produces a vector which is

collinear to the original eigenvector, so if $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal, then so are $\mathbf{v}_1$ and $\mathbf{A}\mathbf{v}_2$ and we have

$$\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_2 = 0 \tag{7.128}$$

which can be seen as a bilinear form applied to the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Actually, (7.127) and (7.128) are almost equivalent, they would be truly equivalent given $\lambda_2 \neq 0$, but we'll ignore that detail, since we are not going to formally rely on the equivalence of these two statements.

Alternatively, we can notice that $\mathbf{A}\mathbf{v}_1$ is collinear to $\mathbf{v}_1$. Replacing $\mathbf{v}_1$ by $\mathbf{A}\mathbf{v}_1$ in (7.127) we obtain $(\mathbf{A}\mathbf{v}_1)^{\mathrm{T}}\mathbf{v}_2 = 0$. The left-hand side of the latter equation can be transformed as

$$(\mathbf{A}\mathbf{v}_1)^{\mathrm{T}}\mathbf{v}_2 = \mathbf{v}_1^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{v}_2 = \mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_2 = 0$$

That is, we have again arrived at the same equation (7.128) but this time we explicitly used the symmetry of $\mathbf{A}$. We might be getting somewhere by comparing the respective two paths connecting (7.127) and (7.128).

Adding detail to the previous observations, let's take the left-hand side of (7.128) and rewrite it in two different ways, each time aiming to arrive at the left-hand side of (7.127). By making use of the fact that $\mathbf{v}_2$ is an eigenvector, we can rewrite the left-hand side of (7.128) as

$$\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_2 = \mathbf{v}_1^{\mathrm{T}} \cdot \lambda_2\mathbf{v}_2 = \lambda_2 \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 \tag{7.129a}$$

Similarly, by making use of the fact that $\mathbf{v}_1$ is an eigenvector and also by making use of the fact that $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$, we can rewrite the left-hand side of (7.128) as

$$\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_2 = \mathbf{v}_1^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{v}_2 = (\mathbf{A}\mathbf{v}_1)^{\mathrm{T}}\mathbf{v}_2 = (\lambda_1\mathbf{v}_1)^{\mathrm{T}}\mathbf{v}_2 = \lambda_1 \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 \tag{7.129b}$$

Since the left-hand sides of (7.129a) and (7.129b) are equal, so are their right-hand sides:

$$\lambda_1 \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 = \lambda_2 \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 \tag{7.130}$$

But this is possible only if either $\lambda_1 = \lambda_2$ or $\mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 = 0$. We thereby showed that eigenvectors corresponding to *different* eigenvalues of a symmetric matrix are indeed orthogonal to each other. However, what about identical eigenvalues $\lambda_1 = \lambda_2$? Theoretically this could either mean that we get an eigenplane, in which we can choose two mutually orthogonal eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$, but we also could obtain a Jordan chain instead. We need to demonstrate that the latter option cannot occur here.

Trying out the same idea, this time, instead of analyzing the bilinear form $\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_2$, which is built around two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ and reduces to $\lambda_{1,2} \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2$, we shall analyze the bilinear form $\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1'$, where $\mathbf{v}_1'$ would be the rank 2 generalized eigenvector from the same Jordan chain as $\mathbf{v}_1$. The idea is that $(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1$ and respectively $\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^{\mathrm{T}} \cdot \mathbf{v}_1$. Thus in place of (7.129a) we obtain

$$\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^{\mathrm{T}} \cdot \mathbf{v}_1 = |\mathbf{v}_1|^2 \tag{7.131a}$$

while in place of (7.129b), noticing that $(\mathbf{A} - \lambda_1)^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} - \lambda_1 = \mathbf{A} - \lambda_1$, we obtain

$$\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)^{\mathrm{T}} \cdot \mathbf{v}_1' = ((\mathbf{A} - \lambda_1)\mathbf{v}_1)^{\mathrm{T}} \cdot \mathbf{v}_1' = 0 \cdot \mathbf{v}_1' = 0 \tag{7.131b}$$

Again, since left-hand sides of (7.131a) and (7.131b) are equal, so are their right-hand sides, yielding

$$|\mathbf{v}_1|^2 = 0$$

But this is impossible, as an eigenvector must have a nonzero length. We must conclude that there is no rank 2 generalized eigenvector $\mathbf{v}_1'$ in the Jordan chain and the Jordan chain associated with $\mathbf{v}_1$ has length 1.

It might look like we have demonstrated that any symmetric matrix $\mathbf{A}$ admits an orthogonal eigenbasis, thereby completing the Goal 7.12. However that conclusion would be a bit premature: we were implicitly assuming that all matrix and vector components are real. But the eigenvectors of an arbitrary real matrix are not guaranteed to have real components. On the contrary, a real matrix can easily have complex eigenvalues and respectively complex-component eigenvectors. And as soon as there are complex numbers involved, we should have used $\mathbf{x}^*\mathbf{y}$ rather than $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ as the dot product (as well as should have used Hermitian transposition $\mathbf{A}^*$ instead of $\mathbf{A}^{\mathrm{T}}$). With $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ as the dot product our results would be questionable, as discussed in Section 3.6. We shall therefore reconsider the previously discussed material once again, this time assuming a complex field.

Hermitian matrices

In case of complex field the same argument can be generalized to Hermitian matrices $\mathbf{A}^* = \mathbf{A}$. In this case we'll have to replace transposition by Hermitian transposition everywhere, in particular the bilinear dot product $\mathbf{v}_1^{\mathrm{T}} \cdot \mathbf{v}_2$ turns into a sesquilinear one $\mathbf{v}_1^* \cdot \mathbf{v}_2$. Thus (7.127) is replaced by

$$\mathbf{v}_1^{\mathrm{T}} \cdot \mathbf{v}_2 \stackrel{?}{=} 0 \tag{7.132}$$

(where the question mark signifies that we are not sure yet whether this equality is going to hold for eigenvectors of an arbitrary Hermitian matrix or not) and (7.128) turns into

$$\mathbf{v}_1^*\mathbf{A}\mathbf{v}_2 \stackrel{?}{=} 0 \tag{7.133}$$

Respectively (7.129) turn into

$$\mathbf{v}_1^*\mathbf{A}\mathbf{v}_2 = \mathbf{v}_1^* \cdot \lambda_2\mathbf{v}_2 = \lambda_2 \cdot \mathbf{v}_1^*\mathbf{v}_2 \tag{7.134a}$$

and

$$\mathbf{v}_1^*\mathbf{A}\mathbf{v}_2 = \mathbf{v}_1^*\mathbf{A}^*\mathbf{v}_2 = (\mathbf{A}\mathbf{v}_1)^*\mathbf{v}_2 = (\lambda_1\mathbf{v}_1)^*\mathbf{v}_2 = \lambda_1^* \cdot \mathbf{v}_1^*\mathbf{v}_2 \tag{7.134b}$$

and (7.130) becomes

$$\lambda_1^* \cdot \mathbf{v}_1^*\mathbf{v}_2 = \lambda_2 \cdot \mathbf{v}_1^*\mathbf{v}_2 \tag{7.135}$$

Now (7.135) looks a little bit awkward. Does it mean that it demonstrates the mutual orthogonality of $\mathbf{v}_1$ and $\mathbf{v}_2$ in case $\lambda_1^* \neq \lambda_2$? But shouldn't we rather expect $\mathbf{v}_1$ and $\mathbf{v}_2$ to be orthogonal if $\lambda_1 \neq \lambda_2$? To answer that question, let's formally let $\mathbf{v}_1 = \mathbf{v}_2$ and respectively $\lambda_1 = \lambda_2$ in (7.135). In this case (7.135) reads

$$\lambda_1^* \cdot \mathbf{v}_1^*\mathbf{v}_1 = \lambda_1 \cdot \mathbf{v}_1^*\mathbf{v}_1$$

or simply

$$\lambda_1 \cdot |\mathbf{v}_1|^2 = \lambda_1^* \cdot |\mathbf{v}_1|^2$$

Since $|\mathbf{v}_1|^2 \neq 0$, it follows that $\lambda_1 = \lambda_1^*$, that is all eigenvalues of a Hermitian matrix are real:

(spectra of Hermitian matrices are real)

$$(\mathbf{A}^* = \mathbf{A}) \implies (\lambda_i \in \mathbb{R}\ \forall i) \tag{7.136}$$

With (7.136) in mind, we can rewrite (7.135) as

$$\lambda_1 \cdot \mathbf{v}_1^* \mathbf{v}_2 = \lambda_2 \cdot \mathbf{v}_1^* \mathbf{v}_2 \tag{7.137}$$

and the mutual orthogonality (with respect to the sesquilinear dot product $\mathbf{v}_1^* \mathbf{v}_2$) of the eigenvectors of a Hermitian matrix follows.

It remains to show that all Jordan chains of a Hermitian matrix have lengths equal to 1, which we do by writing a complex counterpart of (7.131). For a Hermitian $\mathbf{A}$, instead of the bilinear form $\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1'$ we shall consider the bilinear form $\mathbf{v}_1^*(\mathbf{A} - \lambda_1)\mathbf{v}_1'$, obtaining

$$\mathbf{v}_1^*(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^* \cdot (\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^* \cdot \mathbf{v}_1 = |\mathbf{v}_1|^2 \tag{7.138a}$$

and, noticing that $(\mathbf{A} - \lambda_1)^* = \mathbf{A}^* - \lambda_1^* = \mathbf{A} - \lambda_1$ (since $\lambda_1 \in \mathbb{R}$),

$$\mathbf{v}_1^*(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^*(\mathbf{A} - \lambda_1)^* \cdot \mathbf{v}_1' = ((\mathbf{A} - \lambda_1)\mathbf{v}_1)^* \cdot \mathbf{v}_1' = 0 \cdot \mathbf{v}_1' = 0 \tag{7.138b}$$

Equating the right-hand sides of (7.138) yields

$$|\mathbf{v}_1|^2 = 0$$

and we again obtain that all Jordan chains have the length 1.

So, a Hermitian $N \times N$ matrix has no generalized eigenvectors, that is it has N eigenvectors, which are either already mutually orthogonal, or (in case of repeated eigenvalues) can be chosen to be mutually orthogonal (with respect to the sesquilinear dot product $\mathbf{x}^*\mathbf{y}$). Since the lengths of the eigenvectors can be chosen arbitrary, we can force all of them to have unity length, thereby obtaining a unitary eigenbasis. Thus we arrive at

The spectral theorem for Hermitian matrices: a Hermitian matrix admits a unitary eigenbasis.

Since purely real symmetric matrices are a particular case of Hermitian matrices, the statement of the spectral theorem applies to them as well. In this case we have a real matrix with real eigenvalues, which also implies that its eigenvectors have purely real components. So the unitary eigenbasis can be referred to as simply orthonormal in this case and we obtain

The spectral theorem for symmetric matrices: a real symmetric matrix admits a real orthonormal eigenbasis.

Goal 7.12 completed. ■

A little bit earlier, we have found that symmetric and Hermitian matrices have real spectra (that is, all of their eigenvalues are real). It might be interesting to investigate the converse statement: whether a real spectrum implies a symmetric or Hermitian matrix.

Goal 7.13. Investigate whether a real spectrum automatically implies that the matrix is symmetric or Hermitian. *Hint:* diagonalize the matrix first.

Apparently the statement doesn't hold for all matrices with real spectra. E.g. a Jordan block (of size 2 or higher) with a real eigenvalue is not Hermitian or symmetric. So we might already exclude matrices which do not admit an eigenbasis (that is those which have generalized eigenvectors). The remaining matrices all can be brought to their respective diagonal forms by a similarity transformation: given a matrix $\mathbf{A}$, which admits an eigenbasis, we consider its diagonal form $\mathbf{A}' = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$, where $\mathbf{T}$ is the diagonalizing transformation. Being diagonal and having real spectrum, the matrix $\mathbf{A}'$ is both Hermitian and symmetric. The question is, whether this property is preserved by the inverse similarity transformation $\mathbf{A} = \mathbf{T}\mathbf{A}'\mathbf{T}^{-1}$.

We can notice that if $\mathbf{T}$ is unitary, $\mathbf{T}^{-1} = \mathbf{T}^*$, then the diagonalizing similarity transformation turns into a (complex version of) congruence transformation $\mathbf{A} = \mathbf{T}\mathbf{A}'\mathbf{T}^*$. We therefore obtain

(hermiticity is preserved by unitary similarity transformations)

$$\begin{aligned}\mathbf{A}^* &= \left(\mathbf{T}\mathbf{A}'\mathbf{T}^{-1}\right)^* = (\mathbf{T}\mathbf{A}'\mathbf{T}^*)^* = \mathbf{T}(\mathbf{A}')^*\mathbf{T}^* = \mathbf{T}\mathbf{A}'\mathbf{T}^* = \mathbf{T}\mathbf{A}'\mathbf{T}^{-1} \\ &= \mathbf{A}\end{aligned} \tag{7.139}$$

It follows that the matrix $\mathbf{A}$ (admitting a unitary eigenbasis) is Hermitian iff so is its diagonal form $\mathbf{A}'$, which in turn happens iff $\mathbf{A}'$ has a real spectrum, which happens iff $\mathbf{A}$ has a real spectrum (since the spectra of both matrices are identical). We thus arrive at

Corollary of the spectral theorem for Hermitian matrices: a matrix admitting a unitary eigenbasis is Hermitian iff its spectrum is real.

In case of purely real numbers, unitary matrices are replaced by orthogonal matrices ($\mathbf{T}^{-1} = \mathbf{T}^{\mathrm{T}}$), so the diagonalizing similarity transformation turns into a (real) congruence transformation, and therefore

(symmetry is preserved by orthogonal similarity transformations)

$$\begin{aligned}\mathbf{A}^{\mathrm{T}} &= \left(\mathbf{T}\mathbf{A}'\mathbf{T}^{-1}\right)^{\mathrm{T}} = \left(\mathbf{T}\mathbf{A}'\mathbf{T}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{T}(\mathbf{A}')^{\mathrm{T}}\mathbf{T}^{\mathrm{T}} = \mathbf{T}\mathbf{A}'\mathbf{T}^{\mathrm{T}} = \mathbf{T}\mathbf{A}'\mathbf{T}^{-1} \\ &= \mathbf{A}\end{aligned} \tag{7.140}$$

We thereby arrive at a similar corollary, stating that *a real matrix admitting a (real) orthonormal eigenbasis is symmetric iff all its eigenvalues are real.* However a real matrix admitting a real orthonormal eigenbasis automatically has a real spectrum and is thereby automatically symmetric:

Corollary of the spectral theorem for symmetric matrices: a real matrix admitting a (real) orthonormal eigenbasis is symmetric.

▷ It is easy to see from the derivation of (7.139) that hermiticity is preserved by any (complex) congruence transformation, the transformation matrix doesn't need to be unitary. Similarly, from (7.140) one can see that symmetry is preserved by any (real) congruence transformation. Since bilinear forms undergo a congruence transformation under a basis change, it follows that a bilinear form that is symmetric (or conjugate-symmetric, in case of complex numbers) in one basis is simultaneously symmetric in all other bases, including non-orthogonal (or non-unitary) bases.

It remains to find out whether we really need to require a matrix to admit a unitary (or orthonormal) eigenbasis in order to be able to deduce hermiticity (or symmetry) from a real spectrum. We could look at the characteristic polynomial of a generic 2×2 matrix $\mathbf{A} = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ which is

$$\det(\lambda - \mathbf{A}) = (\lambda - a)(\lambda - d) - bc = \lambda^2 - (a + d)\lambda + ad - bc$$

Its determinant is respectively

$$D = (a + d)^2 - 4(ad - bc) = (a - d)^2 - 4bc$$

so $D \geq 0 \iff (a-d)^2 \geq 4bc$ and its easy to see that there are plenty of options for $D \geq 0$ and respectively for the eigenvalues being real while $b \neq c^*$, that is $\mathbf{A}$ being non-Hermitian (or nonsymmetric). Thus we cannot simply drop the requirement of $\mathbf{A}$ admitting a unitary (or orthonormal) eigenbasis.

Goal 7.13 completed. ■

7.9.3 Spectrum of a matrix transpose

Ordinary transpose

In Subsection 7.9.2 we initially attempted to derive the spectral theorem directly for symmetric matrices. This derivation was not very successful due to the possibility of complex eigenvalues, but it provided a good intermediate step to arrive at the spectral theorem for Hermitian matrices. We shall continue using the same derivation as the starting point to gain more insights. In particular, we can notice that the derivation doesn't really require the matrix $\mathbf{A}$ to be symmetric. It might suffice that $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ have the same effect on eigenvectors (and generalized eigenvectors, if any):

$$\mathbf{A}\mathbf{v} = \mathbf{A}^{\mathrm{T}}\mathbf{v} \tag{7.141}$$

where $\mathbf{v}$ is an eigenvector or a generalized eigenvector. Indeed, we have been applying the bilinear form $\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_2$ introduced in (7.128) only to eigenvectors of $\mathbf{A}$. Similarly, the bilinear form $\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1'$ in (7.138) was applied only for an eigenvector $\mathbf{v}_1$ and the respective rank 2 generalized eigenvector $\mathbf{v}_1'$. Thus, the statement about mutual orthogonality of the eigenvectors (with respect to the dot product $\mathbf{x}^{\mathrm{T}}\mathbf{y}$) as well as about the lack of generalized eigenvectors already holds under the condition (7.141).

In regards to the eigenvectors, the condition (7.141) effectively states that $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ share eigenvectors and that each of the eigenvectors corresponds to one and the same eigenvalue for both $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$. Actually, we can notice that $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ always share their eigenvalues, since the eigenvalue equation $\det(\lambda - \mathbf{A}) = 0$ is invariant to matrix transposition. It is however not granted that, for an arbitrary $\mathbf{A}$, the matrices $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ share the eigenvectors.

E.g. we could take the matrix (7.62b) and let $p = 2$, $p^2 - q = 1$ (that is $q = 3$), so that by (7.68) we obtain two non-orthogonal eigenvectors $\mathbf{v}_1 = \left[\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right]$ and $\mathbf{v}_2 = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$, corresponding to the eigenvalues $\lambda_1 = 3$, $\lambda_2 = 1$ given by (7.66). In this case ($p = 2$, $q = 3$) we have $\mathbf{A} = \left[\begin{smallmatrix} 4 & -3 \\ 1 & 0 \end{smallmatrix}\right]$ and $\mathbf{A}^{\mathrm{T}} = \left[\begin{smallmatrix} 4 & 1 \\ -3 & 0 \end{smallmatrix}\right]$. Then by direct computation we obtain $\mathbf{A}^{\mathrm{T}}\mathbf{v}_1 = \left[\begin{smallmatrix} 13 \\ -9 \end{smallmatrix}\right]$ and $\mathbf{A}^{\mathrm{T}}\mathbf{v}_2 = \left[\begin{smallmatrix} 5 \\ -3 \end{smallmatrix}\right]$, that is neither $\mathbf{v}_1$ nor $\mathbf{v}_2$ is an eigenvector of $\mathbf{A}^{\mathrm{T}}$.

Therefore, if we want (7.141) to hold, we have to explicitly assume that $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ share eigenvectors. As mentioned, we also need these eigenvectors to be mapped to the same eigenvalues with $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$. However this latter condition sounds like it might hold automatically, without explicitly requiring it to be the case.

As an intuitive justification for the latter (eigenvectors automatically being mapped to the same eigenvalues) consider the following. Suppose for a particular $\mathbf{A}$ it is given that $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ share their eigenvectors. For simplicity we shall additionally assume for now that there are no generalized eigenvectors, that is both $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ are diagonalizable. As we already know that $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ share their eigenvalues anyway, it would be only a question of whether the mapping between eigenvectors and eigenvalues is identical for $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ or whether it is "shuffled". But thinking about it, it would be difficult to imagine any factors which could define a particular "shuffle order" (different from the "default order" which is an identical mapping for both $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$). With this handwavy justification in mind, let's try to come up with a more rigorous argument.

Goal 7.14. Given $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ which share their eigenvectors, investigate the applicability of the spectral theorem. *Hint:* use the steps of the derivation of the spectral theorem for symmetric and Hermitian matrices as a reference.

Let $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ share all of their eigenvectors. Now consider our "favorite" bilinear form $\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_2$ introduced in (7.128) and let $\mathbf{v}_2 = \mathbf{v}_1$, yielding the quadratic form $\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_1$, where $\mathbf{v}_1$ is a shared eigenvector of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$. Since $\mathbf{v}_1$ is an eigenvector of $\mathbf{A}$, we have

$$\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_1 = \mathbf{v}_1^{\mathrm{T}} \cdot \lambda_1\mathbf{v}_1 = \lambda_1 \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1 \tag{7.142a}$$

Since $\mathbf{v}_1$ is also an eigenvector of $\mathbf{A}^{\mathrm{T}}$, we have

$$\mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_1 = (\mathbf{A}^{\mathrm{T}}\mathbf{v}_1)^{\mathrm{T}}\mathbf{v}_1 = (\tilde{\lambda}_1\mathbf{v}_1)^{\mathrm{T}}\mathbf{v}_1 = \tilde{\lambda}_1 \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1 \tag{7.142b}$$

where $\tilde{\lambda}_1$ is the eigenvalue of $\mathbf{A}^{\mathrm{T}}$ corresponding to the eigenvector $\mathbf{v}_1$. Since the left-hand sides of (7.142a) and (7.142b) are equal, so are their right-hand sides:

$$\lambda_1 \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1 = \tilde{\lambda}_1 \cdot \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1$$

or simply

$$\lambda_1 \cdot |\mathbf{v}_1|^2 = \tilde{\lambda}_1 \cdot |\mathbf{v}_1|^2$$

It follows that $\lambda_1 = \tilde{\lambda}_1$, since $|\mathbf{v}_1| \neq 0$.

So, we have obtained that shared eigenvectors of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ are always mapped to the same eigenvalues no matter what. It remains to check for the existence of generalized eigenvectors. Thinking about it, we could come to the conjecture that maybe there cannot be any generalized eigenvectors, if all eigenvectors of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ are shared. Indeed, imagining a Jordan normal form $\mathbf{A}$ we can notice that, after transposition, each Jordan block (7.104) attains the form

$$\mathbf{J}^{\mathrm{T}} = \begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 & 0 \\ 1 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 1 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 0 \\ 0 & 0 & 0 & \cdots & 1 & \lambda \end{bmatrix}$$

This in turn implies that

$$\begin{aligned} \mathbf{J}\mathbf{e}_1 &= \lambda\mathbf{e}_1 + \mathbf{e}_2 \\ \mathbf{J}\mathbf{e}_2 &= \lambda\mathbf{e}_2 + \mathbf{e}_3 \\ &\vdots \\ \mathbf{J}\mathbf{e}_{\mu-1} &= \lambda\mathbf{e}_{\mu-1} + \mathbf{e}_\mu \\ \mathbf{J}\mathbf{e}_\mu &= \lambda\mathbf{e}_\mu \end{aligned}$$

where

$$\begin{aligned} \mathbf{e}_1 &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}^{\mathrm{T}} \\ \mathbf{e}_2 &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \end{bmatrix}^{\mathrm{T}} \\ &\vdots \\ \mathbf{e}_\mu &= \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^{\mathrm{T}} \end{aligned}$$

are the basis vectors. We can notice that the order of arrangement of $\mathbf{e}_i$ into the Jordan chain is thereby opposite to the usual one. In particular, the leading eigenvector of the chain is $\mathbf{e}_\mu$ rather than $\mathbf{e}_1$. This means that $\mathbf{J}$ and $\mathbf{J}^{\mathrm{T}}$ cannot share an eigenvector, unless the size of the Jordan chain is 1. Our conjecture is that this property (the incompatibility of Jordan chains with shared eigenvectors between a matrix and its transpose) holds not only for Jordan normal form matrices, but generally. Let's verify it.

Taking the bilinear form $\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1'$ introduced in (7.131) we on one hand have

$$\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^{\mathrm{T}} \cdot (\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^{\mathrm{T}} \cdot \mathbf{v}_1 = |\mathbf{v}_1|^2 \tag{7.143a}$$

which simply repeats (7.131a), on the other hand in place of (7.131b) we obtain

$$\mathbf{v}_1^{\mathrm{T}}(\mathbf{A} - \lambda_1)\mathbf{v}_1' = ((\mathbf{A} - \lambda_1)^{\mathrm{T}}\mathbf{v}_1)^{\mathrm{T}} \cdot \mathbf{v}_1' = ((\mathbf{A} - \lambda_1)\mathbf{v}_1)^{\mathrm{T}} \cdot \mathbf{v}_1' = 0 \cdot \mathbf{v}_1' = 0 \tag{7.143b}$$

where we noticed that

$$(\mathbf{A} - \lambda_1)^{\mathrm{T}}\mathbf{v}_1 = \mathbf{A}^{\mathrm{T}}\mathbf{v}_1 - \lambda_1\mathbf{v}_1 = \mathbf{A}\mathbf{v}_1 - \lambda_1\mathbf{v}_1 = (\mathbf{A} - \lambda_1)\mathbf{v}_1$$

Since the left-hand sides of (7.143a) and (7.143b) are equal, so are their right-hand sides:

$$|\mathbf{v}_1|^2 = 0$$

and we again run into a contradiction, implying that $\mathbf{A}$ cannot have generalized eigenvectors.

Thus, if all eigenvectors of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ are shared, then the eigenvectors are also sharing the same eigenvalues and there are also no generalized eigenvectors. The matrix $\mathbf{A}$ thereby behaves as if it was symmetric, as long as the argument of the spectral theorem goes, and we could in principle infer the orthogonality of the eigenbasis of $\mathbf{A}$. However, we already know that the argument of the spectral theorem which we made for symmetric matrices works only if purely real numbers are involved, and a purely real matrix $\mathbf{A}$ admitting a purely real orthogonal eigenbasis is already truly symmetric to begin with (as we found out earlier in a corollary to the spectral theorem for symmetric matrices). Or, even simpler, since in this case (7.141) holds for all vectors of a basis (which is the eigenbasis), it also holds for all vectors, and it follows that $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$. Hence, in the discussion above we didn't really analyze a more general case. All in all it doesn't seem like we have obtained a new useful result. However, same as with our discussion of symmetric matrices in Subsection 7.9.2, we have provided a basis for the further discussion based on the sesquilinear dot product $\mathbf{x}^*\mathbf{y}$, which should be able to handle the complex numbers case, and where we won't be able to automatically infer $\mathbf{A} = \mathbf{A}^*$.

Hermitian transpose

We have attempted to generalize the result of Subsection 7.9.2 to a class of matrices which are not truly symmetric, but behave as if they are symmetric for the purposes of the spectral theorem. This attempt was not too successful, since we have found out that, under the restriction of real numbers, the case we considered turned out to be no more general than the original one. Let's now lift this restriction and repeat the same argument for a matrix $\mathbf{A}$ which shares all of its eigenvectors with its Hermitian transpose $\mathbf{A}^*$, respectively also using $\mathbf{x}^*\mathbf{y}$ as the dot product.

As a counterpart of (7.142) we obtain

$$\mathbf{v}_1^*\mathbf{A}\mathbf{v}_1 = \mathbf{v}_1^* \cdot \lambda_1\mathbf{v}_1 = \lambda_1 \cdot \mathbf{v}_1^*\mathbf{v}_1 \tag{7.144a}$$

and

$$\mathbf{v}_1^*\mathbf{A}\mathbf{v}_1 = (\mathbf{A}^*\mathbf{v}_1)^*\mathbf{v}_1 = (\tilde{\lambda}_1\mathbf{v}_1)^*\mathbf{v}_1 = \tilde{\lambda}_1^* \cdot \mathbf{v}_1^*\mathbf{v}_1 \tag{7.144b}$$

where $\tilde{\lambda}_1$ is the eigenvalue of $\mathbf{A}^*$ corresponding to the eigenvector $\mathbf{v}_1$.

The left-hand sides of (7.144a) and (7.144b) are equal and so must be their right-hand sides:

$$\lambda_1 \cdot \mathbf{v}_1^*\mathbf{v}_1 = \tilde{\lambda}_1^* \cdot \mathbf{v}_1^*\mathbf{v}_1$$

or

$$\lambda_1 \cdot |\mathbf{v}_1|^2 = \tilde{\lambda}_1^* \cdot |\mathbf{v}_1|^2$$

and since $|\mathbf{v}_1| \neq 0$, it follows that $\lambda_1 = \tilde{\lambda}_1^*$, or, generally,

(the eigenvalues corresponding to a shared eigenvector of a matrix and its Hermitian transpose are mutually conjugate)

$$\lambda_i = \tilde{\lambda}_i^* \tag{7.145}$$

This looks very similar to the result we obtained in (7.135), where we had a complex conjugate of an eigenvalue instead of simply having an eigenvalue. Back then we have found out that all eigenvalues are real and could simply drop the complex conjugation. This time however we cannot count on all eigenvalues being real. Indeed, imagine e.g. a matrix which is already diagonal, with essentially complex diagonal components. Apparently, Hermitian transposition of such matrix doesn't change its eigenvectors while it replaces the eigenvalues with their complex conjugates. This means there *are* matrices sharing eigenvectors with their Hermitian transposes such that their respective eigenvalues are not equal (but are related by complex conjugation).

▷ A further observation in this regard is that the solutions of the $\det(\lambda - \mathbf{A}) = 0$ equation and the $\det(\lambda - \mathbf{A}^*) = 0$ are mutual conjugates of each other. Thus, the eigenvalues of $\mathbf{A}^*$ are complex conjugates of those of $\mathbf{A}$.

We thereby obtained that in case of shared eigenvectors the respective eigenvalues of $\mathbf{A}$ and $\mathbf{A}^*$ are complex conjugates of each other. So the effects of $\mathbf{A}$ and $\mathbf{A}^*$ on their shared eigenvectors are not exactly identical. But this might be actually exactly what we need in order to repeat the argument of the spectral theorem for Hermitian matrices. Let's see.

Trying to repeat the steps done in (7.134), we have (7.134a) unchanged:

$$\mathbf{v}_1^* \mathbf{A} \mathbf{v}_2 = \mathbf{v}_1^* \cdot \lambda_2 \mathbf{v}_2 = \lambda_2 \cdot \mathbf{v}_1^* \mathbf{v}_2 \tag{7.146a}$$

while (7.134b) can be changed into

$$\mathbf{v}_1^* \mathbf{A} \mathbf{v}_2 = (\mathbf{A}^* \mathbf{v}_1)^* \mathbf{v}_2 = (\tilde{\lambda}_1 \mathbf{v}_1)^* \mathbf{v}_2 = \tilde{\lambda}_1^* \cdot \mathbf{v}_1^* \mathbf{v}_2 = \lambda_1 \cdot \mathbf{v}_1^* \mathbf{v}_2 \tag{7.146b}$$

where we have used (7.145).

Since the left-hand sides of (7.146a) and (7.146b) are equal, so are their right-hand sides:

$$\lambda_1 \cdot \mathbf{v}_1^* \mathbf{v}_2 = \lambda_2 \cdot \mathbf{v}_1^* \mathbf{v}_2$$

and it follows that the eigenvectors of $\mathbf{A}$ are orthogonal (or unitary, if we require them to have unit lengths). Looks like we exactly managed to repeat the argument of the spectral theorem for Hermitian matrices, as long as generalized eigenvectors are not involved.

To check for the existence of generalized eigenvectors we again use the bilinear form $\mathbf{v}_1^*(\mathbf{A} - \lambda_1)\mathbf{v}_1'$. Using (7.138) as a reference, we have (7.138a) unchanged:

$$\mathbf{v}_1^*(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^* \cdot (\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^* \cdot \mathbf{v}_1 = |\mathbf{v}_1|^2 \tag{7.147a}$$

while (7.138b) becomes

$$\mathbf{v}_1^*(\mathbf{A} - \lambda_1)\mathbf{v}_1' = \mathbf{v}_1^*(\mathbf{A}^* - \lambda_1^*)^* \cdot \mathbf{v}_1' = ((\mathbf{A}^* - \lambda_1^*)\mathbf{v}_1)^* \cdot \mathbf{v}_1'$$

$$= ((\mathbf{A}^* - \tilde{\lambda}_1)\mathbf{v}_1)^* \cdot \mathbf{v}_1' = 0 \cdot \mathbf{v}_1' = 0 \tag{7.147b}$$

Since the left-hand sides of (7.147a) and (7.147b) are equal, so are their right-hand sides:

$$|\mathbf{v}_1|^2 = 0$$

and we have run into the usual contradiction, implying that $\mathbf{A}$ doesn't have generalized eigenvectors.

It follows that the statement of the spectral theorem for Hermitian matrices also applies to the matrices which simply share all their eigenvectors with their Hermitian transpose. The corollary concerning the real spectrum doesn't apply: as we have seen in a simple example of a diagonal matrix with essentially complex components, the eigenvalues of such matrix are not necessarily real.

So, we have found that a matrix sharing its eigenvectors with its transpose is sufficient for admitting a unitary eigenbasis. But is it necessary? Let $\mathbf{A}$ admit a unitary eigenbasis. Then the matrix $\mathbf{T}$ of the diagonalizing similarity transformation of $\mathbf{A}$ is unitary ($\mathbf{T}^{-1} = \mathbf{T}^*$), thus it turns into a (complex) congruence transformation:

$$\mathbf{A}' = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{T}^*\mathbf{A}\mathbf{T}$$

Applying the same transformation to $\mathbf{A}^*$ yields

$$(\mathbf{A}^*)' = \mathbf{T}^*\mathbf{A}^*\mathbf{T} = (\mathbf{T}^*\mathbf{A}\mathbf{T})^* = (\mathbf{T}^{-1}\mathbf{A}\mathbf{T})^* = (\mathbf{A}')^*$$

But $\mathbf{A}'$ is diagonal by construction, therefore so is its Hermitian transpose $(\mathbf{A}')^*$ and therefore so is $(\mathbf{A}^*)'$, which is $\mathbf{A}^*$ brought to the eigenbasis of $\mathbf{A}$. Since $(\mathbf{A}^*)'$ turned out to be diagonal, these new basis vectors are also the eigenvectors of $\mathbf{A}^*$, that is $\mathbf{A}^*$ has the same eigenvectors as $\mathbf{A}$.

Thus the condition is sufficient and necessary, and we obtain

The spectral theorem "for a Hermitian transpose pair": a matrix shares eigenvectors with its Hermitian transpose iff it admits a unitary eigenbasis.

▷ The above is essentially the spectral theorem for normal matrices, which is going to be discussed in Subsection 7.9.6. We just want to arrive at the idea of normal matrices step by step.

Goal 7.14 completed. ■

7.9.4 Orthogonal and unitary matrices

Unitary matrices

We have obtained that a matrix which shares eigenvectors with its Hermitian transpose admits a unitary eigenbasis. One particular case of that are symmetric and Hermitian matrices, where the matrix and its transpose are the same, therefore they are sharing the eigenvectors. But can we think of another type of matrices sharing the eigenvectors with its transpose?

We can realize that $\mathbf{A}$ and $\mathbf{A}^{-1}$ should share eigenvectors. Indeed, $\mathbf{A}^{-1}$ "undoes" the effect of $\mathbf{A}$, therefore if $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, then $\mathbf{A}^{-1}\mathbf{v}_i = \lambda_i^{-1}\mathbf{v}_i$ (thereby the respective eigenvalues are being mutually reciprocal). We can also recall that there is a class of matrices, referred to as unitary matrices, for which $\mathbf{A}^{-1} = \mathbf{A}^*$. Therefore unitary matrices share eigenvectors with their transpose and, by the spectral theorem "for a Hermitian transpose pair" discussed in Subsection 7.9.3, unitary matrices admit unitary eigenbases.

We thereby arrive at

The spectral theorem for unitary matrices: unitary matrices admit unitary eigenbases.

Now we shall study the eigenvalues of unitary matrices. With Hermitian matrices we have arrived at the spectral theorem's corollary, relating the real spectra to hermiticity. Let's see if we can obtain similar results for unitary matrices.

Goal 7.15. Study the spectra of unitary matrices.

As the eigenvalues of unitary matrices and their transposes go, by the previous discussion we have $\tilde{\lambda}_i = \lambda_i^{-1}$, where $\tilde{\lambda}_i$ is the eigenvalue of $\mathbf{A}^*$ corresponding to the same eigenvector as the eigenvalue λ_i of $\mathbf{A}$. Indeed, given $\mathbf{A}\mathbf{v}_i = \lambda \mathbf{v}_i$, we have

$$\mathbf{A}^*\mathbf{v}_i = \mathbf{A}^{-1}\mathbf{v}_i = \lambda_i^{-1}\mathbf{v}_i = \tilde{\lambda}_i$$

Combining with (7.145) we have

$$\begin{cases} \tilde{\lambda}_i = \lambda_i^* \\ \tilde{\lambda}_i = 1/\lambda_i \end{cases} \quad \forall i$$

Thus $\lambda_i^* = 1/\lambda_i$, that is $\lambda_i \lambda_i^* = 1$, that is $|\lambda_i|^2 = 1$ or simply $|\lambda_i| = 1 \ \forall i$.

It follows that all eigenvalues of unitary matrices are complex numbers of unit magnitude:

(spectra of unitary matrices are* unitary*)

$$(\mathbf{A}^* = \mathbf{A}^{-1}) \implies (|\lambda_i| = 1 \ \forall i) \tag{7.148}$$

Now with Hermitian matrices also the converse statement was partially true, where a real spectrum would imply hermiticity under the additional condition that the matrix admits a unitary eigenbasis. Let's see if a similar converse statement holds with unitary matrices.

Goal 7.15a. Investigate the question of whether a matrix $\mathbf{A}$ which has a unitary spectrum and admits a unitary eigenbasis is unitary.

As with Hermitian matrices, we shall start with the diagonalized matrix $\mathbf{A}'$. A diagonalized matrix with a unitary spectrum is unitary, since $\lambda^* = \lambda^{-1}$ $\forall \lambda \colon |\lambda| = 1$. That is $(\mathbf{A}')^* = (\mathbf{A}')^{-1}$. Let's see if we can infer unitarity for the original matrix $\mathbf{A}$.

Since $\mathbf{A}$ admits a unitary eigenbasis, its diagonalizing similarity transformation $\mathbf{A}' = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ is unitary ($\mathbf{T}^{-1} = \mathbf{T}^*$) and can be represented as a (complex) congruence transformation $\mathbf{A}' = \mathbf{T}^*\mathbf{A}\mathbf{T}$. Therefore

(unitarity is preserved by unitary similarity transformations)

$$\begin{aligned}\mathbf{A}^{-1} &= \left(\mathbf{T}\mathbf{A}'\mathbf{T}^{-1}\right)^{-1} = \mathbf{T}(\mathbf{A}')^{-1}\mathbf{T}^{-1} = \mathbf{T}(\mathbf{A}')^*\mathbf{T}^* = (\mathbf{T}\mathbf{A}'\mathbf{T}^*)^* = (\mathbf{T}\mathbf{A}'\mathbf{T}^{-1})^* \\ &= \mathbf{A}^* \end{aligned} \tag{7.149}$$

▷ Differently from hermiticity/symmetry, unitarity is not preserved by arbitrary congruence transformations, since we can notice that in (7.149) we explicitly used the representation of the matrix transformation both as a similarity transformation and as a congruence transformation.

▷ Instead of the derivation (7.149) we could have realized that a unitary matrix $\mathbf{A}$ is one that, being considered as an operator, preserves the dot product $\mathbf{x}^*\mathbf{y} = \sum_i x_i^* y_i$. If an operator preserves the dot product in one basis, it preserves the dot product in any other basis. Respectively the matrix representations of that operator will be unitary, as long as the dot product maintains the same form $\mathbf{x}^*\mathbf{y} = \sum_i x_i^* y_i$. As we know, the dot product maintains that form across all unitary bases, and unitary bases are connected to each other via all possible kinds of unitary transformations. Thus, unitary transformations won't change the unitarity property of $\mathbf{A}$.

It follows by (7.149) that the matrix $\mathbf{A}$ (admitting a unitary eigenbasis) is unitary iff so is its diagonal form $\mathbf{A}'$, which in turn happens iff $\mathbf{A}'$ has a unitary spectrum, which happens iff $\mathbf{A}$ has a unitary spectrum (since the spectra of both matrices are identical). We thus arrive at

Corollary of the spectral theorem for unitary matrices: a matrix admitting a unitary eigenbasis is unitary iff its spectrum is unitary.

Goal 7.15a completed. ■

Goal 7.15 completed. ■

Orthogonal matrices

Orthogonal matrices, being a real-valued case of unitary matrices, exhibit the same property of admitting unitary eigenbases. Notice that, differently from the case of symmetric matrices, this time we cannot claim that orthogonal matrices admit (real) orthonormal eigenbases: since their eigenvalues would usually be essentially complex, so would be their eigenvectors. We therefore won't formulate a separate spectral theorem for orthogonal matrices, since it won't be essentially different from the one for unitary matrices.

Similarly, we could formulate the spectral theorem corollary for real orthogonal matrices, but, since their spectra are complex, we still need to require a unitary (rather than orthogonal) eigenbasis, therefore the corollary wouldn't be essentially different either.

7.9.5 Antisymmetric and anti-Hermitian matrices

Anti-Hermitian matrices

There is one more obvious class of matrices which share eigenvectors with their Hermitian transposes: *anti-Hermitian* matrices $\mathbf{A}^* = -\mathbf{A}$. Indeed, if $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ and $\mathbf{A}^* = -\mathbf{A}$, then $\mathbf{A}^*\mathbf{v}_i = -\mathbf{A}\mathbf{v}_i = -\lambda_i \mathbf{v}$. We thereby arrive at

The spectral theorem for anti-Hermitian matrices: anti-Hermitian matrices admit unitary eigenbases.

As with previous versions of spectral theorem, we shall now study the eigenvalues.

Goal 7.16. Study the spectra of anti-Hermitian matrices

Applying a unitary similarity transformation $\mathbf{A}' = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{T}^*\mathbf{A}\mathbf{T}$ to an anti-Hermitian matrix $\mathbf{A}$, we notice that

(anti-Hermiticity is preserved by unitary similarity transformations)

$$\begin{aligned}(\mathbf{A}')^* &= \left(\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\right)^* = (\mathbf{T}^*\mathbf{A}\mathbf{T})^* = \mathbf{T}^*\mathbf{A}^*\mathbf{T} = -\mathbf{T}^*\mathbf{A}\mathbf{T} = -\mathbf{T}^{-1}\mathbf{A}\mathbf{T} \\ &= -\mathbf{A}'\end{aligned} \tag{7.150}$$

Therefore $\mathbf{A}$ and $\mathbf{A}'$ are simultaneously anti-Hermitian. The eigenvalues of a diagonal anti-Hermitian matrix $\mathbf{A}'$ obviously satisfy $\lambda_i^* = -\lambda_i$, which means that all λ_i are imaginary. The matrix $\mathbf{A}$ has the same eigenvalues, therefore

(spectra of anti-Hermitian matrices are imaginary)

$$(\mathbf{A}^* = \mathbf{A}^{-1}) \implies (\operatorname{Re}\lambda_i = 0\ \forall i) \tag{7.151}$$

Using (7.151) and (7.150) we immediately obtain

Corollary of the spectral theorem for anti-Hermitian matrices: a matrix admitting a unitary eigenbasis is anti-Hermitian iff its spectrum is imaginary.

▷ From the derivation of (7.150) one can see that anti-Hermiticity is preserved by arbitrary congruence transformations.

Goal 7.16 completed. ■

Antisymmetric matrices

Antisymmetric matrices, being a real-valued case of unitary matrices, exhibit the same properties. However since their spectra are imaginary there is not much value of considering them separately from anti-Hermitian matrices.

7.9.6 Normal matrices

We have found several classes of matrices which admit unitary eigenbases, listed in Table 7.1.

Real class	**Complex class**	**Spectrum**
symmetric	Hermitian	real
orthogonal	unitary	unitary
antisymmetric	anti-Hermitian	imaginary

Table 7.1: Special classes of normal matrices.

This suggests that there should be a larger class of matrices, containing all classes from Table 7.1, where the eigenvalues are free to take any complex values. That is, we are simply looking for matrices which admit unitary eigenbases, regardless of their spectrum.

Introducing this as an explicit class of matrices makes it desirable to be able to tell such matrices from the rest in a simple manner. Having to find all of the eigenvectors and check for their mutual orthogonality is a bit too involved. In Subsection 7.9.3 we found that it is sufficient and necessary that the matrix shares eigenvectors with its Hermitian transpose, but this condition per se is not much better, if not worse: not only we have to find the eigenvectors of $\mathbf{A}$, but also those of $\mathbf{A}^*$. On the other hand, this condition was useful in establishing the spectral theorems for matrix classes in Table 7.1.[2] Therefore, let's see if we could use it as a foundation to derive a better equivalent condition.

Goal 7.17. Find a simple test condition telling whether $\mathbf{A}$ and $\mathbf{A}^*$ share eigenvectors. *Hint:* use Cauchy–Schwarz inequality to construct a simple condition for testing whether a vector is an eigenvector of a given matrix.

In order to test whether a particular vector $\mathbf{v}$ is an eigenvector of a matrix $\mathbf{A}$ we could use the standard eigenvector equation $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. However, in its usual form, this equation is rather designed to find an eigenvector for an already given specific value of λ, while we are currently having a kind of an opposite task: given a vector, we want to test whether it is an eigenvector. So, we would need to somehow figure out λ for a given $\mathbf{v}$, which is even not necessarily an eigenvector to begin with. In principle, one could imagine simply "dividing" both sides of the equation by $\mathbf{v}$, getting something like $\lambda = (\mathbf{A}\mathbf{v})/\mathbf{v}$. However, even though one could formally define such vector division (which can be performed only on collinear vectors), this operation would be rather unwieldy to deal with, especially for our current purposes. There is still a simple solution, though: since we don't actually need the value of λ, instead of comparing $\mathbf{A}\mathbf{v}$ and $\lambda\mathbf{v}$ for equality or trying to "divide" $\mathbf{A}\mathbf{v}$ by $\mathbf{v}$, we could simply check whether the vectors $\mathbf{A}\mathbf{v}$ and $\mathbf{v}$ are collinear. To express the latter in a form of equation, we can recall that, by the Cauchy–Schwarz inequality, a dot product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is maximized in the absolute magnitude whenever $\mathbf{x}$ and $\mathbf{y}$ are collinear. In this case we have

$$|\mathbf{x}^*\mathbf{y}| = |\mathbf{x}| \cdot |\mathbf{y}| \tag{7.152a}$$

[2] We didn't use this condition to establish the spectral theorems for Hermitian and symmetric matrices (since we used Hermitian and symmetric matrices as the initial study case), but technically we could.

where $|\mathbf{x}| = \sqrt{\mathbf{x}^*\mathbf{x}}$, $|\mathbf{y}| = \sqrt{\mathbf{y}^*\mathbf{y}}$. In a linear algebra context, square roots are usually a little bit unwieldy, and it might be better to rewrite (7.152a) as

$$|\mathbf{x}^*\mathbf{y}|^2 = |\mathbf{x}|^2 \cdot |\mathbf{y}|^2$$

that is

$$(\mathbf{x}^*\mathbf{y})^* \cdot (\mathbf{x}^*\mathbf{y}) = (\mathbf{x}^*\mathbf{x}) \cdot (\mathbf{y}^*\mathbf{y}) \tag{7.152b}$$

Let's apply (7.152b) to test whether $\mathbf{v}$ is an eigenvector of $\mathbf{Av}$, that is whether $\mathbf{v}$ and $\mathbf{Av}$ are collinear. Substituting $\mathbf{v}$ for $\mathbf{x}$ and $\mathbf{Av}$ for $\mathbf{y}$ yields

$$(\mathbf{v}^*\mathbf{Av})^* \cdot (\mathbf{v}^*\mathbf{Av}) = (\mathbf{v}^*\mathbf{v}) \cdot ((\mathbf{Av})^*\mathbf{Av})$$

that is

$$\mathbf{v}^*\mathbf{A}^*\mathbf{v} \cdot \mathbf{v}^*\mathbf{Av} = \mathbf{v}^*\mathbf{v} \cdot \mathbf{v}^*\mathbf{A}^*\mathbf{Av} \tag{7.153a}$$

We can use the same idea to test whether $\mathbf{v}$ is an eigenvector of $\mathbf{A}^*$ by formally applying (7.153a) to the matrix $\mathbf{A}^*$:

$$\mathbf{v}^*\mathbf{Av} \cdot \mathbf{v}^*\mathbf{A}^*\mathbf{v} = \mathbf{v}^*\mathbf{v} \cdot \mathbf{v}^*\mathbf{AA}^*\mathbf{v}$$

Since $\mathbf{v}^*\mathbf{Av}$ and $\mathbf{v}^*\mathbf{A}^*\mathbf{v}$ are scalars, we can swap their order obtaining

$$\mathbf{v}^*\mathbf{A}^*\mathbf{v} \cdot \mathbf{v}^*\mathbf{Av} = \mathbf{v}^*\mathbf{v} \cdot \mathbf{v}^*\mathbf{AA}^*\mathbf{v} \tag{7.153b}$$

So, (7.153a) is an equality which tests whether $\mathbf{v}$ is an eigenvector of $\mathbf{A}$, while (7.153b) tests whether $\mathbf{v}$ is an eigenvector of $\mathbf{A}^*$.

Since the left-hand sides of (7.153a) and (7.153b) are identical, we can equate their right-hand sides:

$$\mathbf{v}^*\mathbf{v} \cdot \mathbf{v}^*\mathbf{A}^*\mathbf{Av} = \mathbf{v}^*\mathbf{v} \cdot \mathbf{v}^*\mathbf{AA}^*\mathbf{v}$$

or, cancelling the common factor $\mathbf{v}^*\mathbf{v}$,

$$\mathbf{v}^*\mathbf{A}^*\mathbf{Av} = \mathbf{v}^*\mathbf{AA}^*\mathbf{v} \tag{7.154}$$

Now let's take a moment to realize what we have obtained with (7.154). Suppose $\mathbf{A}^*\mathbf{A} = \mathbf{AA}^*$. This will guarantee that the right-hand sides of (7.153a) and (7.153b) are equal, while their left-hand sides were already equal. This would mean that both tests (7.153a) and (7.153b) are simultaneously true or false. That is a vector $\mathbf{v}$ is either simultaneously an eigenvector of both $\mathbf{A}$ and $\mathbf{A}^*$ or it is simultaneously not an eigenvector of those matrices. Since $\mathbf{v}$ can be any vector, it follows that the matrices $\mathbf{A}$ and $\mathbf{A}^*$ share eigenvectors.

This means that the equality

(definition of a normal matrix)

$$\mathbf{A}^*\mathbf{A} = \mathbf{AA}^* \tag{7.155}$$

is a sufficient condition for $\mathbf{A}$ and $\mathbf{A}^*$ to share eigenvectors and respectively to admit a unitary eigenbasis. Is it necessary?

Suppose $\mathbf{A}$ and $\mathbf{A}^*$ share eigenvectors and therefore admit a common unitary eigenbasis. Let $\mathbf{T}$ be the transformation matrix to this eigenbasis, so that $\mathbf{A}' =$

$\mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ and $\mathbf{A}'^* = \mathbf{T}^{-1}\mathbf{A}'\mathbf{T}$ are both diagonal. Therefore $\mathbf{A}'\mathbf{A}'^* = \mathbf{A}'^*\mathbf{A}'$ and therefore

$$\begin{aligned}\mathbf{A}\mathbf{A}^* &= \mathbf{T}\mathbf{A}'\mathbf{T}^{-1} \cdot \mathbf{T}\mathbf{A}'^*\mathbf{T}^{-1} = \mathbf{T} \cdot \mathbf{A}'\mathbf{A}'^* \cdot \mathbf{T}^{-1} = \mathbf{T} \cdot \mathbf{A}'^*\mathbf{A}' \cdot \mathbf{T}^{-1} \\ &= \mathbf{T}\mathbf{A}'^*\mathbf{T}^{-1} \cdot \mathbf{T}\mathbf{A}'\mathbf{T}^{-1} = \mathbf{A}^*\mathbf{A}\end{aligned}$$

Thus (7.155) is a sufficient and necessary condition for $\mathbf{A}$ to admit a unitary eigenbasis. Matrices satisfying (7.155) are referred to as *normal matrices*. We thereby have obtained

The* spectral theorem *for normal matrices: a matrix is normal iff it admits a unitary eigenbasis.

Goal 7.17 completed. ■

7.9.7 Non-unitary bases

As we stated at the beginning of the discussion of the spectral theorem and related topics, we were assuming a unitary (or orthonormal) basis all along this discussion. Let's now see how we can generalize the results to arbitrary bases.

There has to be done a remark in this respect. We were talking all along of eigenvectors of a matrix as well as about matrix transformations under a change of basis. However a change of basis is something that one does to a tensor, e.g. to a bilinear form or a linear operator, not to a matrix. In that sense one doesn't diagonalize a matrix by a similarity transformation, what is actually happening is that a linear operator is being diagonalized by a basis change.

So when we were talking of an eigenbasis of a matrix, we actually were talking of an eigenbasis of a linear operator represented by that matrix. Respectively, our discussion of spectral theorems was actually discussing the spectral theorems of linear operators rather than matrices.

> ▷ Notably, diagonalization of quadratic forms, which we discussed in Subsection 7.9.1, looks a bit awkward from this perspective, since quadratic forms are bilinear forms rather than linear operators, therefore a change of basis should be done by a congruence rather than similarity transformation. However, recall that we were symmetrizing the quadratic forms before applying the diagonalizing basis changes. By the spectral theorem for symmetric matrices, the eigenbasis of a symmetric quadratic form is orthogonal, and therefore there is no difference between the respective similarity and congruence transformations.
>
> From a slightly different angle, we can recall that the diagonalizing transformation of the quadratic form's basis was done as a sequence of rotations. Therefore the total transformation is orthogonal and again, there is no difference between similarity and congruence transformations.

It is however still possible to reinterpret the same topics from a pure matrix perspective. In that case we should drop the idea of a basis and talk strictly about matrix transformations instead. Respectively, the assumption of a unitary basis won't apply at all, since there is no basis to begin with.

Goal 7.18. Reinterpret the previous discussion of spectral theorems in pure matrix terms.

The spectral theorems were essentially concerned with a matrix (or, rather, a linear operator) admitting a unitary eigenbasis, where we implicitly assumed that the eigenbasis doesn't have to coincide with the originally chosen basis of the space, and therefore there must be a transformation from one basis to the other. If the original basis is also unitary (which was our assumption), then the transformation from one basis to the other is a unitary transformation. Conversely, if the eigenbasis is not unitary, then the transformation from the unitary original basis to the non-unitary eigenbasis is not unitary.

We can therefore replace the statement about a linear operator admitting a unitary eigenbasis with a matrix being diagonalizable by a unitary similarity transformation:

The spectral theorem in matrix terms: a matrix is normal iff it is diagonalizable by a unitary transformation.

▷ Notice that in the above we for simplicity omitted the word "similarity". This is justified by a) matrix diagonalization being normally done by a similarity transformation anyway and b) in case of a unitary transformation there is no difference between similarity and congruence transformations.

The other spectral theorems and their corollaries are reformulated in a similar way.

Goal 7.18 completed. ■

7.9.8 Operator classes

As we found in Subsection 7.9.7, our discussion of spectral theorems originally was somewhat nonchalantly mixing up the ideas of matrices and $(1,1)$-type tensors (linear operators). In Subsection 7.9.7 we addressed that issue by reinterpreting our discussion in pure matrix terms. However, instead, we could go into the opposite direction and reinterpret it in purely operator terms. We would therefore need to introduce definitions of Hermitian, unitary, anti-Hermitian and normal operators. Those can be defined as simply those which are represented by Hermitian, unitary, anti-Hermitian and normal matrices in a unitary basis, but such definition would be not too elegant. Ideally we would want to have definitions abstract from any basis.

Goal 7.19. Construct basis-abstract definitions of Hermitian, unitary and anti-Hermitian operators.

Let's start with Hermitian operators. As an idea, consider the following. Lacking the definition of an operator counterpart of Hermitian conjugation (just yet), but wanting to become abstract from the basis, we can realize that the only entity that we're aware of which has a basis-independent component representation is a scalar. If we want to construct a scalar from an operator, we could use a bilinear form $A(\vec{x},\vec{y}) = \vec{x}\cdot\vec{A}(\vec{y})$, where $A(\vec{x},\vec{y})$ is the bilinear form and $\vec{A}(\vec{\cdot})$ is the corresponding operator.

Now a Hermitian matrix $\mathbf{A}$ corresponds to a conjugate-symmetric bilinear form $A(\vec{x},\vec{y}) = \mathbf{x}^*\mathbf{A}\mathbf{y} = (\mathbf{y}^*\mathbf{A}^*\mathbf{x})^* = (\mathbf{y}^*\mathbf{A}\mathbf{x})^* = \big(A(\vec{y},\vec{x})\big)^*$. In the operator terms, $A(\vec{x},\vec{y}) = \vec{x}\cdot\vec{A}(\vec{y})$ and $\big(A(\vec{y},\vec{x})\big)^* = \big(\vec{y}\cdot\vec{A}(\vec{x})\big)^* = \vec{A}(\vec{x})\cdot\vec{y}$, and we obtain

(definition of a Hermitian operator)

$$\vec{x} \cdot \vec{A}(\vec{y}) = \vec{A}(\vec{x}) \cdot \vec{y} \quad \forall \vec{x}, \vec{y} \tag{7.156a}$$

which can be more formally justified by rewriting it in the matrix form as $\mathbf{x}^* \cdot \mathbf{A}\mathbf{y} = (\mathbf{A}^*\mathbf{x})^* \cdot \mathbf{y} = (\mathbf{A}\mathbf{x})^* \cdot \mathbf{y}$.

For unitary matrices we can simply consider that they preserve dot products and so should the unitary operators:

(definition of a unitary operator)

$$\vec{A}(\vec{x}) \cdot \vec{A}(\vec{y}) = \vec{x} \cdot \vec{y} \quad \forall \vec{x}, \vec{y} \tag{7.156b}$$

which can be more formally justified by rewriting it in the matrix form as $(\mathbf{A}\mathbf{x})^* \cdot \mathbf{A}\mathbf{y} = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{y} = \mathbf{x}^*\mathbf{y}$.

Anti-Hermitian operators are obviously given by

(definition of an anti-Hermitian operator)

$$\vec{x} \cdot \vec{A}(\vec{y}) = -\vec{A}(\vec{x}) \cdot \vec{y} \quad \forall \vec{x}, \vec{y} \tag{7.156c}$$

▷ The definitions of different operator classes given here are obtained solely as basis-abstract expressions of the respective matrix classes and thereby, by default, are only applicable to linear operators defined on finite-dimensional linear spaces. These definitions can serve as a starting point for extending these ideas to a wider class of operators, but the details lie outside the scope of the book.

Attempting to define normal operators in a similar fashion, we have some difficulty. In particular, the bilinear form $\mathbf{x}^* \cdot \mathbf{A}^*\mathbf{A} \cdot \mathbf{y} = (\mathbf{A}\mathbf{x})^* \cdot \mathbf{A}\mathbf{y}$ can be easily rewritten in the operator terms as $\vec{A}(\vec{x}) \cdot \vec{A}(\vec{y})$, but how can we write a bilinear form built around the opposite-order matrix product $\mathbf{A}\mathbf{A}^*$ in terms of the same operator $\vec{A}(\vec{\cdot})$? Swapping both terms as $\vec{A}(\vec{y}) \cdot \vec{A}(\vec{x})$ yields the bilinear form $(\mathbf{A}\mathbf{y})^* \cdot \mathbf{A}\mathbf{x} = \mathbf{y}^* \cdot \mathbf{A}^*\mathbf{A} \cdot \mathbf{x}$ with the same matrix $\mathbf{A}^*\mathbf{A}$. The same happens if we try to conjugate the entire bilinear form expression: conjugation of $\mathbf{A}^*\mathbf{A}$ yields the same $\mathbf{A}^*\mathbf{A}$ again. It seems we can't get a hold of $\mathbf{A}\mathbf{A}^*$, at least not in terms of the operator $\vec{A}(\vec{\cdot})$.

On the other hand, it looks like, based on the construction of definitions (7.156), we could generalize the definition of Hermitian conjugation of matrices to operators.

Goal 7.19a. Construct the definition of the operator counterpart of Hermitian conjugation and use it to rewrite the definitions (7.156) in a way which is more resembling the definitions of the respective matrix classes. *Hint:* use the construction process of (7.156a).

Given two matrices $\mathbf{A}$ and $\mathbf{A}^*$, consider two bilinear forms built around these matrices:

$$A(\vec{x}, \vec{y}) = \mathbf{x}^*\mathbf{A}\mathbf{y} = \mathbf{x}^* \cdot \mathbf{A}\mathbf{y} = \vec{x} \cdot \vec{A}(\vec{y})$$
$$A^\dagger(\vec{x}, \vec{y}) = \mathbf{x}^*\mathbf{A}^*\mathbf{y} = \mathbf{x}^* \cdot \mathbf{A}^*\mathbf{y} = \vec{x} \cdot \vec{A}^\dagger(\vec{y})$$

where $\vec{A}(\cdot)$ and $\vec{A}^\dagger(\cdot)$ are the operators corresponding to these two bilinear forms. Noticing that

$$\big(A(\vec{x},\vec{y})\big)^* = (\mathbf{x}^*\mathbf{A}\mathbf{y})^* = \mathbf{y}^*\mathbf{A}^*\mathbf{x} = A^\dagger(\vec{y},\vec{x})$$

we have

$$\vec{x}\cdot\vec{A}(\vec{y}) = A(\vec{x},\vec{y}) = \big(A^\dagger(\vec{y},\vec{x})\big)^* = \big(\vec{y}\cdot\vec{A}^\dagger(\vec{x})\big)^* = \vec{A}^\dagger(\vec{x})\cdot\vec{y}$$

yielding

(definition of a Hermitian adjoint operator)

$$\vec{A}^\dagger(\vec{x})\cdot\vec{y} = \vec{x}\cdot\vec{A}(\vec{y}) \quad \forall\vec{x},\vec{y} \tag{7.157}$$

The operator $\vec{A}^\dagger(\cdot)$ is called the *Hermitian adjoint* of $\vec{A}(\cdot)$.

Using the definition (7.157) we can rewrite the definitions (7.156) simply as

(definitions of operator classes in terms of operator adjoint)

$$\vec{A}^\dagger = \vec{A} \qquad \text{(Hermitian)} \tag{7.158a}$$

$$\vec{A}^\dagger = \vec{A}^{-1} \qquad \text{(unitary)} \tag{7.158b}$$

$$\vec{A}^\dagger = -\vec{A} \qquad \text{(anti-Hermitian)} \tag{7.158c}$$

$$\vec{A}^\dagger\vec{A} = \vec{A}\vec{A}^\dagger \qquad \text{(normal)} \tag{7.158d}$$

Using (7.157) it is easy to demonstrate that the first 3 of the definitions (7.158) coincide with the definitions (7.156):

$$\begin{aligned}
\vec{A}^\dagger = \vec{A} \quad &\Longrightarrow \quad \vec{x}\cdot\vec{A}(\vec{y}) = \vec{A}^\dagger(\vec{x})\cdot y = \vec{A}(\vec{x})\cdot y \\
\vec{A}^\dagger = \vec{A}^{-1} \quad &\Longrightarrow \vec{A}(\vec{x})\cdot\vec{A}(\vec{y}) = \vec{A}^\dagger(\vec{A}(\vec{x}))\cdot y = \vec{A}^{-1}(\vec{A}(\vec{x}))\cdot y = \vec{x}\cdot y \\
\vec{A}^\dagger = -\vec{A} \quad &\Longrightarrow \quad \vec{x}\cdot\vec{A}(\vec{y}) = \vec{A}^\dagger(\vec{x})\cdot y = -\vec{A}(\vec{x})\cdot y
\end{aligned}$$

Goal 7.19a completed. ■

Goal 7.19 completed. ■

7.10 Rotations as exponentials

In Subsection 7.7.3 we have seen that an imaginary-argument exponential can be decomposed into a product of infinitely many infinitesimal rotations, each rotation being represented by a complex number $1 + i\varphi/n$, where φ is the total rotation angle and $n \to \infty$. Multiplication by this complex number converts the point $(1, 0)$ on the complex plane into the point $(1, \varphi/n)$, while the point $(0, 1)$ is converted into $(-\varphi/n, 1)$.

Instead of using complex numbers on the complex plane we can represent the same using vectors and matrices on a 2D plane. Specifically, we want a linear transformation $\mathbf{M}$ which transforms the vector $(1, 0)$ into the vector $(1, \varphi/n)$ and transforms the vector $(0, 1)$ into $(-\varphi/n, 1)$. As the columns of the matrix $\mathbf{M}$ should simply contain the images of the basis vectors, we have

$$\mathbf{M} = \begin{bmatrix} 1 & -\varphi/n \\ \varphi/n & 1 \end{bmatrix} \tag{7.159}$$

Now let $z \in \mathbb{C}$ be some complex number and consider the "rotated" complex number

$$z' = (1 + i\varphi/n)z = z + i\frac{\varphi}{n}z$$

We can see that multiplication by the real part of $1 + i\varphi/n$ simply contributes z to the multiplication result, thereby acting as identity transformation, while the multiplication by the imaginary part of $1+i\varphi/n$ contributes the component $i\frac{\varphi}{n} \cdot z$, which is orthogonal to z. Similarly, we can decompose the matrix $\mathbf{M}$ into the two parts:

$$\mathbf{M} = \mathbf{I} + \frac{\mathbf{A}}{n} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{1}{n}\begin{bmatrix} 0 & -\varphi \\ \varphi & 0 \end{bmatrix} \tag{7.160}$$

Then, given a vector $\mathbf{v}$,

$$\mathbf{M}\mathbf{v} = \left(\mathbf{I} + \frac{\mathbf{A}}{n}\right)\mathbf{v} = \mathbf{v} + \frac{\mathbf{A}}{n}\mathbf{v}$$

where, looking at the components of $\mathbf{A}$ in (7.160), we can realize that $\mathbf{A}\mathbf{v}$ is also orthogonal to $\mathbf{v}$.

Indeed, $\mathbf{A}$ converts the basis vector $(1, 0)$ into $(0, \varphi)$ and converts the basis vector $(0, 1)$ into $(-\varphi, 0)$, thereby both basis vectors are rotated by $90°$ counterclockwise (for $\varphi > 0$). Therefore any other vector is also rotated by $90°$ clockwise by the matrix $\mathbf{A}$, while its length is being multiplied by φ. Respectively, if we compute a matrix exponential of $\mathbf{A}$, this would be pretty much the same as computing $e^{i\varphi}$.

We would be interested to find other matrices, behaving in a way similar to $\mathbf{A}$.

Goal 7.20. Generalize the previous observations.

One essential requirement to the matrix $\mathbf{A}$ is that its exponentiation must yield a rotation matrix. The other requirement is that $\mathbf{A}$ must be real, so that infinitesimal transformations by multiplication by $\mathbf{I} + \mathbf{A}/n$ (at $n \to \infty$) would not leave the space of real vectors and therefore can be considered as infinitesimal rotations.

Goal 7.20a. Find a reasonably wide class of real matrices, which upon exponentiation produce rotation matrices. *Hint:* find a class of matrices which change the vector direction by $90°$ (and possibly change the vector length).

In order to generalize (7.160) to an arbitrary number of dimensions and arbitrary rotation directions we should replace the matrix $\mathbf{A}$ given in (7.160) by a general matrix $\mathbf{A}$ which converts vectors to some other vectors orthogonal to their preimages (where the change of the vector length by the transformation $\mathbf{A}$ can be anything, as long as the resulting length is not zero, since we are going to take the limit at $n \to \infty$ anyway). We have therefore the requirement of mutual orthogonality of $\mathbf{v}$ and $\mathbf{A}\mathbf{v}$ for all $\mathbf{v}$, that is

$$\mathbf{v}^{\mathrm{T}} \cdot \mathbf{A}\mathbf{v} = 0 \quad \forall \mathbf{v} \tag{7.161}$$

where we, as often, assume an orthonormal basis for convenience.

Now (7.161) describes a quadratic form which is always equal to zero. It might seem that this is achieved only at $\mathbf{A} = 0$, but actually it is not. Recall

that by (7.123) a quadratic form is unchanged by symmetrization of its matrix. Therefore the condition $\mathbf{A} = 0$ can be relaxed into

$$\frac{\mathbf{A} + \mathbf{A}^{\mathrm{T}}}{2} = 0$$

that is

$$\mathbf{A}^{\mathrm{T}} = -\mathbf{A} \tag{7.162}$$

That is $\mathbf{A}$ can be any antisymmetric real matrix.

Now we can simply do the same thing as in (7.75). That is we let $n \to \infty$ in $\mathbf{M} = \mathbf{I} + \mathbf{A}/n$, simultaneously raising $\mathbf{M}$ into the n-th power:

$$\mathbf{R} = \lim_{n\to\infty} \left(\mathbf{I} + \frac{\mathbf{A}}{n}\right)^n \tag{7.163a}$$

obtaining some rotation matrix $\mathbf{R}$. By (7.49)

$$\mathbf{R} = e^{\mathbf{A}} \tag{7.163b}$$

Thus, exponentials of antisymmetric real matrices are rotation matrices.

Goal 7.20a completed. ■

So, antisymmetric matrices seem to be similar to imaginary numbers in that, after exponentiation, both generate rotations. This is not too surprising: after all, antisymmetric numbers are the normal matrix counterparts of imaginary numbers (since their spectra are imaginary). Actually, computing the eigenvalues of the matrix $\mathbf{A}$ in (7.160) we obtain the characteristic polynomial $\lambda^2 + \varphi^2$, so the eigenvalues are $\lambda_{1,2} = \pm i\varphi$, which is pretty much the same imaginary number that we were exponentiating in the complex number prototype of $e^{\mathbf{A}}$.

We have found that antisymmetric matrices generate rotation matrices upon exponentiation, but we are not sure if that's the only such class of matrices. That is are there any other matrices also generating rotations upon exponentiation.

Goal 7.20b. Find out whether any rotation matrix is an exponential of some antisymmetric real matrix. *Hint:* start with unitary matrices and then narrow the obtained results to real rotation matrices.

We don't have a handy simple formula which allows to tell rotation matrices from all other matrices. However we know that rotation matrices are a subset of orthogonal matrices. Therefore we can start by considering a wider class of orthogonal matrices. Actually, let's start with an even wider class of unitary matrices.

So let $\mathbf{U}$ be a unitary matrix. Since matrix exponentiation is a basis-independent operation, we could consider it in the unitary eigenbasis of $\mathbf{U}$, where $\mathbf{U}$ would be a diagonal matrix of unit-magnitude values, which are the eigenvalues of $\mathbf{U}$. These diagonal unit-magnitude values can be represented as exponentials of some imaginary values (equal to the logarithms of the eigenvalues of $\mathbf{U}$). Thereby $\mathbf{U} = e^{\mathbf{A}}$, where $\mathbf{A}$ is a diagonal matrix of the mentioned imaginary values. Apparently $\mathbf{A}$ is anti-Hermitian in its diagonal form and therefore is also anti-Hermitian in the original basis of $\mathbf{U}$ (where the equality $\mathbf{U} = e^{\mathbf{A}}$ still holds). It follows that

(unitary matrices are exponentials of anti-Hermitian matrices)

$$\mathbf{U} = e^{\mathbf{A}} \implies \left(\mathbf{U}^* = \mathbf{U}^{-1} \iff \mathbf{A}^* = -\mathbf{A}\right)$$

Now let's restrict $\mathbf{U}$ to be purely real, thus it is now not just unitary, but also an orthogonal matrix. In its diagonalized form, however, $\mathbf{U}$ will most likely contain complex components, and so will the diagonalizing transformation matrix $\mathbf{T}$. Therefore, we still obtain an anti-Hermitian matrix $\mathbf{A}$ in the eigenbasis of $\mathbf{U}$, but we cannot automatically assume that $\mathbf{A}$ becomes purely real (and hence antisymmetric) in the original basis of $\mathbf{U}$.

Nevertheless, the restriction of $\mathbf{U}$ being purely real still results in certain restrictions on $\mathbf{A}$. Specifically, the characteristic polynomial of $\mathbf{U}$ (written in the original basis) has purely real coefficients. Therefore all eigenvalues of $\mathbf{U}$ are either purely real or come in complex-conjugate pairs. Notice that, since $\mathbf{U}$ is unitary, its eigenvalues all have unit magnitude, therefore purely real eigenvalues of $\mathbf{U}$ are equal to 1 or -1. Respectively, the eigenvalues of $\mathbf{A}$, being obtained by taking the logarithm of the eigenvalues of $\mathbf{U}$, are either coming in complex-conjugate pairs of imaginary numbers or they are equal to 0 or πi.

▷ More precisely, in the latter statement the equality of eigenvalues of $\mathbf{A}$ to any number is to be understood modulo $2\pi i$, since the complex logarithm is defined only modulo $2\pi i$. E.g. two values λ_1 and λ_2 would be considered mutually complex-conjugate if $(\lambda_1 - \lambda_2^*) = 2\pi i k$ for some $k \in \mathbf{N}$. Similarly, we would consider $\lambda = 2\pi k i$ to be equal to 0, and $\lambda = (2k+1)\pi i$ to be equal to πi.

This has to be understood in the sense that there is a "modulo $2\pi i$ freedom" in the computation of the eigenvalues of $\mathbf{A}$. We can choose any of the values within the correct equivalence class.

Let's first take a look at the complex-conjugate eigenvalue pairs. Suppose λ and λ^* is a complex-conjugate pair of eigenvalues of $\mathbf{U}$. The respective eigenvectors (in the original basis) will be $\mathbf{v}$ and $\overline{\mathbf{v}}$, where $\overline{\mathbf{v}}$ denotes the vector whose components are complex conjugates of the components of $\mathbf{v}$. Indeed, since $\mathbf{U}$ is purely real, $\mathbf{U} \cdot \overline{\mathbf{v}} = \overline{\mathbf{U}\mathbf{v}} = \overline{\lambda \mathbf{v}} = \lambda^* \cdot \overline{\mathbf{v}}$. Switching to the eigenbasis of $\mathbf{U}$, we obtain the submatrix of the form $\mathbf{Q} = \left[\begin{smallmatrix} \lambda & 0 \\ 0 & \lambda^* \end{smallmatrix}\right]$ in the respective 2 dimensions of the entire space. This submatrix is an exponential of the 2×2 matrix $\mathbf{\Lambda} = \left[\begin{smallmatrix} \log\lambda & 0 \\ 0 & \log(\lambda^*) \end{smallmatrix}\right]$, or, since $|\lambda| = 1$, $\mathbf{\Lambda} = i \cdot \left[\begin{smallmatrix} \arg\lambda & 0 \\ 0 & -\arg\lambda \end{smallmatrix}\right]$, or simply $\mathbf{\Lambda} = i \arg\lambda \cdot \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ (where for the sake of simplicity we so far ignored the "modulo $2\pi i$ rule"). Now, instead of the basis vectors $\mathbf{v}$ and $\overline{\mathbf{v}}$, let's switch to the "real basis" consisting of $(\mathbf{v} + \overline{\mathbf{v}})/2$ and $(\mathbf{v} - \overline{\mathbf{v}})/2i$, that is the components of the new basis vectors are respectively the real and the imaginary parts of the components of $\mathbf{v}$ and $\overline{\mathbf{v}}$. The components of these basis vectors are thereby purely real and thereby so is the transformation from the original basis of $\mathbf{U}$ to the "real basis".

The transformation matrix from the eigenbasis of $\mathbf{U}$ to the "real basis" basis and the inverse transformation matrix are thereby

$$\mathbf{T} = \frac{1}{2} \cdot \begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix} \qquad \mathbf{T}^{-1} = \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$$

and

$$\mathbf{\Lambda}' = \mathbf{T}^{-1}\mathbf{\Lambda}\mathbf{T} = \frac{i \arg\lambda}{2} \cdot \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix}$$

$$
\begin{aligned}
&= \frac{i \arg \lambda}{2} \cdot \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix} \cdot \begin{bmatrix} 1 & -i \\ -1 & -i \end{bmatrix} = \frac{i \arg \lambda}{2} \cdot \begin{bmatrix} 0 & -2i \\ 2i & 0 \end{bmatrix} \\
&= \arg \lambda \cdot \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \arg \lambda \\ -\arg \lambda & 0 \end{bmatrix}
\end{aligned} \tag{7.164}
$$

We therefore obtain that $\mathbf{\Lambda}'$ is an antisymmetric real matrix in the "real basis" consisting of $(\mathbf{v} + \overline{\mathbf{v}})/2$ and $(\mathbf{v} - \overline{\mathbf{v}})/2i$. Notably, the transformation $\mathbf{T}$ is not exactly unitary, but it can be made unitary by redistributing the normalization between $\mathbf{T}$ and $\mathbf{T}^{-1}$ as

$$
\mathbf{T} = \frac{1}{\sqrt{2}} \cdot \begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix} \qquad \mathbf{T}^{-1} = \frac{1}{\sqrt{2}} \cdot \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}
$$

This doesn't change the result of the transformation $\mathbf{\Lambda}' = \mathbf{T}^{-1}\mathbf{\Lambda}\mathbf{T}$.

▷ To properly account for the "modulo $2\pi i$ rule", we need to add a term $2\pi i k$ (where $k \in \mathbb{Z}$) to either of the diagonal elements of $\mathbf{\Lambda}$. This however immediately turns the trace of $\mathbf{\Lambda}$ from zero to a nonzero imaginary number. Since the trace is invariant under similarity transformations, $\mathbf{\Lambda}$ becomes essentially complex in any basis.

The same happens if we add $2\pi i k_1$ to the first diagonal element and $-2\pi i k_2$ to the second diagonal element, unless $k_1 = k_2$.

Thus, in the 2D real "eigenplane" corresponding to the conjugate pair λ and λ^*, the respective submatrix of the matrix $\mathbf{U}$ is obtained by exponentiation of an antisymmetric matrix $\mathbf{\Lambda}'$. The same argument can be made for all complex-conjugate eigenvalue pairs, effectively building a submatrix of the matrix $\mathbf{A}$ (that we're trying to construct) in some basis, related to our original basis via an orthogonal transformation. It remains to complete the matrix $\mathbf{A}$ by adding all dimensions corresponding to the real eigenvalues.

Actually, we can tackle the largest part of the real eigenvalues by formally arranging them into complex-conjugate pairs. Indeed, we mentioned that the real eigenvalues of $\mathbf{U}$ must be all equal to 1 or -1. Now any pair of eigenvalues equal to 1 can be formally seen as a complex-conjugate pair, and so can be any pair of eigenvalues equal to -1. So, let's "eliminate" all such pairs from the set of real eigenvalues, by treating them as complex-conjugate pairs. This leaves at most one real eigenvalue which is equal to 1 or -1, or at most one pair of eigenvalues equal to 1 and -1 respectively, all other eigenvalues being organized into complex-conjugate pairs.

If no eigenvalues are being left, then we have constructed a real antisymmetric matrix $\mathbf{A}$ consisting (in some real orthonormal basis) of subblocks of the form (7.164). E.g. in 4 dimensions we could obtain something like

$$
\mathbf{A} = \left[\begin{array}{cc|cc} 0 & \arg \lambda_1 & 0 & 0 \\ -\arg \lambda_1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \arg \lambda_2 \\ 0 & 0 & -\arg \lambda_2 & 0 \end{array}\right] \tag{7.165a}
$$

If we are being left with a single eigenvalue equal to 1, this eigenvalue is an exponential of 0 (all other "modulo $2\pi i$ equivalents" $2\pi i k$ are purely imaginary and therefore do not belong to our purely real matrix $\mathbf{A}$). We therefore are having an exponential of a 1×1 submatrix, the component of this submatrix

being equal to 0. This matrix is formally antisymmetric, and therefore so is the entire matrix $\mathbf{A}$. We have thereby constructed a real antisymmetric $\mathbf{A}$ such that $e^{\mathbf{A}} = \mathbf{U}$. E.g. in 3 dimensions we could have something like

$$\mathbf{A} = \left[\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & 0 & \arg\lambda_1 \\ 0 & -\arg\lambda_1 & 0 \end{array}\right] \tag{7.165b}$$

On the other hand if we are left with a single eigenvalue equal to -1, this eigenvalue is an exponential of πi (or, accounting for the "modulo $2\pi i$ equivalents", $(2k+1)\pi i$), so we're having an exponential of a 1×1 matrix with its only component being a nonzero imaginary number. E.g. in 3 dimensions we could have something like

$$\mathbf{A} = \left[\begin{array}{c|cc} \pi i & 0 & 0 \\ \hline 0 & 0 & \arg\lambda_1 \\ 0 & -\arg\lambda_1 & 0 \end{array}\right] \tag{7.165c}$$

This matrix is anti-Hermitian but not antisymmetric. It also cannot be brought to a real antisymmetric form by any unitary transformation of basis, since its trace is a nonzero imaginary number.

Similarly, if we are left with a pair of eigenvalues 1 and -1, upon taking the logarithm we obtain $2\pi k i$ and $(2l+1)\pi i$ respectively. E.g.

$$\mathbf{A} = \left[\begin{array}{cc|cc} 2\pi k i & 0 & 0 & 0 \\ 0 & (2l+1)\pi i & 0 & 0 \\ \hline 0 & 0 & 0 & \arg\lambda_1 \\ 0 & 0 & -\arg\lambda_1 & 0 \end{array}\right] \tag{7.165d}$$

The trace of this matrix is $(2k+2l+1)\pi i$, which is again a nonzero imaginary number, therefore the matrix cannot be made real antisymmetric.

We can notice that in the situations like (7.165a) and (7.165b) we had $\det \mathbf{U} = 1$. Indeed, in (7.165a) we had all eigenvalues of $\mathbf{U}$ organized in complex-conjugate pairs, and, since the eigenvalues all had unit magnitude, their product is 1. In (7.165b) on top of that we had to additionally multiply this product by the "standalone" real eigenvalue equal to 1, which leaves the entire product equal to 1. On the other hand, in (7.165c) and (7.165d) we additionally had to multiply by -1 or by $1 \cdot (-1)$ respectively, which gives $\det \mathbf{U} = -1$. It turns out that only orthogonal matrices with unit determinants can be represented as exponentials of antisymmetric matrices. But wait, those are exactly the rotation matrices! Therefore

(rotation matrices are exponentials of antisymmetric matrices)

$$\mathbf{R} = e^{\mathbf{A}} \implies \Big(\big(\mathbf{R}^{\mathrm{T}} = \mathbf{R}^{-1} \ \wedge\ \det \mathbf{R} = 1\big) \iff \mathbf{A}^{\mathrm{T}} = -\mathbf{A}\Big)$$

where all matrices are real.

Goal 7.20b completed. ■

Goal 7.20 completed. ■

Summary of Chapter 7

- We have used differential equations as the motivation to introduce the exponential function as well as real powers of real values.
- We have constructed solutions of differential equations as limiting cases of a finite-step scheme, effectively obtaining the exponential function e^x as the main element of such solutions. In case of a differential equation system the same idea was generalized to the idea of a matrix exponential.
- We have use matrix exponentials as the motivation for matrix diagonalization. We then have studied the Jordan normal form as a "diagonalization fallback" emerging as a limiting case of diagonalizable matrices.
- We "discovered" the spectral theorems by analyzing the graphical behavior of the level sets of quadratic forms.
- We have obtained the motivation for the introduction of complex numbers from solving differential equations by diagonalization.
- We have found that the rotational behavior of imaginary exponents comes from the respective rotational behavior of multiplication by $1 + i\xi$ where ξ is a small real number. We also found that, there is a matrix counterpart to that, represented by rotation matrices being exponentials of real antisymmetric matrices.

Chapter 8

Superpositions of sinusoids

Four types of Fourier transform: discrete Fourier series and the other three types as its limiting cases. Normalization placement. Transform domains and their dual symmetry. Transform properties. Kronecker and Dirac delta functions. Convergence. Dual domain effects of discontinuities and features. Energy and its momenta. Reciprocal duality of transform domains. Sampling and STFT as a means to finite computability. Application to random processes.

8.1 Discrete Fourier series . . . 381
8.2 Discrete Fourier transform . . . 385
8.3 Domain semantics . . . 388
8.4 Duality . . . 391
8.4.1 Symmetric version . . . 391
8.4.2 Amplitude/density duality . . . 393
8.4.3 Application principles . . . 394
8.4.4 Application example: the constant term . . . 396
8.5 Real discrete Fourier transform . . . 398
8.6 Symmetries . . . 402
8.7 Discrete-time Fourier transform . . . 404
8.8 Fourier series . . . 407
8.9 Fourier transform . . . 412
8.10 Unitarity . . . 417
8.11 Convolution . . . 423
8.12 Shift and modulation . . . 430
8.13 Delta function . . . 432
8.13.1 Nonperiodic Kronecker delta . . . 432
8.13.2 Periodic Kronecker delta . . . 435
8.13.3 DFT roundtrip . . . 437
8.13.4 DTFT roundtrip . . . 438
8.13.5 Periodic Dirac delta . . . 439
8.13.6 Fourier series roundtrip . . . 446
8.13.7 Nonperiodic Dirac delta . . . 447
8.13.8 Fourier transform roundtrip . . . 453
8.13.9 Convergence as filtering . . . 454
8.13.10 Continuous representation of discrete functions . . . 456
8.13.11 Normalization checks . . . 457
8.14 Correlation . . . 459

8.15 Differentiation and integration 461
8.16 Discontinuities 464
8.16.1 Sawtooth function 464
8.16.2 Periodic piecewise parabolic function 467
8.16.3 Periodic piecewise polynomial functions 468
8.16.4 Periodic arbitrary piecewise functions 471
8.16.5 Discontinuities in nonperiodic functions 473
8.16.6 Integral sine function 475
8.16.7 Compactly supported spectra 478
8.16.8 Spectral discontinuities 479
8.16.9 Dual-domain effects of features 481
8.17 Domain scaling 482
8.18 Power and energy 484
8.18.1 Power and energy in two domains 484
8.18.2 Energy center 485
8.18.3 Energy concentration 489
8.18.4 Energy concentration bound 490
8.18.5 Maximally concentrated energy 494
8.18.6 Energy center shift 498
8.19 Finite computability 503
8.19.1 Sampling 503
8.19.2 Domain truncation 508
8.19.3 Windowing 510
8.19.4 Window width and height 512
8.19.5 Spectral-domain sampling 514
8.19.6 Localization of frequency 516
8.19.7 Discrete-domain windowing 517
8.20 Stochastic Fourier transform 519
8.20.1 Discrete Fourier transform 519
8.20.2 Discrete-time Fourier transform 523
8.20.3 Fourier transform 528
8.20.4 Distribution of spectral values 530
Summary of Chapter 8 533

The treatment of Fourier theory topics (which in math texts usually cover continuous-argument Fourier series and Fourier transform) varies from one text to the other. Some texts may use real sine and cosine functions, some may employ complex sinusoids instead. The convergence conditions of Fourier series and Fourier transform are really complicated, which respectively makes a rigorous treatment of the topic really complicated as well, going into the depths of calculus and advanced linear algebra and also often avoiding certain areas altogether anyway. With the requirement of the rigor being noticeably relaxed in this book, we shall take a (hopefully) much more accessible and intuitive approach to the topic.

The central idea of Fourier theory is to represent a function as a linear combination of (complex or real) sinusoids, e.g.:

$$x(t) = \sum_k a_k e^{i\omega_k t} \tag{8.1}$$

▷ We could have called this function $f(x)$ instead of $x(t)$, but in this chapter we would really like to reserve the letter f for frequency. Simultaneously and similarly to the case of differential equations, we would like to suggest time semantics for t in (8.1), at least as one possible interpretation. Using letters x and y to denote functions of time is common in signal processing, so it's not that we are using a highly non-standard notation. On the other hand, in signal processing it is also common to denote the imaginary unit as j rather than i, but we would like not to deviate too far from the generic math into the signal processing area, so let's keep i as the imaginary unit and respectively use j as an index letter.

Complex sinusoids, being a particular case of exponentials, are eigenfunctions of the differentiation operator:

$$\frac{\mathrm{d}}{\mathrm{d}t}e^{i\omega t} = i\omega\frac{\mathrm{d}}{\mathrm{d}t}e^{i\omega t}$$

and this property can be utilized to convert differential equations to algebraic ones (by simply cancelling the common eigenfunction multiplier). Historically, this was the basis for the motivation to introduce Fourier theory in the first place, just as an aid in solving differential equations, although just being able to see a function as a linear combination of sinusoids is already valuable per se, e.g. because it is closely related to how humans perceive light and sound.

Motivational starter for Chapter 8

Develop an approach/framework for representing a given function by a superposition (linear combination) of sinusoids.

8.1 Discrete Fourier series

In this book we are going to approach the topic of Fourier transform by first looking at the discrete argument case. This will allow us to utilize the basic linear algebra framework, instead of requiring some totally advanced linear algebra areas or not being able to use linear algebra at all as an aid in this context otherwise. We are also going to use complex sinusoids, which should result in a much simpler presentation of the topic compared to using real sinusoids. One could even say that using real sinusoids in this context would simply present unnecessary difficulties, and we would rather avoid those.

Thus, suppose we are given an (infinitely) periodic function $x[n]$ of a discrete integer argument n.

▷ Usually in the context of mathematics in this case we would rather say that we are given an infinitely periodic sequence $\{x_n\}$. Since we want $n \in \mathbb{Z}$ rather than $n \in \mathbb{N}$, the sequence $\{x_n\}$ is also a *bilateral* one, extending from $-\infty$ to $+\infty$. However, since our goal is to promote later to functions of a continuous argument, we would like to make the parallels between continuous and discrete argument more apparent, so we would rather talk of functions of a discrete argument. To highlight the function aspect we are going to notate the argument function-style, except that, to also highlight the discreteness of the argument, we are going to use square rather than round brackets, thus writing $x[n]$ instead of $x(n)$.

▷ We are also going to look at the values of $x[\cdot]$ evaluate at specific values of the function's argument as vector components. In this case we will denote $x[\cdot]$ as a column matrix $\mathbf{x}$ and utilize the index notation x_n for the components of $\mathbf{x}$.

Let N be the period of this function: $x[n+N] = x[n]\ \forall n$. We wish to represent the function $x[\cdot]$ as a linear combination of complex sinusoids.

Goal 8.1. Represent $x[\cdot]$ as a linear combination of complex sinusoids. *Hint:* the problem can be reduced to solving a system of N linear equations.

The sum (8.1) in case of discrete argument turns into

$$x[n] = \sum_k a_k e^{i\omega_k n} \tag{8.2}$$

where i is the imaginary unit, a_k are the linear combination coefficients and ω_k are the coefficients affecting the periods of the respective sinusoids (in fact, ω is the only parameter which we have available to ensure there is a difference between different terms of the linear combination). We assume $x[n], a_k \in \mathbb{C}$, $\omega_k \in \mathbb{R}$ and $n, k \in \mathbb{Z}$, although we're not sure yet what the summation range over k is.

▷ In equations (8.1) and (8.2), from the signal processing perspective, t is time, n is discrete time, a_k are the (complex) amplitudes of the sinusoidal partials, and ω_k are the cyclic frequencies.

Now, since $x[n]$ is N-periodic, it might be logical to try to utilize sinusoids which are also N-periodic. In such case, having ensured the equality (8.2) on just one function period, we thereby would ensure that it holds everywhere. Note that there is more than one (actually infinitely many) N-periodic sinusoids. Besides the trivial choice $\omega = 0$, which yields $e^{i\omega n} \equiv 1$, the next simplest choice is $e^{2\pi i n/N}$, that is $\omega = 2\pi/N$. However any integer multiple of that value of ω will also do. Indeed, if we take $k\omega$ instead of ω (where $k \in \mathbb{Z}$, rather than just $k \in \mathbb{N}$), the resulting sinusoid $e^{2\pi i k n/N}$ will be N/k-periodic (where N/k doesn't have to be an integer per se!) and thereby also N-periodic.

So, we want to represent our function $x[n]$ as a linear combination of sinusoids $e^{2\pi i k n/N}$. Let's introduce some notation for those as functions of n, e.g.

$$h_k[n] = e^{2\pi i k n/N} \tag{8.3}$$

Thereby (8.2) turns into

$$x[n] = \sum_k a_k h_k[n] \tag{8.4}$$

The equation set (8.4) is a system of linear equations with respect to the unknowns a_k. The number of unknowns as well as the number of equations are yet undecided, however we shall recall what we have said earlier: it must be sufficient to ensure the equalities just on one period, e.g. from 0 to $N-1$, the equality at all other values of n will be ensured automatically by the N-periodicity of the involved functions $x[n]$ and $h_k[n]$. Thus, we can restrict n in (8.4) to the range 0 to $N-1$, thereby we have N different equations in (8.4).

Normally, a system of N linear equations should have N unknowns, therefore we should also limit k to taking N different values, resulting in N different unknowns a_k and N different functions $h_k[n]$. However, which specific N values

of k shall we leave? To answer the latter question we can make one more observation: n and k are participating in the expression (8.3) for $h_k[n]$ on a 100% equal footing:

$$h_k[n] = h_n[k] \tag{8.5}$$

Therefore, since $h_k[n]$ are (by construction) N-periodic with respect to n:

$$h_k[n+N] = h_k[n] \tag{8.6a}$$

they also should be N-periodic with respect to k:

$$h_{k+N}[n] = h_k[n] \tag{8.6b}$$

and therefore there are only N different functions $h_k[n]$ anyway!

> ▷ In case of continuous argument the functions $h_k(t) = e^{2\pi ikft}$ would be of course all different. However, in case of discrete argument, it just so happens that the values of each given function h_k, taken at the specific discrete subset of points where they are evaluated, begin to repeat after k goes through the "first N values". They would still produce different values at other points, but these other points are not a part of the discrete domain of h_k. In signal processing this is known as *aliasing* and is the reason for the so-called Nyquist limit on the bandwidth of discrete-time signal media.

Therefore, we shall take both n and k going from 0 to $N-1$ in (8.4), thereby obtaining N equations with respect to the N unknowns a_k:

$$x[n] = \sum_{k=0}^{N-1} a_k h_k[n] \qquad n = 0, 1, \ldots, N-1 \tag{8.7}$$

Let's rewrite (8.7) in matrix notation:

$$\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} = \begin{bmatrix} h_0[0] & h_1[0] & \cdots & h_{N-1}[0] \\ h_0[1] & h_1[1] & \cdots & h_{N-1}[1] \\ \vdots & \vdots & \ddots & \vdots \\ h_0[N-1] & h_1[N-1] & \cdots & h_{N-1}[N-1] \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{N-1} \end{bmatrix}$$

or, concisely

$$\mathbf{x} = \mathbf{H} \cdot \mathbf{a} \tag{8.8}$$

The solution of (8.8) is simply $\mathbf{a} = \mathbf{H}^{-1} \cdot \mathbf{x}$. That is, provided $\mathbf{H}$ is not singular. Therefore, we need to study the matrix $\mathbf{H}$.

A straightforward way to check the singularity of a matrix is to compute its determinant. However it looks like computing the determinant of $\mathbf{H}$ might be a bit involved. Maybe there is a simpler way? Fortunately we can notice that the rows of $\mathbf{H}$ are linearly independent. Indeed, taking a sesquilinear (since we're dealing with complex numbers) dot product of j-th and k-th columns yields

$$\sigma_{jk} = \sum_{n=0}^{N-1} h_j^*[n] h_k[n] = \sum_{n=0}^{N-1} e^{-2\pi ijn/N} e^{2\pi ikn/N} = \sum_{n=0}^{N-1} e^{2\pi i(k-j)n/N} = \sum_{n=0}^{N-1} q^n \tag{8.9}$$

where

$$q = e^{2\pi i(k-j)/N}$$

Now if $j = k$, we have $q = e^0 = 1$ and

$$\sigma_{jj} = \sum_{n=1}^{N} e^0 = N$$

(more precisely, this occurs if $k - j$ is a multiple of N, but since k and j are both limited to the range 0 to $N - 1$, it's only possible for $j = k$). Otherwise $q \neq 1$ and we can use the formula for the sum of geometric series, obtaining

$$\sigma_{jk} = \sum_{n=0}^{N-1} q^n = \frac{q^N - 1}{q - 1} = \frac{e^{2\pi i(k-j)} - 1}{q - 1} = \frac{1 - 1}{q - 1} = 0$$

Thus

$$\sigma_{jk} = N\delta_{jk} \tag{8.10}$$

where δ_{jk} is the Kronecker delta symbol introduced in (4.32).

▷ Instead of using the geometric series formula we could just use the considerations of symmetry to obtain the result $\sigma_{jk} = 0$ for $j \neq k$. Indeed, the terms of the sum (8.9) are equally spread on a unit circle on the complex plane. There are N such terms, so the picture is rotationally symmetric with respect to rotations by a $2\pi/N$ angle or by its multiple. However, any nonzero value of the sum (8.9) is not symmetric with respect to such rotation (for $N > 1$) and thus cannot be the value of the sum (8.9).

The result (8.10) means that $\mathbf{H} \cdot \mathbf{H}^* = N$ (where N stands for N multiplied by the identity matrix and $\mathbf{H}^*$ denotes the Hermitian transpose of the matrix $\mathbf{H}$).[1] It follows that $\mathbf{H}^* \cdot \mathbf{H} = N$ as well and we obtain

$$\mathbf{H} \cdot \mathbf{H}^* = \mathbf{H}^* \cdot \mathbf{H} = N \tag{8.11}$$

Therefore $\mathbf{H}$ is "almost unitary". The matrix $\mathbf{H}$ itself is not exactly unitary, but the matrix $\mathbf{H}/\sqrt{N}$ (where N is just a number) is:

$$\left(\frac{\mathbf{H}}{\sqrt{N}}\right) \cdot \left(\frac{\mathbf{H}}{\sqrt{N}}\right)^* = \left(\frac{\mathbf{H}}{\sqrt{N}}\right)^* \cdot \left(\frac{\mathbf{H}}{\sqrt{N}}\right) = 1$$

Rewriting (8.11) as

$$\mathbf{H} \cdot \left(\frac{\mathbf{H}}{N}\right)^* = \left(\frac{\mathbf{H}}{N}\right)^* \cdot \mathbf{H} = 1$$

we conclude that

$$\mathbf{H}^{-1} = \mathbf{H}^*/N \tag{8.12}$$

and thus the solution of (8.8) is

$$\mathbf{a} = \mathbf{H}^{-1} \cdot \mathbf{x} = \frac{1}{N} \cdot \mathbf{H}^* \cdot \mathbf{x} \tag{8.13}$$

Rewriting (8.8) explicitly instead of the concise matrix form yields

[1] Notice that by (8.5), $\mathbf{H}^{\mathrm{T}} = \mathbf{H} \neq \mathbf{H}^*$.

(discrete Fourier series)

$$x[n] = \sum_{k=0}^{N-1} a_k e^{2\pi ikn/N} \tag{8.14a}$$

whereas (8.13) in explicitly written out form gives

(discrete Fourier series coefficient formula)

$$a_k = \frac{1}{N} \sum_{n=0}^{N-1} x[n] e^{-2\pi ikn/N} \tag{8.14b}$$

(where $e^{-2\pi ikn/N}$ are the elements of $\mathbf{H}^*$). The representation of the function $x[n]$ in the form (8.14a) with the coefficients a_k defined by (8.14b) is known as *discrete Fourier series*, or, shortly, *DFS*. This is the discrete-argument version of the commonly known Fourier series (where the argument is continuous).

Goal 8.1 completed. ■

8.2 Discrete Fourier transform

Let's rewrite (8.8) and (8.13) together

$$\mathbf{a} = \frac{\mathbf{H}^*}{N} \cdot \mathbf{x} \tag{8.15a}$$

$$\mathbf{x} = \mathbf{H} \cdot \mathbf{a} \tag{8.15b}$$

(where we also swapped the relative order of (8.8) and (8.13), the reason for doing so is about to become clear next). We could see (8.15) as a pair of mutually inverse linear transformations. Given the vector $\mathbf{x}$ we employ (8.15a) to obtain some other vector $\mathbf{a}$. Then we could employ (8.15b) to obtain the original vector $\mathbf{x}$ from our vector $\mathbf{a}$ back.

The transformation pair (8.15) would look even nicer in some sense, if both the forward and inverse transformations were truly unitary, which we can still achieve by dividing the matrix $\mathbf{H}$ by $\sqrt{N}$:

$$\mathbf{a} = \frac{\mathbf{H}^*}{\sqrt{N}} \cdot \mathbf{x} \tag{8.16a}$$

$$\mathbf{x} = \frac{\mathbf{H}}{\sqrt{N}} \cdot \mathbf{a} \tag{8.16b}$$

Note that in this case we obtain a different (by the $\sqrt{N}$ factor) vector $\mathbf{a}$ than in (8.15)! But otherwise we still get a pair of mutually inverse linear transformations, and in this case both transformations are simply unitary rotations of the linear space.

Apparently, the difference between (8.15) and (8.16) is that in the latter we distributed the $1/N$ normalization factor equally between the forward and inverse transformations. We could go even further and move the $1/N$ normalization factor completely into the inverse transformation:

$$\mathbf{a} = \mathbf{H}^* \cdot \mathbf{x} \tag{8.17a}$$

$$\mathbf{x} = \frac{\mathbf{H}}{N} \cdot \mathbf{a} \tag{8.17b}$$

Let's write once again the equations (8.15), (8.16) and (8.17) explicitly for better visibility:

$$a[k] = \frac{1}{N} \sum_{n=0}^{N-1} x[n] e^{-2\pi i k n/N} \tag{8.18a}$$

$$x[n] = \sum_{k=0}^{N-1} a[k] e^{2\pi i k n/N} \tag{8.18b}$$

$$a[k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-2\pi i k n/N} \tag{8.19a}$$

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a[k] e^{2\pi i k n/N} \tag{8.19b}$$

$$a[k] = \sum_{n=0}^{N-1} x[n] e^{-2\pi i k n/N} \tag{8.20a}$$

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} a[k] e^{2\pi i k n/N} \tag{8.20b}$$

Notice that we also switched the sequence notation a_k to the (discrete-argument) function notation $a[k]$. The transformations (8.18), (8.19) and (8.20) thereby can be seen as transforming the function $x[n]$ to the function $a[k]$ and back. Except that... the function $x[n]$ is N-periodic, but the function $a[k]$ has been defined only on the range from $k = 0$ to $k = N - 1$, for which we decided to solve the equations (8.4). However, we can recall that the function $h_k[n]$ in (8.4) are also N-periodic with respect to k (as we stated in (8.6b)). This means that the equation (8.4) at some given k is exactly the same as the equation at $k+N$ and thereby, had we written this latter equation, we would have obtained $a_{k+N} = a_k$. Therefore the function $a[k]$ is also N-periodic. The transformations (8.18), (8.19) and (8.20) thereby transform an N-periodic function $x[n]$ to an N-periodic function $a[k]$ and back.

From that perspective we are going to apply one more notation change: instead of $a[k]$ we shall write $X[k]$, so that the transformations now transform from $x[n]$ to $X[k]$ and back:

(amplitude-version DFT)

$$X[k] = \frac{1}{N} \sum_{n=0}^{N-1} x[n] e^{-2\pi i k n/N} \tag{8.21a}$$

$$x[n] = \sum_{k=0}^{N-1} X[k]e^{2\pi ikn/N} \tag{8.21b}$$

(symmetric-version DFT)

$$X[k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n]e^{-2\pi ikn/N} \tag{8.22a}$$

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X[k]e^{2\pi ikn/N} \tag{8.22b}$$

(density-version DFT)

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-2\pi ikn/N} \tag{8.23a}$$

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k]e^{2\pi ikn/N} \tag{8.23b}$$

The transformations (8.21), (8.22) and (8.23) are referred to as (3 different versions of) *discrete Fourier transform*, or, shortly, *DFT*. The first of each pair of equations (transforming $x[n]$ to $X[k]$) is referred to as *direct transform* and the second of each pair of equations (transforming $X[k]$ back to $x[n]$) is referred to as *inverse transform*.

Conventional usage of the term DFT

Conventionally, the term *discrete Fourier transform* is used in a somewhat more restrictive way. The pair of transformations normally associated with the term *discrete Fourier transform* is (8.23). So one could kind of think of (8.21) as discrete Fourier series and (8.23) as discrete Fourier transform, but that could be still misleading. One further essential feature of discrete Fourier transform, which distinguishes it from discrete Fourier series, is that conventionally it is often seen as being applied to a function $x[n]$ defined on the range 0 to $N-1$ yielding the function $X[k]$ defined on the same range, thereby ignoring the periodicity of these functions. This conventional usage of the term *DFT* will be encountered in the material of Subsection 8.19.7.

Usage of the term DFT in this book

In this book, however, we mostly will use the terms *discrete Fourier transform* and *discrete Fourier series* interchangeably, essentially implying more or less what is conventionally referred to as *discrete Fourier series*. When referring to it as series, we will be primarily interested in representing a function as a linear combination of sinusoids, which is what (8.21) is explic-

itly doing, and therefore the usage of (8.21) (or, even better, (8.14), which is using the a_k notation instead of $X[k]$) would be appropriate. On the other hand, when referring to it as a transform, we will be putting the focus on going from $x[n]$ to $X[k]$ and back. In this case it often doesn't matter much, which of (8.21), (8.22) and (8.23) is being used, since the different placements of the normalization coefficient simply result in differently scaled $X[k]$.

▷ The reason for the way the term *DFT* is used conventionally is probably that DFT is considered as a (rough) approximation of the integral (continuous-argument) Fourier transform, where instead of continuous argument we take discrete argument and instead of an infinitely wide range from $-\infty$ to $+\infty$ the functions are defined on a finite range from 0 to $N-1$.

We can distinguish between (8.21), (8.22) and (8.23) by highlighting the slightly different semantics of $X[k]$ arising from different placement of the normalization factor. In (8.21) the values of $X[k]$ directly represent the amplitudes of the sinusoidal components into which $x[n]$ is being expanded, therefore we might refer to (8.21) as the *amplitude version* of discrete Fourier transform.

On the other hand, in (8.23) the amplitudes of the sinusoidal components are effectively equal to $\frac{1}{N}X[k]$. We shall see in Section 8.3 that this suggests that $X[k]$ has the *amplitude density* semantics. Therefore we shall refer to (8.23) as the *density version*. The density version is also the one which makes most sense if instead of discrete-argument functions $X[k]$ we are having continuous-argument functions $X(\omega)$, which is the case with *Fourier transform* and *discrete-time Fourier transform* (where the latter, despite a very similar name, is not the same as *discrete Fourier transform* which we are currently discussing). Conversely, the amplitude version is better suitable for discrete Fourier series and Fourier series, where $X[k]$ are discrete-argument.

The remaining version (8.22) can be referred to as *symmetric version*. Its main feature is that it is using truly unitary transformations, thereby preserving the lengths of vectors and the values of their dot products.

▷ In this book we are going to use the discrete Fourier transform versions (8.21), (8.22) and (8.23) depending on which of the versions is more convenient for our current purposes. The readers are asked to pay attention to which of the three versions is being used at the moment.

8.3 Domain semantics

Discrete Fourier transform performs a 1:1 mapping of the function $x[n]$ to the function $X[k]$ and back. From some perspective, this can be seen as two different representations of the same function. This latter point of view becomes particularly natural in the interpretation of Fourier transform as a linear space basis change.

Indeed, by (8.15), (8.16) and (8.17), Fourier transform can be seen as just a multiplication of a vector by a matrix, and we already know that such an operation can either represent a linear transformation of the linear space, or a change of basis in that space. In particular, (8.16) can be seen as a change from one unitary basis to another unitary basis.

Time/frequency semantics

In practical usage the arguments n and k of $x[n]$ and $X[k]$ are typically endowed with some semantics. E.g. in signal processing n would normally represent (discrete) time. The integer time argument n of $x[n]$ would thereby have a dimension of "unit time steps", or simply "samples". Let's see what would be the semantics of k in this case.

The terms $a_k e^{2\pi ikn/N}$ of discrete Fourier series in (8.14a) are occasionally referred to as *partials*, the function $x[n]$ thereby being represented as a sum of its partials. The partials all have (complex) sinusoidal shape, their (complex) amplitudes being given by a_k. The set of a_k (which, as we know, fully determines $x[\cdot]$) is respectively referred to as the *signal's spectrum*.

Looking at the partials making up $x[\cdot]$, we notice that at $k = 1$ the respective complex sinusoid $e^{2\pi ikn/N}$ turns into $e^{2\pi in/N}$. It completes a single 2π radian cycle as n goes from 0 to N, corresponding to the frequency $1/N$ cycles per unit time. At $k = 2$ the sinusoid moves twice as fast, corresponding to the frequency $2/N$ cycles per unit time. Generally, the k-th partial is a complex sinusoid of frequency $f_k = k/N$.

Thus all partial frequencies f_k are integer multiples of the base (a.k.a. *fundamental*) frequency $F = 1/N$. We can reflect this by rewriting (8.14a) as

(DFS in terms of fundamental frequency)

$$x[n] = \sum_{k=0}^{N-1} a_k e^{2\pi ikFn}$$

(and rewriting other DFS and DFT formulas in the same fashion).

▷ We also say that f_k are *harmonically related*, since each ratio of f_k/f_m is equal to the ratio of two integers k/m, and more specifically refer to the sinusoidal partials of $x[\cdot]$ as *harmonic partials* or shortly *harmonics*.

Note that in case of DFS the frequency f_k of each partial is only determined up to an integer additive value, since sinusoids of frequencies f_k and f_{k+N} (where $f_{k+N} = (k + N)/N = k/N + 1 = f_k + 1$) are indistinguishable due to *aliasing*, as mentioned in Section 8.1. Possibly because of that, the term *harmonics* is much more commonly used with continuous Fourier series, where each partial's frequency is uniquely defined.

Thus, k represents frequency, measured in units of base frequency (1 unit = F). From that perspective we can consider k to be the (discrete) frequency and respectively refer to the domain of $X[\cdot]$ (where $X[k] = a_k$) as *frequency domain*. The domain of $x[\cdot]$ is time. So we'd be speaking of a *time-domain representation* $x[\cdot]$ and a *frequency-domain representation* $X[k]$ of one and the same signal.

The introduced discrete time and frequency domain semantics of DFT will generalize in other types of Fourier transform to continuous time t and continuous frequency f (or, in cyclic frequency terms, ω).

Amplitude densities

Now we shall convince ourselves that $X[k]$ in the "density DFT version" (8.23) are indeed having density semantics. Let $X[k]$ denote the values obtained from

(8.23) and let a_k denote the partial amplitudes obtained via (8.14) (or, equivalently, via (8.21)).

Since the frequencies of the partials are equal to $f_k = kF = k/N$, the frequency step from one partial to the other is equal to $f_{k+1} - f_k = F = 1/N$. We could picture this as if each partial occupied a frequency band of width $\Delta f = 1/N$. Now looking at (8.23b) we can notice that we are effectively computing a sum of terms $\frac{1}{N}X[k] \cdot e^{2\pi ikn/N}$, which means that the partial amplitudes are $\frac{1}{N}X[k]$. However partial amplitudes must be the same in all of (8.14), (8.21), (8.22) and (8.23), since those are always the same partials. It follows that

(amplitude density DFT semantics)

$$a_k = X[k] \cdot \frac{1}{N} = X[k] \cdot \Delta f \tag{8.24}$$

where a_k are by (8.14) and $X[k]$ are by (8.23).

This suggests that $X[k]$ have the semantics of density (with respect to the frequency), which justifies referring to them as *amplitude densities*.

> ▷ Indeed, this can be compared to the mass density function $\rho(\vec{r})$, where $\vec{r}$ is the position vector. The real mass of a small volume element ΔV at position $\vec{r}$ in this case is given by $\rho(\vec{r}) \cdot \Delta V$. If instead of an essentially 3D object we are given an essentially 1D object, e.g. a rod, we could introduce a 1D density $\rho(x)$, where x is the position coordinate. The mass of a small 1D segment of the rod would then be given by $\rho(x) \cdot \Delta l$, where Δl is the segment's length. Compare this to $a_k = X[k] \cdot \Delta f$ in (8.24).

Other domain semantics

We will often, for the sake of demonstration, assume the time and frequency domain semantics in the discussion of Fourier transform, however there are many situations where the domains have other semantics. We can more abstractly refer to the two domains as the *original domain* and *transform domain* respectively. Or, we could refer to the domains by their variable names, thus speaking of the n-domain and the k-domain (or, in case of continuous domains, of t-domain and f- or ω-domain respectively).

Generally, whenever the exponential terms in Fourier transform have the form e^{ixy} (or, more generally $e^{i\mu xy}$, where μ is some real constant, e.g. 2π or $2\pi/N$), one domain will be related to x and the other to y. In signal processing we are dealing with the exponentials of the form $e^{2\pi ift}$ or $e^{i\omega t}$, so one domain is related to the time t and the other to the frequency f (or ω). When dealing with waves, we are having exponentials of the form $e^{i(kx-\omega t)}$ (where k is the wavenumber rather than the partial index). Respectively there are simultaneously two Fourier transform relationships: one between time t and frequency ω and the other between the position x and the wavenumber k.

> ▷ Indeed, if we fix the value of t at zero ($t = 0$), we obtain $e^{i(kx-\omega t)} = e^{ikx}$. So at $t = 0$ we have a Fourier transform relationship between x and k. Similarly at $x = 0$ we have a Fourier transform relationship between t and ω. For non-zero fixed values of t or non-zero fixed values of x we can factor the exponential as $e^{ikx} \cdot e^{-i\omega t}$. One of the multipliers can then be factored into the partial amplitude, while the other represents the sinusoidal shape of the partial.

In quantum mechanics the wavenumber effectively turns into momentum and frequency into energy, while the exponentials look like $e^{i(px-Et)/\hbar}$. It is therefore common to have the position domain and the momentum domain related via Fourier transform. Quantum mechanics also utilizes the basis change interpretation of Fourier transform by talking of the position basis and the momentum basis.

▷ There is also simultaneously a similar relationship between energy and time, but the time is treated on a special footing in quantum mechanics, in particular one doesn't normally speak of a "time basis" there. Still, some Fourier-like relationships between energy and time manifest themselves. E.g. in case of time-independent potential energy, the expansion of a solution of Schrödinger equation in terms of stationary energy levels is essentially Fourier series (except that the partials are not necessarily harmonically related), the stationary solutions acting similarly to Fourier series coefficients. Or the famous Heisenberg uncertainty relationship holding not only between position and momentum, but also between energy and time.

8.4 Duality

Looking at the direct/inverse transform formulas, especially in their symmetric version (8.22) we can notice that they are almost identical, except for the different sign of the exponent. This suggests that the properties of direct and inverse transforms must be also almost identical. In case of a non-symmetric version there could be additional differences due to asymmetric normalization but that should be it. Let's try to get the details.

Goal 8.2. Formalize the similarities between the direct and inverse transform formulas. *Hint:* use operator notation, or, if one wishes to be less abstract, matrix notation, to express the results.

8.4.1 Symmetric version

Taking (8.22b) we formally replace n by $-n$:

$$x[-n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X[k] e^{-2\pi i k n/N}$$

This looks very close to (8.22b), except that we have $X[k]$ in place of $x[n]$, have $x[-n]$ in place of $X[k]$ and the summation is done over k rather than over n. The latter distinction is removed by formally swapping the k and n variables:

$$x[-k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} X[n] e^{-2\pi i k n/N} \tag{8.25}$$

Now formally the transformation defined by (8.25) looks exactly as (8.22a) except that we have $X[k]$ in place of $x[n]$ and $x[-n]$ in place of $X[k]$. So, if we apply direct Fourier transform (8.22a) to $X[\cdot]$ instead of applying it to $x[\cdot]$, we obtain $x[\cdot]$ back, almost as from an inverse transform application, but with a sign-inverted argument. That is, we can use direct transform instead of inverse one, if we also remember to sign-invert the argument afterwards.

It might be easier to express the same in matrix notation. Let $\mathbf{F} = \mathbf{H}^*/\sqrt{N}$ ("F" standing for "Fourier", not for "frequency), where $\mathbf{H}$ is as in (8.16), therefore $\mathbf{F}^{-1} = \mathbf{H}/\sqrt{N}$. We can rewrite (8.16) as

$$\mathbf{X} = \mathbf{F}\mathbf{x} \tag{8.26a}$$

$$\mathbf{x} = \mathbf{F}^{-1}\mathbf{X} \tag{8.26b}$$

where

$$\mathbf{x} = \begin{bmatrix} x[0] & x[1] & \cdots & x[N-1] \end{bmatrix}^{\mathrm{T}}$$
$$\mathbf{X} = \begin{bmatrix} X[0] & X[1] & \cdots & X[N-1] \end{bmatrix}^{\mathrm{T}}$$

are vectors representing $x[\cdot]$ and $X[\cdot]$.

We also wish to express the argument sign inversion of $x[\cdot]$ in matrix notation. The sign inversion of $x[\cdot]$ means

$$x[0] \leftarrow x[0],\ x[1] \leftarrow x[-1],\ x[-1] \leftarrow x[1],\ x[2] \leftarrow x[-2],\ x[-2] \leftarrow x[2],\ \ldots$$

However, since $x[\cdot]$ is N-periodic, we can describe the same sign-inversion as

$$x[0] \leftarrow x[0],\ x[1] \leftarrow x[N-1],\ x[2] \leftarrow x[N-2],\ \ldots,\ x[N-1] \leftarrow x[1]$$

which means that in vector notation we simply have

$$\mathbf{x} \leftarrow \mathbf{P}\mathbf{x} \tag{8.27a}$$

where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \iddots & \vdots & \vdots \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{8.27b}$$

Combining (8.26) and (8.27) we can reexpress (8.25) as

$$\mathbf{x} = \mathbf{P}\mathbf{F}\mathbf{X} \tag{8.28}$$

Comparing (8.28) to (8.26b) we conclude that

$$\mathbf{F}^{-1} = \mathbf{P}\mathbf{F} \tag{8.29}$$

that is inverse Fourier transform is equal to the direct Fourier transform followed by the argument sign inversion.

Instead of replacing n by $-n$ in (8.22b) we could replace k by $-k$:

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=-(N-1)}^{0} X[-k]e^{-2\pi ikn/N} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X[-k]e^{-2\pi ikn/N}$$

Swapping k and n yields

$$x[k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} X[-n]e^{-2\pi ikn/N} \tag{8.30}$$

instead of (8.25). That is, instead of inverting the argument sign of $x[\cdot]$, we could invert the argument sign of $X[\cdot]$. Respectively in matrix notation we obtain

$$\mathbf{F}^{-1} = \mathbf{FP}$$

in place of (8.29) and thus (8.29) can be augmented into

(duality relationship of direct and inverse Fourier transform)

$$\mathbf{F}^{-1} = \mathbf{PF} = \mathbf{FP} \tag{8.31}$$

that is matrices $\mathbf{F}$ and $\mathbf{P}$ commute.

Exactly the same argument can be made in the opposite direction, obtaining

(duality relationship of inverse and direct Fourier transform)

$$\mathbf{F} = \mathbf{PF}^{-1} = \mathbf{F}^{-1}\mathbf{P} \tag{8.32}$$

Formula (8.32) can be also formally obtained from (8.31) by noticing that $\mathbf{P}$ is its own inverse.

Formulas (8.31) and (8.32) can be also expressed by a single formula

(repeated Fourier transform inverts the argument sign)

$$\mathbf{F}^2 = \mathbf{P} \tag{8.33}$$

which in combination with $\mathbf{P}^{-1} = \mathbf{P}$ can be used to infer (8.31) and (8.32). It also follows that $\mathbf{F}^4 = 1$ since $\mathbf{P}^2 = 1$.

So basically we can always replace (the symmetric version of) direct Fourier transform by inverse Fourier transform or vice versa, as long as we remember to also invert the argument sign either before of after the transform. The property of direct and inverse Fourier transforms turning into each other (as described by (8.31) and (8.32)) is referred to as *duality* of direct and inverse Fourier transforms. Since swapping the direct and inverse transforms also swaps $\mathbf{x}$ and $\mathbf{X}$, we can also speak of duality of the two respective domains (e.g. time and frequency or position and momentum).

8.4.2 Amplitude/density duality

If amplitude or density version is used, we need to additionally mind the different placement of the normalization coefficient $1/N$.

Let

$$\mathbf{F}_A = \frac{\mathbf{H}^*}{N} \qquad \mathbf{F}_A^{-1} = \mathbf{H}$$
$$\mathbf{F}_D = \mathbf{H}^* \qquad \mathbf{F}_D^{-1} = \frac{\mathbf{H}}{N}$$

be the amplitude and density versions of the transformation, as per (8.15) and (8.17).

In order to obtain dual relationships for $\mathbf{F}_A$ and $\mathbf{F}_D$ we can reexpress $\mathbf{F}_A$ and $\mathbf{F}_D$ in terms of the symmetric version $\mathbf{F}$:

$$\mathbf{F}_A = \frac{\mathbf{F}}{\sqrt{N}} \qquad \mathbf{F}_A^{-1} = \sqrt{N} \cdot \mathbf{F}^{-1}$$
$$\mathbf{F}_D = \sqrt{N} \cdot \mathbf{F} \qquad \mathbf{F}_D^{-1} = \frac{\mathbf{F}^{-1}}{\sqrt{N}}$$

Then (8.31), by respectively multiplying or dividing by $\sqrt{N}$, turns into one of

$$\mathbf{F}_A^{-1} = \mathbf{P}\mathbf{F}_D = \mathbf{F}_D\mathbf{P} = N \cdot \mathbf{P}\mathbf{F}_A = N \cdot \mathbf{F}_A\mathbf{P} \tag{8.34a}$$
$$\mathbf{F}_D^{-1} = \mathbf{P}\mathbf{F}_A = \mathbf{F}_A\mathbf{P} = \frac{1}{N}\mathbf{P}\mathbf{F}_D = \frac{1}{N}\mathbf{F}_D\mathbf{P} \tag{8.34b}$$

and (8.32) turns into one of

$$\mathbf{F}_A = \mathbf{P}\mathbf{F}_D^{-1} = \mathbf{F}_D^{-1}\mathbf{P} = \frac{1}{N}\mathbf{P}\mathbf{F}_A^{-1} = \frac{1}{N}\mathbf{F}_A^{-1}\mathbf{P} \tag{8.35a}$$
$$\mathbf{F}_D = \mathbf{P}\mathbf{F}_A^{-1} = \mathbf{F}_A^{-1}\mathbf{P} = N \cdot \mathbf{P}\mathbf{F}_D^{-1} = N \cdot \mathbf{F}_D^{-1}\mathbf{P} \tag{8.35b}$$

That is amplitude version is turned into density version and vice versa, or, if we want to stay within one and the same version, we have to use additional normalization by N or $1/N$.

Goal 8.2 completed. ■

8.4.3 Application principles

The duality property of Fourier transform can be applied to obtains duals of other properties of Fourier transform. Specifically, let's suppose we have demonstrated some property $\mathcal{P}(\mathbf{x}, \mathbf{X})$ of Fourier transform, describing some specific relationship between $\mathbf{x}$ and $\mathbf{X}$. Or more formally

$$\mathcal{P}(\mathbf{x}, \mathbf{X}) \quad \forall \mathbf{x}, \mathbf{X} \colon \mathbf{X} = \mathbf{F}\mathbf{x} \tag{8.36}$$

where $\mathcal{P}(\cdot, \cdot)$ is a Boolean predicate holding for any $\mathbf{x}$, $\mathbf{X}$ which are related via symmetric Fourier transform $\mathbf{F}$.

We can rewrite (8.36) more concisely as

$$\mathcal{P}(\mathbf{x}, \mathbf{F}\mathbf{x}) \quad \forall \mathbf{x} \tag{8.37}$$

By (8.32), $\mathbf{F} = \mathbf{P}\mathbf{F}^{-1}$, therefore (8.37) can be further rewritten as

$$\mathcal{P}(\mathbf{x}, \mathbf{P}\mathbf{F}^{-1}\mathbf{x}) \quad \forall \mathbf{x}$$

Formally replacing the notation $\mathbf{x}$ by $\mathbf{X}$ (which can be done since $\mathbf{x}$ simply takes arbitrary values) yields

$$\mathcal{P}(\mathbf{X}, \mathbf{P}\mathbf{F}^{-1}\mathbf{X}) \quad \forall \mathbf{X}$$

Re-introducing $\mathbf{x} = \mathbf{F}^{-1}\mathbf{X}$ yields

$$\mathcal{P}(\mathbf{X}, \mathbf{P}\mathbf{x}) \quad \forall \mathbf{x}, \mathbf{X} \colon \mathbf{X} = \mathbf{F}\mathbf{x}$$

We have therefore obtained that

(application of Fourier transform's duality)

$$\mathcal{P}(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}(\mathbf{X}, \mathbf{P}\mathbf{x}) \tag{8.38a}$$

(where the left-hand side of (8.38) is understood in the sense of (8.36), that is it must hold for any $\mathbf{x}$, $\mathbf{X}$ related via symmetric Fourier transform). In a similar fashion, we can also obtain a variation of (8.38a):

(application of Fourier transform's duality, alternative version)

$$\mathcal{P}(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}(\mathbf{P}\mathbf{X}, \mathbf{x}) \tag{8.38b}$$

Relationships (8.38a) and (8.38b) can be used to replace any statement $\mathcal{P}(\mathbf{x}, \mathbf{X})$ about two functions $x[n]$ and $X[k]$ related via (symmetric) Fourier transform with its dual $\mathcal{P}(\mathbf{X}, \mathbf{P}\mathbf{x})$ or $\mathcal{P}(\mathbf{P}\mathbf{X}, \mathbf{x})$, where we simply swap $x[\cdot]$ and $X[\cdot]$ and additionally invert the argument sign of one of the functions.

If amplitude or density versions of Fourier transform are involved, (8.38) needs to be adapted to account for different normalizations between direct and inverse transformation formulas. The adaptations are rather straightforward and probably don't bring any new insights, but it might be helpful to write them out for reference. So, let

$$\mathcal{P}_A(\mathbf{x}, \mathbf{X}) \quad \forall \mathbf{x}, \mathbf{X} \colon \mathbf{X} = \mathbf{F}_A \mathbf{x} \tag{8.39}$$

Then, using (8.35a),

$$\mathcal{P}_A(\mathbf{x}, \mathbf{X}) = \mathcal{P}_A(\mathbf{x}, \mathbf{F}_A \mathbf{x}) = \mathcal{P}_A(\mathbf{x}, \mathbf{P}\mathbf{F}_A^{-1}\mathbf{x}/N)$$

Formally replacing $\mathbf{x}$ with $\mathbf{X}$ we continue the above transformations as

$$\mathcal{P}_A(\mathbf{X}, \mathbf{P}\mathbf{F}_A^{-1}\mathbf{X}/N) = \mathcal{P}_A(\mathbf{X}, \mathbf{P}\mathbf{x}/N)$$

So we obtain

$$\mathcal{P}_A(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_A(\mathbf{X}, \mathbf{P}\mathbf{x}/N) \tag{8.40a}$$

in place of (8.38a). Similarly, in place of (8.38b) we obtain

$$\mathcal{P}_A(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_A(\mathbf{P}\mathbf{X}, \mathbf{x}/N) \tag{8.40b}$$

Or, since Fourier transform is a linear transformation, we have $\mathbf{X} = \mathbf{F}\mathbf{x} \iff N\mathbf{X} = \mathbf{F} \cdot N\mathbf{x}$, therefore $\mathcal{P}_A(\mathbf{x}, \mathbf{X})\ \forall \mathbf{x}, \mathbf{X} \iff \mathcal{P}_A(N\mathbf{x}, N\mathbf{X})\ \forall \mathbf{x}, \mathbf{X}$, and therefore (8.40a) and (8.40b) can be rewritten as

$$\mathcal{P}_A(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_A(N\mathbf{X}, \mathbf{P}\mathbf{x}) \tag{8.40c}$$

$$\mathcal{P}_A(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_A(N \cdot \mathbf{P}\mathbf{X}, \mathbf{x}) \tag{8.40d}$$

That is, the duals of $\mathcal{P}_A(\mathbf{x}, \mathbf{X})$ need to divide the replacement of $\mathbf{X}[k]$ by N or multiply the replacement of $x[n]$ by N.

For the density version we respectively have

$$\mathcal{P}_D(\mathbf{x}, \mathbf{X}) = \mathcal{P}_D(\mathbf{x}, \mathbf{F}_D \mathbf{x}) = \mathcal{P}_D(\mathbf{x}, N \cdot \mathbf{P}\mathbf{F}_D^{-1}\mathbf{x})$$

Replacing $\mathbf{x}$ with $\mathbf{X}$,

$$\mathcal{P}_D(\mathbf{X}, N \cdot \mathbf{P}\mathbf{F}_D^{-1}\mathbf{X}) = \mathcal{P}_D(\mathbf{X}, N \cdot \mathbf{P}\mathbf{x})$$

yielding

$$\mathcal{P}_D(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_D(\mathbf{X}, N \cdot \mathbf{P}\mathbf{x}) \tag{8.41a}$$
$$\mathcal{P}_D(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_D(\mathbf{P}\mathbf{X}, N\mathbf{x}) \tag{8.41b}$$

or

$$\mathcal{P}_D(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_D(\mathbf{X}/N, \mathbf{P}\mathbf{x}) \tag{8.41c}$$
$$\mathcal{P}_D(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_D(\mathbf{P}\mathbf{X}/N, \mathbf{x}) \tag{8.41d}$$

That is, the duals of $\mathcal{P}_D(\mathbf{x}, \mathbf{X})$ need to multiply the replacement of $\mathbf{X}[k]$ by N or divide the replacement of $x[n]$ by N.

If we don't need or want to stay within the same version of Fourier transform, we can recall that by (8.34) and (8.35), the density version is the dual of the amplitude version and vice versa. We obtain

$$\mathcal{P}_A(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_D(\mathbf{X}, \mathbf{P}\mathbf{x}) \tag{8.42a}$$
$$\mathcal{P}_A(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_D(\mathbf{P}\mathbf{X}, \mathbf{x}) \tag{8.42b}$$
$$\mathcal{P}_D(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_A(\mathbf{X}, \mathbf{P}\mathbf{x}) \tag{8.42c}$$
$$\mathcal{P}_D(\mathbf{x}, \mathbf{X}) \implies \mathcal{P}_A(\mathbf{P}\mathbf{X}, \mathbf{x}) \tag{8.42d}$$

That is, by swapping $x[\cdot]$ and $X[\cdot]$ and inverting the argument sign in one of the functions, a statement about $x[\cdot]$ and $X[\cdot]$ related via amplitude version of Fourier transform is turned into a statement about $x[\cdot]$ and $X[\cdot]$ related via density version of Fourier transform and vice versa.

8.4.4 Application example: the constant term

As some basic practice of application of Fourier transform's duality property we shall be looking at the "constant term" of the Fourier series, occurring at $k = 0$ (or, in the dual version, at $n = 0$ respectively).

The discrete Fourier series formula (8.14a) can be rewritten as

$$x[n] = a_0 + \sum_{k=1}^{N-1} a_k e^{2\pi ikn/N}$$

since at $k = 0$ we have $e^{2\pi ikn/N} \equiv 1$. In terms of the amplitude version of inverse Fourier transform (8.21b) the same formula looks as

(DFT with an explicit constant term)

$$x[n] = X[0] + \sum_{k=1}^{N-1} X[k] e^{2\pi ikn/N} \tag{8.43a}$$

This justifies referring to a_0 (or $X[0]$) as "the constant term".

▷ In signal processing the term $X[0]$ is often referred to as *DC offset* or *DC term*, where ''DC" stands for "direct current".

The term $X[0]$ can be evaluated by the direct transform formulas, where we notice that at $k = 0$ all exponential factors in these formulas turn to 1. Specifically, in the amplitude version of the transform, (8.21a) turns simply into an arithmetic average of all $x[n]$:

(computation of the DFT's constant term)

$$X[0] = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \tag{8.43b}$$

▷ In the density version of the transform, instead of an arithmetic average of all $x[n]$, (8.23a) at $k = 0$ delivers the total sum of $x[n]$ as $X[0]$. This sum, acting as a "density value" produces a term $X[0]/N$ in the inverse formula (8.23b):

$$x[n] = \frac{X[0]}{N} + \frac{1}{N}\sum_{k=1}^{N-1} X[k]e^{2\pi kn/N}$$

effectively amounting to exactly the same constant term.

Taking (8.43b) we shall now apply the duality property.

Goal 8.3. Use the duality property to obtain the dual of (8.43b).

There are 4 different variants of the duality property (8.40) for the amplitude version of Fourier transform, which one should we pick? For one, it would be easier to pick one where the replacement of $x[\cdot]$ will be multiplied by N, therefore cancelling the $1/N$ factor occurring in the right-hand side of (8.43b). Also, we'd like to pick one where the argument sign inversion occurs during the replacement of $X[\cdot]$, because doing so for the replacement of $x[n]$ will require a reindexing (since the sum is running over n). These two conditions taken simultaneously suggest using (8.40c). Applying (8.40c) to (8.43b) we replace $X[0]$ with $x[-0]$ and replace $x[\cdot]$ with $NX[\cdot]$ yielding

$$x[-0] = \frac{1}{N}\sum_{n=0}^{N-1} NX[n]$$

or

$$x[0] = \sum_{n=0}^{N-1} X[n]$$

or

$$x[0] = \sum_{k=0}^{N-1} X[k] \tag{8.44}$$

which is the dual of (8.43b). The property (8.44) essentially states that at $n = 0$ all sinusoidal terms "sync up in phase", contributing their total sum to the value $x[0]$.

▷ To verify that the result (8.44) obtained by the application of duality is correct, we could directly reobtain it by letting $n = 0$ in (8.21b).

Goal 8.3 completed. ■

It would also be instructive to formally apply the duality to (8.43a).

Goal 8.4. Use the duality property to obtain the dual of (8.43a).

To choose one of the 4 options (8.40) we take the following considerations. We'd like to invert the argument sign in the replacement of $x[n]$, since inverting it in the replacement of $X[k]$ will require a reindexing. As multiplication or division by N goes, we probably should pick division by N of the replacement of $X[k]$, so that the formula would look close to the usual division of a sum by N in the transform formulas. We therefore pick (8.40b). Applying (8.40b) to (8.43a) yields

$$X[-n] = \frac{x[0]}{N} + \sum_{k=1}^{N-1} \frac{x[k]}{N} e^{2\pi ikn/N}$$

Replacing n by $-n$:

$$X[n] = \frac{x[0]}{N} + \sum_{k=1}^{N-1} \frac{x[k]}{N} e^{-2\pi ikn/N}$$

Formally swapping the free variables n and k we obtain the final result

$$X[k] = \frac{x[0]}{N} + \sum_{n=1}^{N-1} \frac{x[n]}{N} e^{-2\pi ikn/N} \tag{8.45}$$

which can be verified by direct comparison to (8.21a).

Goal 8.4 completed. ■

▷ Of course, it would have been much simpler to obtain (8.44) and (8.45) directly from the transform formulas. However we wanted to use this example as an exercise demonstrating the application of the duality of Fourier transform. Some other Fourier transform properties can have noticeably more involved derivations and obtaining them by applying the duality could be saving quite a bit of time.

It is left as an exercise to the reader to apply (8.42) to obtain the density versions of the amplitude version formulas (8.43b) and (8.44).

8.5 Real discrete Fourier transform

In the beginning of this chapter we agreed to use complex sinusoids for developing Fourier theory, because they make the respective math much simpler. This also led to us implicitly assuming that the involved functions $x[n]$ and $X[k]$ can attain complex values. Fields like quantum mechanics may be completely fine with complex sinusoid-based Fourier transform, since the quantum-mechanical wavefunctions are complex valued anyway. Waves or signals in classical physics are a slightly different story, since one deals with physical variables taking real

values. This can make it desired to represent the respective functions as sums of *real* sinusoidal partials. Let's see how this can work out.

Goal 8.5. Obtain a real form of discrete Fourier transform, where a real signal is expanded into a linear combination of real sinusoids.

If we numerically compute the discrete Fourier transform of some randomly picked real functions $x[\cdot]$, we can start noticing that the obtained $X[\cdot]$ always satisfy the property $X[-k] = X^*[k]$ (where the asterisk denotes complex conjugation), which is referred to as $X[k]$ being *Hermitian*. We can then realize that this should always be the case. Indeed, e.g. by (8.21a), assuming real $x[n]$,

(spectra of real signals are Hermitian)

$$\begin{aligned} X[-k] &= \frac{1}{N}\sum_{n=0}^{N-1} x[n]e^{2\pi ikn/N} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\big(e^{-2\pi ikn/N}\big)^* \\ &= \left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]e^{-2\pi ikn/N}\right)^* = X^*[k] \end{aligned} \tag{8.46}$$

There's one problem with the property (8.46). Normally we assume that k varies from 0 to $N-1$ which doesn't sit too well (but only on an intuitive level, formally there is no issue) with $X[k]$ being Hermitian: $X[-k] = X^*[k]$. However, since $X[k]$ is periodic anyway, we could make k go over a bilateral range. If N is odd, this range can be symmetric: from $-(N-1)/2$ to $(N-1)/2$. E.g. if $N = 5$ we can have k varying from -2 to 2. A difficulty, however, arises for an even N, where we cannot have a symmetric bilateral range. E.g. for $N = 6$ we can have k varying from -2 to 3 or from -3 to to 2, but in neither case the range is fully symmetric. This problem is something we have to live with in case of discrete Fourier transform (it gets easier with other Fourier transform types where the domains either become continuous or infinitely long). As a "minor consolation point", for an even N we can notice that in this case we are having $X[-N/2] = X^*[N/2]$ due to being Hermitian, but simultaneously $X[N/2] = X[-N/2]$ due to N-periodicity and it follows that $X[N/2]$ is purely real. On the other hand, the value $X[0]$, defining the constant term, is always real regardless of N being odd or even, since $X[0] = X[-0] = X^*[0]$.

The property (8.46) is what allows us to write a real $x[n]$ as a sum of real sinusoidal partials, despite $X[k]$ being complex. The idea is to notice that not only $X[k]$ and $X[-k]$ are mutually conjugate, but so are the respective complex sinusoids $e^{2\pi ikn/N}$ and $e^{-2\pi ikn/N}$ and thereby so are the entire sinusoidal partials $X[k]e^{2\pi ikn/N}$ and $X[-k]e^{-2\pi ikn/N}$. Adding them both together we obtain a purely real value

$$\begin{aligned} X[k]\cdot e^{2\pi ikn/N} + X[-k]\cdot e^{-2\pi ikn/N} &= X[k]\cdot e^{2\pi ikn/N} + \big(X[k]\cdot e^{2\pi ikn/N}\big)^* \\ &= 2\,\mathrm{Re}\Big(X[k]\cdot e^{2\pi ikn/N}\Big) \end{aligned}$$

Further, consider the product $X[k]\cdot e^{2\pi ikn/N}$ from the complex polar form perspective:

$$X[k]\cdot e^{2\pi ikn/N} = |X[k]|\cdot e^{i\arg X[k]}\cdot e^{2\pi ikn/N} = |X[k]|\cdot e^{i(2\pi kn/N+\arg X[k])}$$

That is the complex arguments add, while the complex magnitude $|X[k]|$ simply determines the complex magnitude of the entire expression. If we take a real part of that product, we should be left simply with a real sinusoidal partial. Specifically we have:

$$\begin{aligned} X[k] \cdot e^{2\pi ikn/N} + X[-k] \cdot e^{-2\pi ikn/N} &= 2\,\mathrm{Re}\Big(X[k] \cdot e^{2\pi ikn/N}\Big) \\ &= 2\,\mathrm{Re}\Big(|X[k]| \cdot e^{i(2\pi kn/N + \arg X[k])}\Big) \\ &= 2 \cdot |X[k]| \cdot \cos\left(\frac{2\pi kn}{N} + \arg X[k]\right) = a_k \cdot \cos\left(2\pi kn/N + \varphi_k\right) \end{aligned} \tag{8.47}$$

where

$$\begin{aligned} a_k &= 2 \cdot |X[k]| \\ \varphi_k &= \arg X[k] \end{aligned}$$

are real numbers.[2] Thus, *each pair of mutually conjugate complex partials adds up to a real sinusoidal partial* $a_k \cos\left(2\pi kn/N + \varphi_k\right)$. Except the partial at $k = 0$ and the partial at $k = N/2$ (optionally, if N is even), which do not have a pair, but they are real by themselves. Indeed, at $k = 0$ we have

$$X[k] \cdot e^{2\pi ikn/N} = X[0] \cdot e^{2\pi i \cdot 0 \cdot n/N} = X[0] \cdot e^0 = X[0] = a_0 \cos(2\pi \cdot 0 \cdot n/N + \varphi_0)$$

where $a_0 = |X[0]|$, $\varphi_0 = \arg X_0$, while at $k = N/2$ we have

$$\begin{aligned} X[k] \cdot e^{2\pi ikn/N} &= X[N/2] \cdot e^{2\pi i(N/2)n/N} = X[N/2] \cdot e^{\pi in} = X[N/2] \cdot (-1)^n \\ &= X[N/2] \cdot \cos \pi n = a_{N/2} \cos\big(2\pi \cdot (N/2) \cdot n/N + \varphi_{N/2}\big) \end{aligned}$$

where $a_{N/2} = |X[N/2]|$, $\varphi_{N/2} = \arg X_{N/2}$.

Thus we obtain

(real-form DFS)

$$x[n] = \sum_{k=0}^{N_r} a_k \cos(2\pi kn/N + \varphi_k) \tag{8.48}$$

where

$$N_r = \begin{cases} (N-1)/2 & \text{if } N \text{ is odd} \\ N/2 & \text{if } N \text{ is even} \end{cases}$$

and

$$\begin{aligned} a_k &= \begin{cases} |X[k]| & \text{if } k = 0 \text{ or } k = N/2 \\ 2 \cdot |X[k]| & \text{otherwise} \end{cases} \\ \varphi_k &= \arg X[k] \end{aligned}$$

The coefficients a_k in (8.48) are referred to as (real) amplitudes. The term φ_k is often referred to as *phase*, but actually the term "phase" denotes the entire

[2] Apparently a_k here are different from (although somewhat related to) the components of the vector $\mathbf{a}$ introduced in (8.8).

argument of the cosine in (8.48). Referring to φ_k as "phase" is just a shortening of the full term "initial phase". Apparently, φ_k is what the argument of the cosine is equal to at $n = 0$, hence it *is* the initial value of this argument, thereby justifying the term "initial phase".

Goal 8.5 completed. ■

Other conventions

There are a number of variations in how the formula (8.48) may be written. One variation arises out of letting $a_0 = X[0]$, $\varphi_0 = 0$ and $a_{N/2} = X[N/2]$, $\varphi_0 = 0$, that is we drop the absolute value operation and formally impose zero initial phases on the purely real partials at $k = 0$ and $k = N/2$ by incorporating the sign into the amplitude coefficients.

Another convention defines $a_0 = 2 \cdot X[0]$, $\varphi_0 = 0$. In this case $X[0]$ is multiplied by 2, similarly to $X[k]$ at other k, but then the first term of (8.48) needs to be written out explicitly:

(real-form DFS with an explicit constant term, one possible version)

$$x[n] = \frac{a_0}{2} + \sum_{k=1}^{N_r} a_k \cos(2\pi kn/N + \varphi_k)$$

A similar convention may be used at $k = N/2$ as well.

Yet another convention can be commonly found in calculus-related treatments of Fourier theory, either as a way to avoid complex numbers or due to historical reasons (or both). The form (8.48) is highly appropriate for interpretations with signal or similar kinds of semantics, since it contains an explicit initial phase term φ_k. For the purposes of solving differential equations and for general theoretical treatment, short of using the complex version, it could be more practical to represent the phase-shifted cosine partials in (8.48) as linear combinations of cosine and sine functions:

$$\begin{aligned} a_k \cdot \cos(2\pi kn/N + \varphi_k) &= a_k \cos(2\pi kn/N)\cos(\varphi_k) - a_k \sin(2\pi kn/N)\sin(\varphi_k) \\ &= \alpha_k \cos(2\pi kn/N) + \beta_k \sin(2\pi kn/N) \end{aligned}$$

where

$$\begin{aligned} \alpha_k &= a_k \cos(\varphi_k) = 2 \cdot |X[k]| \cdot \cos(\arg X[k]) = 2 \operatorname{Re} X[k] \\ \beta_k &= -a_k \sin(\varphi_k) = -2 \cdot |X[k]| \cdot \sin(\arg X[k]) = -2 \operatorname{Im} X[k] \end{aligned}$$

except at $k = 0$ or $k = N/2$ where we obtain

$$\begin{aligned} \alpha_k &= a_k \cos(\varphi_k) = |X[k]| \cdot \cos(\arg X[k]) = \operatorname{Re} X[k] \\ \beta_k &= -a_k \sin(\varphi_k) = -|X[k]| \cdot \sin(\arg X[k]) = -\operatorname{Im} X[k] \end{aligned}$$

Thereby (8.48) turns into

(real-form DFS without initial phases)

$$x[n] = \sum_{k=0}^{N_r} \Big(\alpha_k \cos(2\pi kn/N) + \beta_k \sin(2\pi kn/N) \Big) \tag{8.49}$$

8.6 Symmetries

We have established in (8.46) that Fourier transform $X[k]$ of a real sequence $x[n]$ is Hermitian. However from (8.47) we can also notice that the converse is true as well: if $X[k]$ is Hermitian, then $x[n]$ is real. Indeed, given a Hermitian $X[k]$, the inverse Fourier transform sum (8.21b) can be organized into pairs of mutually conjugate partials (plus one or two purely real partials at $k = 0$ and $k = N/2$) yielding a purely real sum (8.48). Thereby

$$X[-k] = X^*[k] \ \forall k \iff x[n] \in \mathbb{R} \ \forall n \tag{8.50}$$

that is a purely real $x[\cdot]$ corresponds to a Hermitian $X[\cdot]$. It can be shown in a similar fashion that a purely imaginary $x[\cdot]$ corresponds to an anti-Hermitian $X[\cdot]$:

$$X[-k] = -X^*[k] \ \forall k \iff \operatorname{Re} x[n] = 0 \ \forall n$$

▷ Of course, one could simply demonstrate (8.50) by a direct computation of a Hermitian conjugate of (8.21b) under the assumption of a Hermitian $X[k]$ and showing that $x^*[n] = x[n]$ in this case. However, inferring it from (8.47) instead, as we did it, gives some additional insight about how the mutually-conjugate pairing of the Fourier series terms results in a real $x[n]$.

Further, we can notice that in (8.49) the cosine terms define the even component of the function and the sine terms define the odd component of the function. It follows that a purely real $x[n]$ is even iff $\beta_k = 0 \ \forall n$ and is odd if $\alpha_k = 0 \ \forall n$. Or, in terms of $X[k]$, a purely real $x[n]$ is even iff all $X[k]$ are real and is odd iff all $X[k]$ are imaginary.

We can ask ourselves, if the latter property also holds for complex $x[n]$. A bit of quick thinking should reveal that it doesn't. Indeed, applying the duality of Fourier transform to (8.50) yields

$$X[k] \in \mathbb{R} \ \forall k \iff x[-n] = x^*[n] \ \forall n$$

(applying (8.40d) to (8.50) we replace $x[n]$ by $N \cdot X[-n]$ and replace $X[k]$ by $x[k]$, yielding $x[k] = x^*[-k] \ \forall k \iff N \cdot X[-n] \in \mathbb{R} \ \forall n$). Therefore real $X[\cdot]$ corresponds to a Hermitian rather than even $x[\cdot]$, and only if $x[\cdot]$ is also real, being Hermitian implies being even. A similar conclusion is obtained for an imaginary $X[\cdot]$, where we get an anti-Hermitian $x[\cdot]$:

$$\operatorname{Re} X[k] = 0 \ \forall k \iff x[-n] = -x^*[n] \ \forall n$$

Summing the previous findings up, we get

(Hermitian and anti-Hermitian spectra)

$$\text{x[n] is real} \iff \text{X[k] is Hermitian} \tag{8.51a}$$
$$\text{x[n] is imaginary} \iff \text{X[k] is anti-Hermitian} \tag{8.51b}$$

and their duals

(real and imaginary spectra)

$$\text{x[n] is Hermitian} \iff \text{X[k] is real} \tag{8.52a}$$
$$\text{x[n] is anti-Hermitian} \iff \text{X[k] is imaginary} \tag{8.52b}$$

while imposing (8.51a) on top of (8.52) yields

(spectra of even/odd real signals)

$$\text{x[n] is real even} \iff \text{X[k] is real even} \tag{8.53a}$$
$$\text{x[n] is real odd} \iff \text{X[k] is imaginary odd} \tag{8.53b}$$

Instead of analyzing $x[\cdot]$ and $X[\cdot]$ which are already known to have some symmetries, let's consider the transformations behind those symmetries. E.g. let's ask ourselves, what is going to happen if we replace $x[n]$ with $x[-n]$. If $x[n]$ is real even, obviously nothing would happen, but what if it's not? So let $x'[n] = x[-n]$. By (8.21a)

$$\begin{aligned} X'[k] &= \frac{1}{N}\sum_{n=0}^{N-1} x'[n]e^{-2\pi ikn/N} = \frac{1}{N}\sum_{n=0}^{N-1} x[-n]e^{-2\pi ikn/N} \\ &= \frac{1}{N}\sum_{n=-(N-1)}^{0} x[n]e^{2\pi ikn/N} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]e^{2\pi ikn/N} \\ &= \frac{1}{N}\sum_{n=0}^{N-1} x[n]e^{-2\pi i\cdot(-k)\cdot n/N} = X[-k] \end{aligned}$$

where we replaced the summation over the range from $-(N-1)$ to 0 with the summation over the range from 0 to $(N-1)$ by the N-periodicity of each of the terms of the sum.

Similarly, for $x'[n] = x^*[n]$ we obtain

$$\begin{aligned} X'[k] &= \frac{1}{N}\sum_{n=0}^{N-1} x'[n]e^{-2\pi ikn/N} = \frac{1}{N}\sum_{n=0}^{N-1} x^*[n]e^{-2\pi ikn/N} \\ &= \left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]e^{2\pi ikn/N}\right)^* = \left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]e^{-2\pi i\cdot(-k)\cdot n/N}\right)^* = X^*[-k] \end{aligned}$$

We also can obtain by duality of Fourier transform (or explicitly, or by combining the already obtained results) that for $X'[k] = X^*[k]$ we have $x'[n] = x^*[-n]$.

Summing up,

(dual-domain relationships for inversions of the argument axis and/or conjugation)

$$x'[n] = x[-n] \iff X'[k] = X[-k] \tag{8.54a}$$
$$x'[n] = x^*[n] \iff X'[k] = X^*[-k] \tag{8.54b}$$
$$x'[n] = x^*[-n] \iff X'[k] = X^*[k] \tag{8.54c}$$

If $x[\cdot]$ is real, there is no difference between $x[n]$ and $x^*[n]$, while $X[\cdot]$ is Hermitian, so the transformations in (8.54b) do not do anything, while the remaining two properties (8.54a) and (8.54c) collapse into a single one:

(time-inversion for real signals inverts the initial phases of the partials)

$$x'[n] = x[-n] \iff X'[k] = X^*[k] \qquad \text{for real } x[\cdot] \tag{8.55}$$

This means that $|X[k]|$ is unchanged, while $\arg X[k]$ inverts its sign. This, in turn, means that in the real Fourier series form (8.48) a_k are unchanged while φ_k invert their signs.

Apparently, there is also a dual of (8.55), which is not so interesting, since there is nothing too special about purely real spectrum. The set of duals of (8.54) coincides with itself: (8.54a) is its own dual, while (8.54b) and (8.54c) are duals of each other.

The symmetry and related transformation properties of DFT analyzed in this section generalize to other types of Fourier transform (discrete-time Fourier transform, Fourier series and integral Fourier transform) discussed later in this chapter, but these properties look pretty much the same, and there are no new insights to be obtained, so we are not going to repeat the same discussion again.

8.7 Discrete-time Fourier transform

Math texts typically do not cover *discrete-time Fourier transform* (or, shortly, *DTFT*), probably considering the DTFT to simply be a dual of the (continuous) Fourier series. The term "discrete-time Fourier transform" itself belongs more to texts focusing on signal processing, which is also reflected in the term itself, by suggesting that one of the two domains has time semantics. In this book, however, we believe that some additional insights can be gained by considering the DTFT in addition to the continuous Fourier series. The key distinction between DTFT and continuous Fourier series is which of the two domains is considered to be the original ("preimage") domain, and which one is considered to contain the transform's image. Formally this, again, doesn't make much difference, but intuitively puts the things into quite a different perspective, particularly in regards to the choice of the domain scale, normalization placement and convergence conditions.

As DTFT is commonly used in signal processing, in order to make things intuitive we will also assume the time and frequency semantics for the two domains. In discrete Fourier transform (DFT) both domains were discrete and the representations of functions in both domains were N-periodic. Say, we wish

to lift the requirement of periodicity in the time domain and somehow become able to deal with nonperiodic functions of discrete time. So let $N \to \infty$ in (8.21), (8.22) and (8.23) (thereby treating a nonperiodic function $x[n]$ as if it was periodic with an infinitely large period) and then let's see what happens.

Goal 8.6. Design a way to apply Fourier transform to a nonperiodic discrete-argument function $x[n]$.

Let's start with (8.21a). The normalization by $\frac{1}{N}$ gives a good chance that the right-hand side of (8.21a) is not going to explode to infinity (due to an infinite sum from 0 to $N-1$), but will stay within finite bounds, however that's actually all. While it is staying within finite bounds, there is no reason to expect that the expression (8.21a) is going to converge to any particular value. In comparison, (8.23a) is likely to simply explode... except, what if, say, $x[n]$ is absolutely summable, e.g. $x[n]$ decay as $1/n^2$? In that case, actually (8.23a) *is* going to converge to a finite value. We are still not sure about the convergence of (8.23b), but at least we might be getting somewhere.

▷ Before we get too focused on (8.23a), let's brush (8.21a) and (8.22a) completely off for now, since they are just going to be converging to zero in case of an absolutely summable $x[n]$.

The sum in (8.23a) is running from 0 to $N-1$, which in case of $N \to \infty$ runs from 0 to ∞, suggesting that $x[n]$ should be a unilateral nonperiodic sequence. However, we can also rewrite (8.23a) in a bilateral form. With discrete Fourier transform on its own, the bilateral form might be not so popular due to the symmetry problems occurring for even N mentioned in Section 8.5. However, as we shall see next, as we get to other Fourier transform types besides DFT, these symmetry problems are going to disappear. We therefore assume that $x[n]$ is a bilateral sequence and simultaneously make the summation of k bilateral, so that we can also better exploit the mutual conjugate symmetry of the sinusoidal terms in case of real $x[n]$ (although the latter is not a must for the subsequent discussion). Rewriting both (8.23a) and (8.23b) in the bilateral form we obtain

$$X[k] = \sum_{n=-N/2}^{N/2} x[n] e^{-2\pi i k n/N} \tag{8.56a}$$

$$x[n] = \frac{1}{N} \cdot \sum_{k=-N/2}^{N/2} X[k] e^{2\pi i k n/N} \tag{8.56b}$$

▷ We haven't been too careful with the summation index ranges in (8.56), actually we should have written $\pm(N-1)/2$ for an odd N and something like $-N/2+1$ to $N/2$ for an even N, but this detail is about to disappear in the subsequent transformations anyway. Let's just understand the summation index ranges in (8.56) as approximately symmetric bilateral ranges of total length N.

So, say (8.56a) converges at $N \to \infty$ (due to the absolute summability of $x[n]$, or, even if $x[n]$ is not absolutely summable, maybe it's still "nice enough" to get (8.56a) to converge). Now recall that the sinusoidal partials in the discrete Fourier transform are harmonics with the base frequency $1/N$, where k is the multiplier. That is, they are having frequencies $f_k = k/N$, so we may rewrite

(8.56) as

$$X[k] = \sum_{n=-N/2}^{N/2} x[n] e^{-2\pi i f_k n} \tag{8.57a}$$

$$x[n] = \sum_{k=-N/2}^{N/2} X[k] e^{2\pi i f_k n} \Delta f \tag{8.57b}$$

where $\Delta f = 1/N$. Now recall that $X[k]$ are N-periodic. Respectively, the corresponding period of values of f_k has a length of 1. As k goes from $-N/2$ to $N/2$, f_k go from $-\frac{1}{2}$ to $\frac{1}{2}$ with a step of $\Delta f = 1/N$. As $N \to \infty$, this discrete range of $-\frac{1}{2}$ to $\frac{1}{2}$ turns into a continuous range from $-\frac{1}{2}$ to $\frac{1}{2}$ in which a continuous variable f is varying, and (8.57a) becomes

(direct DTFT)

$$X(f) = \sum_{n=-\infty}^{\infty} x[n] e^{-2\pi i f n} \tag{8.58a}$$

while (8.57b) turns into an integral

(inverse DTFT)

$$x[n] = \int_{-1/2}^{1/2} X(f) e^{2\pi i f n} \, \mathrm{d}f \tag{8.58b}$$

The transformation from $x[n]$ to $X(f)$ and back, described by (8.58), is known as *discrete-time Fourier transform*, or, shortly, *DTFT*.

▷ The difference between the names *discrete Fourier transform (DFT)* and *discrete-time Fourier transform (DTFT)* is that the latter specifically states the discreteness of the time only (while the frequency is continuous), whereas in discrete Fourier transform both time and frequency are discrete.

▷ As (8.58a) was obtained as a limiting case of (8.57a), the infinite sum in (8.58a) is generally understood in the Cauchy principal value sense:

$$X(f) = \text{p.v.} \sum_{n=-\infty}^{\infty} x[n] e^{-2\pi i f n} = \lim_{N\to\infty} \sum_{n=-N}^{N} x[n] e^{-2\pi i f n}$$

As the N-period of $X[k]$ turns into the 1-period of $X(f)$, the function $X(f)$ is 1-periodic (and so is the entire expression under the integral sign in (8.58b), thereby it also doesn't matter, over which period the integral (8.58b) is taken, e.g. it could be taken over $[0, 1]$ instead). The function $x[n]$, as we should remember, is an infinitely long function, decaying with a sufficient speed to zero as $n \to \infty$.

▷ The exact answer to the question, whether (8.58a) converges and whether, what is equally important, (8.58b) actually delivers back the original $x[n]$, can get pretty complicated (some further details will be covered in Section 8.13.4). Intuition, however, should tell us that potentially problematic cases are probably occurring closely to the boundaries of convergence, e.g. if $x[n]$ is not absolutely summable.

So, while discrete Fourier transform, being fully finite, is guaranteed to work for any $x[n]$ or any $X[k]$, the discrete-time Fourier transform doesn't offer such guarantee (although usually it works for "reasonable" functions).

It is also common to write (8.58) in terms of cyclic frequency, as an alternative. Substituting $f = \omega/2\pi$ into (8.58) we obtain

(DTFT in terms of cyclic frequency)

$$X(\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-i\omega n} \tag{8.59a}$$

$$x[n] = \int_{-\pi}^{\pi} X(\omega)e^{i\omega n}\,\frac{\mathrm{d}\omega}{2\pi} = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)e^{i\omega n}\,\mathrm{d}\omega \tag{8.59b}$$

▷ Notice that $X(\cdot)$ in (8.59) is not the same one as in (8.58), the former is 2π-periodic, the latter is 1-periodic. Formally we should have labelled them as two different functions, e.g. the one in (8.58) could have been labelled $X_f(\cdot)$ and the one in (8.59) $X_\omega(\cdot)$, where we would have $X_\omega(2\pi f) = X_f(f)$.

Now, we have arrived at (8.58) and (8.59) starting from the density version (8.23) of discrete Fourier transform. Respectively the functions $X(f)$ and $X(\omega)$ work as amplitude densities rather than amplitudes. Indeed, if we consider an integral as a sum of infinitely small sinusoidal terms, the "linear combination coefficients" in (8.58b) and (8.59b) will be respectively $X(f)\,\mathrm{d}f$ and $X(\omega)\,\frac{\mathrm{d}\omega}{2\pi}$. Which means that the complex amplitudes themselves are given by these expressions $X(f)\,\mathrm{d}f$ and $X(\omega)\,\frac{\mathrm{d}\omega}{2\pi}$, while $X(f)$ and $X(\omega)$ are just amplitude densities.

Notice that in each of (8.58b) and (8.59b) the integration is done over $\mathrm{d}f$ (which in (8.59b) is notated as $\frac{\mathrm{d}\omega}{2\pi}$), that is we are integrating with respect to ordinary rather than cyclic frequency.

▷ Similarly to the three different normalization versions which we introduced with discrete Fourier transform, the f- and ω-based discrete-time Fourier transform (as well as Fourier series and integral Fourier transform later) will be both encountered throughout the book. Sometimes it will be more convenient to use the f-based transform formulas, sometimes ω-based transform will be more convenient. The readers are asked to pay attention to this detail. Normally one would be able to tell between $X(f)$ and $X(\omega)$ by the letter used for the argument.

Goal 8.6 completed. ■

8.8 Fourier series

As we mentioned that (continuous) Fourier series is a dual of discrete-time Fourier transform, we could just take the duals of the results obtained in Section 8.7 in order to obtain Fourier series formulas. However, as we also mentioned, we believe that there is value in looking at Fourier series from a different

perspective. So instead of simply taking the duals of the DTFT formulas (8.58) and/or (8.59), we shall reobtain the Fourier series from scratch.

> ▷ When the term "Fourier series" is used without additional qualification or context, continuous Fourier series is implied. Therefore, whenever we say "Fourier series" we are going to mean "continuous Fourier series", unless the context clearly suggests discrete Fourier series.

Goal 8.7. Design a way to apply Fourier transform to a periodic continuous-argument function $x(t)$. *Hint:* Take the steps dual to the derivation of DTFT.

Applying the duality principle to the derivation of the discrete-time Fourier transform, we can take the discrete Fourier transform setup and this time turn $X[k]$ into an infinite bilateral nonperiodic sequence, with k going from $-\infty$ to ∞. While we have used (8.23) to derive discrete-time Fourier transform, for a dual derivation we shall use the dual of (8.23), which is (8.21) (specifically, the normalization placement makes these two versions the duals of each other). Rewriting (8.21) in the bilateral form:

$$X[k] = \frac{1}{N}\sum_{n=-N/2}^{N/2} x[n]e^{-2\pi ikn/N} \tag{8.60a}$$

$$x[n] = \sum_{k=-N/2}^{N/2} X[k]e^{2\pi ikn/N} \tag{8.60b}$$

we want to turn n/N into a continuous variable, same as we previously turned k/N into a continuous variable f. Let's turn n/N into a continuous variable τ (without being certain of its exact semantics or units yet). Letting $N \to \infty$ in (8.60) we obtain

$$X[k] = \int_{-1/2}^{1/2} x(\tau)e^{-2\pi ik\tau}\,\mathrm{d}\tau \tag{8.61a}$$

$$x(\tau) = \sum_{k=-\infty}^{\infty} X[k]e^{2\pi ik\tau} \tag{8.61b}$$

where the entire period of n going from $-N/2$ to $N/2$ turns into τ going from $-\frac{1}{2}$ to $\frac{1}{2}$. The function $x(\tau)$ is thereby 1-periodic. The variable τ, taking over the function of n, thereby, in the time/frequency interpretation of Fourier transform, takes the time semantics.

With DTFT, having "natural discrete time units" (one discrete time step being equal to one time unit) was relatively natural. After all, why not "measure discrete time in samples"? A unit time step in DTFT corresponds to the 1-periodic frequency-domain counterpart, which is a common (although not necessary) assumption with DTFT. In comparison, the assumption of 1-periodicity of the original function with Fourier series is less common. While of course we can convert all functions of different periods to a unit period each, this may be not so intuitive or convenient. Let's therefore replace τ with t/T where T is some non-unit period:

$$X[k] = \int_{-T/2}^{T/2} x(t)e^{-2\pi ikt/T}\,\frac{\mathrm{d}t}{T} \tag{8.62a}$$

$$x(t) = \sum_{k=-\infty}^{\infty} X[k] e^{2\pi i k t/T} \tag{8.62b}$$

▷ Notice that $x(t)$ in (8.62) is not the same one as in (8.61), the former is T-periodic, the latter is 1-periodic. Formally we should have labelled them as two different functions, e.g. the one in (8.61) could have been labelled $x_1(t)$ and the one in (8.62) $x_T(t)$, where we would have $x_1(t) = x_T(t/T)$.

Letting $F = 1/T$ in (8.62) (where, as it is not difficult to realize, F is the base frequency, given the time semantics of t) yields

(Fourier series, bracket index notation)

$$X[k] = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-2\pi i k F t} \, \mathrm{d}t \tag{8.63a}$$

$$x(t) = \sum_{k=-\infty}^{\infty} X[k] e^{2\pi i k F t} \tag{8.63b}$$

The equation (8.63b) represents a T-periodic function $x(t)$ as a linear combination of sinusoidal components of frequencies kF where $F = 1/T$, thereby $X[k]$ are the amplitudes of these sinusoids. It might be more appropriate therefore to switch to the index notation X_k instead of $X[k]$ and we obtain

(Fourier series, subscript index notation)

$$X_k = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-2\pi i k F t} \, \mathrm{d}t \tag{8.64a}$$

$$x(t) = \sum_{k=-\infty}^{\infty} X_k e^{2\pi i k F t} \tag{8.64b}$$

The right-hand side of (8.64b) is referred to as (continuous) *Fourier series*. The formula (8.64a) gives the coefficients X_k of the linear combination (8.64b), also known as *Fourier series coefficients*. Notice that due do T-periodicity of the entire expression under the integral sign in (8.64a), the integral can be taken over any other period, e.g. $[0, T]$.

▷ Similarly to the direct DTFT formula (8.58a), the sum in (8.64b) is understood in the Cauchy principal value sense:

$$X(t) = \text{p.v.} \sum_{k=-\infty}^{\infty} X_k e^{2\pi i k F t} = \lim_{N\to\infty} \sum_{k=-N}^{N} X_k e^{2\pi i k F t}$$

▷ This time we started from the amplitude version (8.21) of discrete Fourier transform and respectively X_k in (8.64b) and (8.65b) give the amplitudes of the sinusoidal components rather than amplitude densities.

▷ By (8.64a), X_0 is simply the average value of $x(t)$, computed over its period.

It is also common to write (8.64) in terms of cyclic frequency, which we can obtain by substituting $F = \Omega/2\pi$ into (8.64) yielding

(Fourier series, cyclic-frequency form)

$$X_k = \frac{1}{T}\int_{-T/2}^{T/2} x(t)e^{-ik\Omega t}\,\mathrm{d}t \tag{8.65a}$$

$$x(t) = \sum_{k=-\infty}^{\infty} X_k e^{ik\Omega t} \tag{8.65b}$$

▷ Similarly to discrete-time Fourier transform, the question of convergence of Fourier series is rather complicated, and the answers also may vary from one source to the other due to the formulated conditions being sufficient rather than sufficient and necessary. Intuitively, problems are more likely to occur in boundary cases.

Being only sufficient, the conditions for the Fourier series convergence are also unlikely to be the duals of the conditions of discrete Fourier series convergence, due to a different purpose. With discrete-time Fourier transform we are usually given an infinitely long sequence $x[n]$ and want to convert it to $X(\omega)$ and back, while with Fourier series we usually start with a periodic function $x(t)$ and want to convert it to X_k and back. So, for discrete-time Fourier transform the convergence conditions usually specify the requirements for $x[n]$, while for Fourier series the convergence conditions usually specify the requirements for $x(t)$ (rather than for X_k, which would be the dual of specifying the requirements for $x[n]$).

▷ It might seem that Fourier series loses information during the direct transform (8.64a) and restores the information "out of nowhere" during the inverse transform (8.64b). Indeed the function $x(t)$ is specified on an uncountable set of points $[0, T)$, while the sequence X_k is specified on a countable set of points $k \in \mathbb{Z}$, therefore $x(t)$ must have "infinitely more" degrees of freedom than X_k. This is, however, not true, since $x(t)$, roughly speaking, must be a continuous function and thus it suffices to define it at the rational points, the remaining values can be established from the function's continuity.

The same consideration also applies to the discrete-time Fourier transform. It also applies to continuous Fourier transform, where, despite dealing with functions of continuous argument in both domains, the de-facto amount of freedom degrees is countable.

▷ Fourier series could have been referred to as *discrete-frequency Fourier transform* or *DFFT*, using an obvious similarity to how the term DTFT compares to DFT. This term is, however, pretty uncommon.

Goal 8.7 completed. ■

Goal 8.8. Obtain a real form of the Fourier series, similar to the real form of DFT.

The real version of Fourier series can be obtained along the lines of Section 8.5. Assuming real $x(t)$, the coefficients obtained by (8.64a) will make up a Hermitian sequence: $X_{-k} = X_k^*$. Respectively we can rewrite (8.64b) as

$$x(t) = \sum_{k=-\infty}^{\infty} X_k e^{2\pi ikFt} = X_0 + \sum_{k=1}^{\infty}(X_k e^{2\pi ikFt} + X_{-k}e^{-2\pi ikFt})$$

$$= X_0 + \sum_{k=1}^{\infty}\Big(X_k e^{2\pi i k F t} + (X_k e^{2\pi i k F t})^*\Big) = X_0 + 2\sum_{k=1}^{\infty}\mathrm{Re}(X_k e^{2\pi i k F t})$$
$$= X_0 + 2\sum_{k=1}^{\infty}\Big(|X_k| \cdot \cos(2\pi k F t + \arg X_k)\Big)$$

finally yielding

(real form of Fourier series)

$$x(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(2\pi k F t + \varphi_k) = \sum_{k=0}^{\infty} a_k \cos(2\pi k F t + \varphi_k) \tag{8.66}$$

where $a_0 = X_0$, $\varphi_0 = 0$ and $a_k = 2 \cdot |X_k|$, $\varphi_k = \arg X_k$ for $k > 0$.

▷ The term a_0 in (8.66), if written out separately, is often notated as $\frac{a_0}{2}$, so that $a_0 = 2X_0$ which is more consistent with $a_k = 2 \cdot |X_k|$:

$$x(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(2\pi k F t + \varphi_k)$$

In terms of cyclic frequency ω, (8.66) looks like

(real form of Fourier series in terms of cyclic frequency)

$$x(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(k\Omega t + \varphi_k) = \sum_{k=0}^{\infty} a_k \cos(k\Omega t + \varphi_k) \tag{8.67}$$

▷ Discrete-time Fourier transform obviously also has a real version, we just didn't bother to derive it here. The readers might do so, if they wish, as an exercise.

Goal 8.8 completed. ■

Even and odd harmonics

Not unlike the DFS case, the frequencies of the components of Fourier series (8.64b) and (8.65b) are equal to kF and $k\Omega$ respectively, where F or Ω is the base (fundamental) frequency, thereby, like in the DFS case, we are having a sum of harmonic partials. Unlike the DFS case, the frequencies of these partials do not alias, therefore each partial has a uniquely defined frequency, thereby making the whole picture quite a bit more clear compared to DFS.

It is particularly interesting to look at even and odd harmonics separately. It is not difficult to notice that all even harmonics are actually not only (formally) T-periodic, but also $T/2$-periodic. Indeed, the 1st harmonic completes its cycle over a time period of T. The 2nd harmonic completes 2 cycles over the same period. The 4th harmonic completes 4 cycles, etc. So all even harmonics complete an even number of cycles over a period of T. Over a period of $T/2$ they complete half as many cycles, however, since the numbers of cycles were

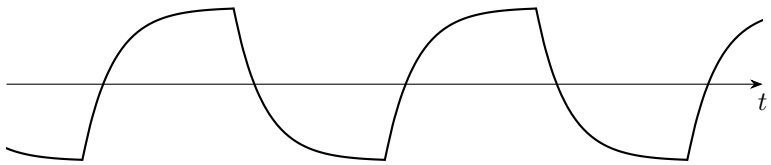


Figure 8.1: A function (signal) consisting only of odd harmonics.

originally all even, we still get integer numbers of cycles over a period of $T/2$. Formally

$$x_{2k}(t + T/2) = x_{2k}(t)$$

where $x_k(t) = a_k \cos(2\pi kFt + \varphi_k)$ is the k-th partial of the Fourier series expansion.

Similarly, all odd harmonics complete an integer number of cycles and a half over a time period of $T/2$. Therefore they are "$T/2$-antiperiodic":

$$x_{2k+1}(t + T/2) = -x_{2k+1}(t)$$

Now imagine the function $x(t)$ consists of only even harmonics. Then

$$x(t + T/2) = \sum_{k=-\infty}^{\infty} x_k(t + T/2) = \sum_{k=-\infty}^{\infty} x_{2k}(t + T/2) = \sum_{k=-\infty}^{\infty} x_{2k}(t) = x(t)$$

Conversely, if $x(t)$ consists only of odd harmonics,

$$x(t + T/2) = \sum_{k=-\infty}^{\infty} x_k(t + T/2) = \sum_{k=-\infty}^{\infty} x_{2k+1}(t + T/2) = -\sum_{k=-\infty}^{\infty} x_{2k+1}(t) = -x(t)$$

We obtain

(odd/even-harmonic signals)

$x(t + T/2) = x(t)$	if $x(t)$ consists of even harmonics only
$x(t + T/2) = -x(t)$	if $x(t)$ consists of odd harmonics only

So, a function consisting only of even harmonics repeats itself after $T/2$, a function consisting only of odd harmonics changes its sign after $T/2$ (as illustrated in Fig. 8.1).

8.9 Fourier transform

The (continuous, a.k.a. integral) Fourier transform combines the "efforts" of discrete-time Fourier transform and Fourier series. The discrete-time Fourier transform made the time-domain representation infinitely long and nonperiodic, while its dual, Fourier series, made the frequency-domain representation infinitely long and nonperiodic. Simultaneously, the dual domain stayed periodic, but became continuous. In Fourier transform we are aiming to make the representations in both domains infinitely long, nonperiodic and continuous.

▷ Same as with Fourier series, unless there is an explicit qualification or context, whenever the term "Fourier transform" is encountered, by default it implies continuous (integral) Fourier transform. In this book, however, an extra caution has to be exercised in this regard, since we are also collectively referring to discrete Fourier series, discrete-time Fourier transform, (continuous) Fourier series and (continuous) Fourier transform as 4 different types of Fourier transform.

Goal 8.9. Design a way to apply Fourier transform to a non-periodic, potentially infinitely long, continuous-argument function $x(t)$.

It doesn't matter much where we start, we could start from discrete-time Fourier transform, or we could start from Fourier series. In the first case we would have to stretch the frequency-domain representation making it infinitely long, in the second case we would have to do the same to the time-domain representation.

Let's start with Fourier series (8.63)[3] and let $T \to \infty$. Similarly to our derivation of the discrete-time Fourier transform in Section 8.7, we notice that the expression in the right-hand side of (8.63a), while staying bounded due to the normalization by $1/T$, is unlikely to converge to a specific value. So we need to do the same as in Section 8.7: we assume that $x(t)$ is absolutely integrable (or is otherwise "nice enough" to let the integral in the right-hand side of (8.63a) converge). Except that we notice that in this case, while the integral in the right-hand side of (8.63a) is going to converge to some finite value, the entire right-hand side of (8.63a) is going to converge to zero due to the normalization by $1/T$. At this point we shall recall that in Section 8.7 we were using the density version of discrete Fourier transform, while (8.63) is a variation of the amplitude version. This is not a problem, since we can move the normalization by $1/T$ over to (8.63b), obtaining a "density version of Fourier series":

$$X[k] = \int_{-T/2}^{T/2} x(t) e^{-2\pi i k F t} \, \mathrm{d}t \tag{8.68a}$$

$$x(t) = \frac{1}{T} \cdot \sum_{k=-\infty}^{\infty} X[k] e^{2\pi i k F t} \tag{8.68b}$$

▷ The amplitudes of the sinusoids in (8.68b) are effectively $X[k]/T = X[k] \cdot F$, where $F = 1/T$ is the step between the neighboring frequencies of the sinusoidal components of the Fourier series. Therefore $X[k]$ are having density semantics.

We thereby assume that (8.68a) converges at $T \to \infty$. It remains to find out what happens at the same time to (8.68b), where we can notice that $1/T = F$ is exactly the step between the frequencies of the complex sinusoidal partials of (8.68b), the frequencies themselves being $f_k = kF = k/T$. At $T \to \infty$ this step becomes infinitely small, so the sum in (8.68b) must be turning into an integral. Letting $\Delta f = f_{k+1} - f_k = F = 1/T$, we rewrite (8.68b) as

$$x(t) = \sum_{k=-\infty}^{\infty} X[k] e^{2\pi i f_k t} \Delta f$$

[3] We picked the function-style notation version (8.63) instead of the index-style notation version (8.64) because our intention is to convert the sequence $X[k]$ into a function.

Replacing $X[k]$ with $X(f_k)$, where $X(f)$ is the continuous-argument function that we're aiming to obtain from $X[k]$, gives

$$x(t) = \sum_{k=-\infty}^{\infty} X(f_k) e^{2\pi i f_k t} \Delta f$$

As T grows indefinitely, the points f_k become infinitely dense and we obtain

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{2\pi i f t} \, \mathrm{d}f \tag{8.69a}$$

Now let's return to (8.68a) which we neglected for a while. Actually, we only have to replace kF with f_k and then f_k with f. Not forgetting that $T \to \infty$, we obtain

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} \, \mathrm{d}t \tag{8.69b}$$

and we pretty much obtained both transform formulas we were looking for.

Rewriting (8.69a) and (8.69b) once again in the "correct order", we have

(Fourier transform)

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} \, \mathrm{d}t \tag{8.70a}$$

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{2\pi i f t} \, \mathrm{d}f \tag{8.70b}$$

Equations (8.70) specify the *direct and inverse Fourier transform* respectively. As with Fourier series and discrete-time Fourier transform, the convergence conditions are complicated (actually even more than in the cases of the the former two).

▷ Similarly to the direct DTFT formula (8.58a) and the Fourier series formula (8.64b), the integrals (8.70) are understood in the Cauchy principal value sense:

$$X(f) = \text{p.v.} \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} \, \mathrm{d}t = \lim_{T \to \infty} \int_{-T}^{T} x(t) e^{-2\pi i f t} \, \mathrm{d}t$$

$$x(t) = \text{p.v.} \int_{-\infty}^{\infty} X(f) e^{2\pi i f t} \, \mathrm{d}f = \lim_{F \to \infty} \int_{-F}^{F} X(f) e^{2\pi i f t} \, \mathrm{d}f$$

It remains to obtain the cyclic frequency version of (8.70), where, surprisingly, we get 3 different variants, differing by the $1/2\pi$ normalization placement. Substituting $f = \omega/2\pi$ into (8.70) yields

(Fourier transform, density-like cyclic-frequency form)

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} \, \mathrm{d}t \tag{8.71a}$$

$$x(t) = \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} \, \mathrm{d}\omega \tag{8.71b}$$

The coefficient $1/2\pi$ can also be moved to the direct transform, yielding

(Fourier transform, amplitude-like cyclic-frequency form)

$$X(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} x(t) e^{-i\omega t} \, \mathrm{d}t \tag{8.72a}$$

$$x(t) = \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} \, \mathrm{d}\omega \tag{8.72b}$$

And of course there is the symmetric version

(Fourier transform, symmetric cyclic-frequency form)

$$X(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t) e^{-i\omega t} \, \mathrm{d}t \tag{8.73a}$$

$$x(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} \, \mathrm{d}\omega \tag{8.73b}$$

▷ As with discrete-time Fourier transform, $X(\omega)$ is a different function than $X(f)$. More rigorously, we should have used different notations, e.g. $X_\omega(\cdot)$ and $X_f(f)$. Actually $X(\omega)$ is a different function in each of (8.71), (8.72) and (8.73), where in (8.71) we have $X_\omega(2\pi f) = X_f(f)$, in (8.72) we have $X_\omega(2\pi f) = X_f(f)/2\pi$ and in (8.73) we have $X_\omega(2\pi f) = X_f(f)/\sqrt{2\pi}$.

Differently from (8.21), (8.22) and (8.23), the variants (8.71), (8.72) and (8.73) cannot be really referred to as "density", "amplitude" and "symmetric" versions respectively, because all of them are variants of the density version, since the actual amplitudes of the components are given by $X(\omega)\,\mathrm{d}\omega$ (up to an optional coefficient of $1/2\pi$ or $1/\sqrt{2\pi}$). We can still, however, refer to (8.71) as "density-like", (8.72) as "amplitude-like" and (8.73) as "symmetric" versions of Fourier transform.

Actually, there is not that much difference between (8.71) and (8.72), it's just that in the latter the density is specified in the ordinary frequency scale, while the former, in case the coefficient $1/2\pi$ is written in front of the integral, should be understood as using the cyclic frequency-scale density $X(\omega)$, which respectively requires the division of the result by 2π. However, as with discrete-time Fourier transform, the coefficient $1/2\pi$ in front of (8.71b) can be "glued" with the differential $\mathrm{d}\omega$, yielding $\frac{\mathrm{d}\omega}{2\pi}$, which can be simply understood as integration over ordinary rather than cyclic frequency, and from that perspective $X(\omega)$ is just the ordinary frequency-scale density expressed as a function of ω.

The variant (8.73) (which still can be referred to as "symmetric", just in a different sense) however still has its own special benefits. Most importantly, it represents a truly unitary transformation (of an infinitely-dimensional linear space), and thereby, same as (8.22), preserves the lengths and other dot products. For that reason, it is the preferred variant in quantum mechanics. We shall discuss the Fourier transform properties arising out of unitarity in Section 8.10.

▷ The readers are asked to pay attention to which of the three different normalizations (8.71), (8.72) and (8.73) is assumed at any given time. There are no notational differences and the type of normalization used needs to be known from the context. Often, integration being done over $\mathrm{d}\omega/2\pi$ can be a good hint suggesting (8.71), although it needs to be taken with caution.

Goal 8.9 completed. ■

We have thus introduced

The 4 types of Fourier transform:

- discrete Fourier transform/series: the function is discrete and periodic in both domains
- discrete-time Fourier transform: the function is discrete infinitely long in the original domain and continuous periodic in the transform domain
- Fourier series: the function is continuous periodic in the original domain and discrete infinitely long in the transform domain
- Fourier transform: the function is continuous infinitely long in both domains

Thereby the function is discrete in one domain iff it's periodic in the dual domain.

Goal 8.10. Obtain a real form of the integral Fourier transform.

To obtain the real version of Fourier transform, assume real $x(t)$. Thereby (8.70a) yields a Hermitian function $X(-f) = X^*(f)$. Respectively we can rewrite (8.70b) as

$$\begin{aligned} x(t) &= \int_{-\infty}^{\infty} X(f)e^{2\pi ift}\,\mathrm{d}f = \int_0^{\infty} \big(X(f)e^{2\pi ift} + X(-f)e^{-2\pi ift}\big)\,\mathrm{d}f \\ &= \int_0^{\infty} \Big(X(f)e^{2\pi ift} + \big(X(f)e^{2\pi ift}\big)^*\Big)\,\mathrm{d}f = 2\int_0^{\infty} \mathrm{Re}\big(X(f)e^{2\pi ift}\big)\,\mathrm{d}f \\ &= 2\int_0^{\infty} \Big(|X(f)| \cdot \cos\big(2\pi ft + \arg X(f)\big)\Big)\,\mathrm{d}f \end{aligned}$$

finally yielding

(real-form Fourier transform)

$$x(t) = \int_0^{\infty} a(f)\cos\big(2\pi ft + \varphi(f)\big)\,\mathrm{d}f \tag{8.74}$$

where $a(f) = 2 \cdot |X(f)|$, $\varphi(f) = \arg X(f)$, or, in terms of ω

(real-form Fourier transform, cyclic-frequency version)

$$x(t) = \int_0^\infty a(\omega)\cos\bigl(\omega t + \varphi(\omega)\bigr)\,\frac{\mathrm{d}\omega}{2\pi} \tag{8.75}$$

(where $a(\omega)$ and $\varphi(\omega)$ have been stretched 2π times compared to $a(f)$ and $\varphi(f)$).

▷ There is no separable term corresponding to a_0 from the real Fourier series in (8.74) and (8.75). In real Fourier series the term a_0 represents a non-zero average value of $x(t)$ on its period (in signal processing terms this value is known as "DC offset"). However, in case of Fourier transform the infinitely long $x(t)$ cannot have a non-zero average, otherwise the integral (8.70a) will diverge.

The value $X(0)$ per se, even if it is non-zero, contributes only an infinitely small amount to the entire integral and therefore doesn't result in a need for a separate term like $a(0)$.

▷ Sometimes it is possible and even desirable for the infinitely long $x(t)$ in Fourier transform to have a "non-zero average". This is achieved by introducing the *Dirac delta* function, discussed in Section 8.13. However this goes beyond the Fourier transform basics which we are considering here.

Goal 8.10 completed. ■

8.10 Unitarity

Having introduced all 4 types of Fourier transform, we are going to start looking into various properties of Fourier transform, where we shall try to compare how these properties look with different Fourier transform types. We have mentioned that truly unitary versions of Fourier transform preserve the length and dot products and this is what we shall concentrate on in this section.

Let's return to the matrix notation (8.16) of the symmetric version of the discrete Fourier transform. Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors and let $\mathbf{a} = N^{-1/2}\mathbf{H}^*\cdot\mathbf{x}$ and $\mathbf{b} = N^{-1/2}\mathbf{H}^*\cdot\mathbf{y}$ be their transformed versions. If we consider the dot product $\mathbf{x}^*\cdot\mathbf{y}$ of $\mathbf{x}$ and $\mathbf{y}$ (where we have to use the Hermitian transpose $\mathbf{x}^*$ instead of simple transpose $\mathbf{x}^\mathrm{T}$ since we are dealing with complex numbers[4]), this product will be unchanged after multiplying each of the vectors by the unitary matrix $N^{-1/2}\mathbf{H}^*$. Therefore

$$\mathbf{x}^*\cdot\mathbf{y} = \mathbf{a}^*\cdot\mathbf{b} \tag{8.76a}$$

▷ Same as elsewhere in this book, boldface implies matrix notation, therefore the dot product of two vectors is notated as $\mathbf{x}^*\cdot\mathbf{y}$ (or $\mathbf{x}^\mathrm{T}\cdot\mathbf{y}$) rather than $\mathbf{x}\cdot\mathbf{y}$, the latter expression denoting a product of two $N\times 1$ matrices, thereby being undefined.

▷ Given a unitary matrix $\mathbf{U}$: $\mathbf{U}^{-1} = \mathbf{U}^*$,

$$(\mathbf{U}\cdot\mathbf{x})^*\cdot(\mathbf{U}\cdot\mathbf{y}) = \mathbf{x}^*\cdot\mathbf{U}^*\cdot\mathbf{U}\cdot\mathbf{y} = \mathbf{x}^*\cdot\mathbf{U}^{-1}\cdot\mathbf{U}\cdot\mathbf{y} = \mathbf{x}^*\cdot\mathbf{y}$$

[4]Some further discussion of the need to use Hermitian transpose in dot product is found in Section 3.6.

In case of two equal vectors $\mathbf{x} = \mathbf{y}$, (8.76a) turns into

$$\mathbf{x}^* \cdot \mathbf{x} = \mathbf{a}^* \cdot \mathbf{a}$$

and since $\mathbf{x}^* \cdot \mathbf{x} = |\mathbf{x}|^2$ and $\mathbf{a}^* \cdot \mathbf{a} = |\mathbf{a}|^2$ we obtain the property of the preservation of the vector length by Fourier transform:

$$|\mathbf{x}|^2 = |\mathbf{a}|^2 \tag{8.76b}$$

▷ In the context of Fourier transform the length preservation property is often written in terms of squared lengths as in (8.76b), because squared lengths often have more useful semantics in specific applications of Fourier transform. However, of course, it can also be written as

$$|\mathbf{x}| = |\mathbf{y}|$$

The property (8.76a), often including its particular version (8.76b), is referred to as *Parseval's theorem* (or also as *Plancherel theorem*).

Discrete Fourier transform

We would like now to reexpress (8.76) in the functional notation, which is more in line with transform formulas like (8.22) and (8.73). In order to do so we need to introduce a "dot product of functions". A "dot product of functions x and y" is more commonly referred to as an *inner product* of functions x and y and notated as $\langle x, y \rangle$. Thus, the dot product

$$\mathbf{x}^* \cdot \mathbf{y} = \sum_{n=0}^{N-1} x_n^* y_n \tag{8.77a}$$

in terms of functions $x[n]$, $y[n]$ is notated as

$$\langle x, y \rangle = \sum_{n=0}^{N-1} x^*[n] y[n] \tag{8.77b}$$

Notice the lack of complex conjugation or Hermitian transpose symbol in $\langle x, y \rangle$: the inner product notation already assumes that this conjugation is performed "under the hood".

In terms of the inner product (8.77b), the property (8.76a) becomes written as

$$\langle x, y \rangle = \langle X, Y \rangle \tag{8.78a}$$

and respectively (8.76b) becomes written as

$$\|x\|^2 = \|X\|^2 \tag{8.78b}$$

where the norm $\|x\|$ is defined to have the "length" semantics:

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{n=0}^{N-1} \big|x[n]\big|^2} \tag{8.79}$$

Writing (8.78) out explicitly we obtain

(Parseval's theorem for symmetric-version DFT)

$$\sum_{n=0}^{N-1} x^*[n]y[n] = \sum_{k=0}^{N-1} X^*[k]Y[k] \tag{8.80a}$$

$$\sum_{n=0}^{N-1} \left|x[n]\right|^2 = \sum_{k=0}^{N-1} \left|X[k]\right|^2 \tag{8.80b}$$

Notice that (8.78) and (8.80) assume the symmetric version (8.22) of the discrete Fourier transform, which ensures a truly unitary transformation. In other versions of discrete Fourier transform, the formulas (8.78) and (8.80) are going to look different. Specifically, in the amplitude version (8.21), for a given pair of $x[n]$ and $y[n]$, the functions $X[k]$ and $Y[k]$ obtained by (8.21a) are $\sqrt{N}$ times smaller compared to (8.22), hence their product is N times smaller and (8.78) and (8.80) become

$$\frac{1}{N}\langle x, y\rangle = \langle X, Y\rangle \tag{8.81a}$$

$$\frac{1}{N}\|x\|^2 = \|X\|^2 \tag{8.81b}$$

and

(Parseval's theorem for amplitude-version DFT)

$$\frac{1}{N}\sum_{n=0}^{N-1} x^*[n]y[n] = \sum_{k=0}^{N-1} X^*[k]Y[k] \tag{8.82a}$$

$$\frac{1}{N}\sum_{n=0}^{N-1} \left|x[n]\right|^2 = \sum_{k=0}^{N-1} \left|X[k]\right|^2 \tag{8.82b}$$

respectively. Conversely, for the density version (8.23) we obtain

$$\langle x, y\rangle = \frac{1}{N}\langle X, Y\rangle \tag{8.83a}$$

$$\|x\|^2 = \frac{1}{N}\|X\|^2 \tag{8.83b}$$

and

(Parseval's theorem for density-version DFT)

$$\sum_{n=0}^{N-1} x^*[n]y[n] = \frac{1}{N}\sum_{k=0}^{N-1} X^*[k]Y[k] \tag{8.84a}$$

$$\sum_{n=0}^{N-1} \left|x[n]\right|^2 = \frac{1}{N}\sum_{k=0}^{N-1} \left|X[k]\right|^2 \tag{8.84b}$$

▷ Notice that the placement of the $1/N$ normalization factor in all of the versions of Parseval's theorem listed above coincides with the placement of the $1/N$ normalization factor in the transform itself. E.g. in the amplitude version of the transform (8.21) the $1/N$ factor is placed in front of the sum involving $x[n]$ (in the direct transform), and so it is in the amplitude version of Parseval's theorem. The same applies to the density version. One can also think of the symmetric version case as having a $1/\sqrt{N}$ normalization factor on each side, these equal factors simply being cancelled. This can be used as some kind of mnemonic hint to remember the positioning of the $1/N$ factor.

Technically, the $1/N$ factor appearing in Parseval's theorem emerges as a product of two additional $1/\sqrt{N}$ factors (appearing in comparison to the symmetric version), rather than the "original" $1/N$ factor from the transform formulas themselves.

▷ The different options to place the $1/N$ coefficient in Parseval's theorem may motivate a modification to the definition of the inner product (8.77b), which may be made to include the $1/N$ coefficient. This will effectively "change the units of length" by a factor of $1/\sqrt{N}$.

One could also consider letting the $1/N$ factor appear or not appear in the inner product depending on the domain in which the inner product is computed, effectively "using different units of length" in different domains.

Goal 8.11. Extend the just obtained detailed results about the unitarity property to the other 3 types of Fourier transform. *Hint:* take the limits (taken in the derivation of these 3 types) again.

Discrete-time Fourier transform

In case of discrete-time Fourier transform the two domains are highly asymmetric: in one we are having an infinitely long discrete-argument function, in the other one we are having a continuous-argument periodic function. So the inner product of two functions is probably going to attain some other form than (8.77b) in each of those cases.

Let's recall that discrete-time Fourier transform is obtained as a limiting case of the density version (8.23) of discrete Fourier transform. This suggests that we might be able to obtain Parseval's theorem for discrete-time Fourier transform as a limiting case of the density version (8.84a) of Parseval's theorem for discrete Fourier transform.

On the left-hand side of (8.84a) we have the inner product (8.77b) defined for two discrete-argument periodic functions. Let's see what happens with it at $N \to \infty$. Actually, we have to remember to switch to the bilateral version first, as we did with discrete-time Fourier transform, yielding

$$\langle x, y \rangle = \sum_{n=-N/2}^{N/2} x^*[n] y[n] \tag{8.85}$$

and then let $N \to \infty$, yielding

$$\langle x, y \rangle = \sum_{n=-\infty}^{\infty} x^*[n] y[n] \tag{8.86}$$

Remember that in case of discrete-time Fourier transform we usually assume that $x[\cdot]$ and $y[\cdot]$ decay sufficiently quickly at $n \to \infty$, therefore there are good chances that the infinite sum in (8.86) will converge.

On the right-hand side of (8.84a) we also have the inner product (8.77b), this time divided by N. Replacing $1/N$ by $\mathrm{d}f$ and $X[k]$ and $Y[k]$ by $X(f)$ and $Y(f)$, in the limit at $N \to \infty$ we obtain

$$\langle X, Y \rangle = \int_0^1 X^*(f) Y(f) \, \mathrm{d}f \tag{8.87a}$$

or, in the bilateral form

$$\langle X, Y \rangle = \int_{-1/2}^{1/2} X^*(f) Y(f) \, \mathrm{d}f \tag{8.87b}$$

Apparently the two forms (8.87a) and (8.87b) are identical due to 1-periodicity of $X(f)$ and $Y(f)$.

The Parseval's theorem formulas (8.84) are thereby attaining the form

(Parseval's theorem for DTFT)

$$\langle x, y \rangle = \sum_{n=-\infty}^{\infty} x^*[n] y[n] = \int_{-1/2}^{1/2} X^*(f) Y(f) \, \mathrm{d}f = \langle X, Y \rangle \tag{8.88a}$$

$$\|x\|^2 = \sum_{n=-\infty}^{\infty} \big|x[n]\big|^2 = \int_{-1/2}^{1/2} \big|X(f)\big|^2 \, \mathrm{d}f = \|X\|^2 \tag{8.88b}$$

▷ If $X(\cdot)$ and $Y(\cdot)$ are specified in the ω scale as in (8.59), then we have to remember to still integrate with respect to the ordinary frequency $\frac{\mathrm{d}\omega}{2\pi}$ in (8.87). Alternatively, we could integrate with respect to $\mathrm{d}\omega$, but then we should divide the right-hand sides of (8.88) by 2π.

Fourier series

Fourier series is obtained as a limiting case of the amplitude version (8.21) of discrete Fourier transform. Respectively, we shall obtain Parseval's theorem for Fourier series as a limiting case of the amplitude version (8.82a) of Parseval's theorem for discrete Fourier transform.

Starting with (8.82a) we do pretty much the same steps as we did starting with (8.84a) in the derivation of Parseval's theorem for discrete-time Fourier transform. The only essential difference is that the left- and right-hand sides are swapped compared to the previous time, and respectively in place of (8.88a) we obtain

$$\langle x, y \rangle = \int_{-1/2}^{1/2} x^*(t) y(t) \, \mathrm{d}t = \sum_{k=-\infty}^{\infty} X^*[k] Y[k] = \langle X, Y \rangle \tag{8.89}$$

for 1-periodic $x(t)$ and $y(t)$. Switching from 1-periodic to T-periodic $x(t)$ and $y(t)$ we notice that this switch shouldn't affect the sum in the right-hand side (8.89), while we need to replace $\mathrm{d}t$ by $\mathrm{d}t/T$ to leave the integral in the left-hand side unchanged, yielding

(Parseval's theorem for Fourier series)

$$\langle x, y\rangle = \frac{1}{T}\int_{-T/2}^{T/2} x^*(t)y(t)\,\mathrm{d}t = \sum_{k=-\infty}^{\infty} X_k^* Y_k = \langle X, Y\rangle \tag{8.90a}$$

$$\|x\|^2 = \frac{1}{T}\int_{-T/2}^{T/2} \big|x(t)\big|^2\,\mathrm{d}t = \sum_{k=-\infty}^{\infty} \big|X_k\big|^2 = \|X\|^2 \tag{8.90b}$$

(where we also switched the notation from $X[k]$ and $Y[k]$ to X_k and Y_k).

▷ As in the previous cases, the placement of the normalization coefficient in the Parseval's theorem formula for Fourier series matches the placement of that coefficient in the transform formulas themselves. Specifically, the $1/T$ coefficient present in the direct transform formula (8.64a) of the Fourier series is also present on the left side of (8.90). Notice that this time, differently from (8.81), we incorporated this coefficient into the inner product itself:

$$\langle x, y\rangle = \frac{1}{T}\int_{-T/2}^{T/2} x^*(t)y(t)\,\mathrm{d}t$$

but that's essentially just a notational convention.

Fourier transform

Fourier transform can be obtained as a limiting case of Fourier series, therefore we shall derive Parseval's theorem for Fourier transform as a limiting case of Parseval's theorem for Fourier series.

Recalling that Fourier transform is obtained as a limiting case of the "density version" of Fourier series, we need to obtain the density version of (8.90a). In the density version we don't have the normalization factor $1/T$ in front of the integral in the direct transform formula (8.64a), therefore X_k and Y_k become T times larger each, compared to the "amplitude version". Therefore the product $X_k^* Y_k$ becomes T^2 times larger and we need to compensate by dividing the right-hand side of (8.90a) by T^2:

$$\frac{1}{T}\int_{-T/2}^{T/2} x^*(t)y(t)\,\mathrm{d}t = \frac{1}{T^2}\sum_{k=-\infty}^{\infty} X_k^* Y_k$$

or simply

$$\int_{-T/2}^{T/2} x^*(t)y(t)\,\mathrm{d}t = \frac{1}{T}\sum_{k=-\infty}^{\infty} X_k^* Y_k$$

which is the "density version" counterpart of (8.90a). Recalling that $T = 1/F$, where F is the base frequency, we introduce $\Delta f = f_{k+1} - f_k = F = 1/T$ (where $f_k = kF = k/T$), as well as continuous-time functions $X(f_k) = X_k$, $Y(f_k) = Y_k$, yielding

$$\int_{-T/2}^{T/2} x^*(t)y(t)\,\mathrm{d}t = \sum_{k=-\infty}^{\infty} X^*(f_k)Y(f_k)\Delta f$$

and taking a limit at $T \to \infty$ yields

$$\int_{-\infty}^{\infty} x^*(t)y(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} X^*(f)Y(f)\mathrm{d}f$$

Thus, the Fourier transform counterpart of (8.90) is

(Parseval's theorem for Fourier transform)

$$\langle x, y\rangle = \int_{-\infty}^{\infty} x^*(t)y(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} X^*(f)Y(f)\,\mathrm{d}f = \langle X, Y\rangle \tag{8.91a}$$

$$\|x\|^2 = \int_{-\infty}^{\infty} \left|x(t)\right|^2 \mathrm{d}t = \int_{-\infty}^{\infty} \left|X(f)\right|^2 \mathrm{d}f = \|X\|^2 \tag{8.91b}$$

Reexpressing (8.91) in terms of ω and noticing that $\mathrm{d}f = \mathrm{d}\omega/2\pi$ yields

(Parseval's theorem for density-like cyclic-frequency version of Fourier transform)

$$\int_{-\infty}^{\infty} x^*(t)y(t)\,\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} X^*(\omega)Y(\omega)\,\mathrm{d}\omega \tag{8.92a}$$

$$\int_{-\infty}^{\infty} \left|x(t)\right|^2 \mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|X(\omega)\right|^2 \mathrm{d}\omega \tag{8.92b}$$

for the "density-like" version (8.71) of Fourier transform. For other versions of Fourier transform we obtain similar results, the placement of the $1/2\pi$ coefficient matching the one in the transform formulas, as usual. For the symmetric version we thus obtain $1/\sqrt{2\pi}$ on both sides of the equality, which cancel each other and we're left with a formula without extra normalization coefficients, similar to (8.91).

Goal 8.11 completed. ■

8.11 Convolution

In this section we shall (re-)discover the convolution property of Fourier transform, starting pretty much from scratch. The initial motivation for this discovery might be coming from signal processing, although maybe there are also other motivations. We are going to use the signal processing context to set up the motivation for the discussion.

If $x[n]$ or $x(t)$ is a signal, then $X[k]$ or $X(f)$ represents its frequency spectrum. Indeed, $X[k]$ or $X(f)$ are complex amplitudes (or amplitude densities) of the sinusoidal components of various frequencies making up the signal. We might therefore be interested in performing various operations on the signal's spectrum. One of the most prominent of such operations is filtering, where we might want to dampen or completely remove some of the sinusoidal components from the signal. Mathematically this can be described by multiplying the spectrum $X[k]$ or $X(f)$ with some other function $Y[k]$ or $Y(f)$ respectively,

this latter function defining how much to dampen (or to boost) each of the frequencies.[5] We would be interested in the time-domain description of the same process.

Goal 8.12. Find the Fourier transform dual-domain counterparts of multiplication of functions (for all 4 Fourier tranform types). *Hint:* use the transform formulas.

Discrete Fourier transform

Let $x[n]$ and $y[n]$ be two N-periodic discrete-time signals and let $X[k]$ and $Y[k]$ be their spectra. We are interested in the signal corresponding to the spectrum $X[k]Y[k]$. Let's see if we can rewrite the expression $X[k]Y[k]$ so that it looks like a discrete Fourier transform of some signal. By (8.21a)

$$\begin{aligned} X[k]Y[k] &= \left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]e^{-2\pi ikn/N}\right)\left(\frac{1}{N}\sum_{m=0}^{N-1} y[m]e^{-2\pi ikm/N}\right) \\ &= \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} x[n]y[m]e^{-2\pi ikn/N}e^{-2\pi ikm/N} \\ &= \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} x[n]y[m]e^{-2\pi ik(n+m)/N} \end{aligned} \tag{8.93a}$$

Introducing $l = n + m$ and recalling that by periodicity we can sum over any interval of length N, we further transform (8.93a) as

$$\begin{aligned} X[k]Y[k] &= \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1} x[n]y[l-n]e^{-2\pi ikl/N} \\ &= \frac{1}{N^2}\sum_{l=0}^{N-1}\sum_{n=0}^{N-1} x[n]y[l-n]e^{-2\pi ikl/N} \\ &= \frac{1}{N}\sum_{l=0}^{N-1}\left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]y[l-n]\right)e^{-2\pi ikl/N} \\ &= \frac{1}{N}\sum_{l=0}^{N-1} u[l]e^{-2\pi ikl/N} \end{aligned} \tag{8.93b}$$

where $u[l]$ is the signal equal to the expression in the parentheses in the line above.

▷ More specifically, what we did in (8.93b) was

$$\begin{aligned} \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} x[n]y[m]e^{-2\pi ik(n+m)/N} &= \sum_{n=0}^{(N-1)}\sum_{l=n}^{(n+N-1)} x[n]y[l-n]e^{-2\pi ikl/N} \\ &= \sum_{n=0}^{N-1}\sum_{l=0}^{N-1} x[n]y[l-n]e^{-2\pi ikl/N} \end{aligned}$$

where we replaced the summation range going from n to $n+N-1$ by the one going from 0 to $N-1$ due to N-periodicity of the expression under the summation sign.

[5] $Y[k]$ or $Y(f)$ is referred to as the *frequency response* of the respective filter.

Now (8.93b) is discrete Fourier transform of $u[\cdot]$. This means that the spectrum $U[k] = X[k]Y[k]$ corresponds to the signal

$$u[n] = \frac{1}{N}\sum_{j=0}^{N-1} x[j]y[n-j]$$

The expression

(discrete-argument periodic convolution)

$$x * y = \sum_{j=0}^{N-1} x[j]y[n-j] \tag{8.94}$$

is referred to as *convolution* of two N-periodic discrete-time signals $x[\cdot]$ and $y[\cdot]$, or, more precisely, as *discrete-time periodic convolution* of these two signals.

▷ Sometimes a subscript on the convolution sign might be used to highlight the periodicity: $x *_N y$. Also sometimes the normalization by $\frac{1}{N}$ might be implicitly included into the convolution operation, that is the right-hand side of (8.94) contains an additional factor of $\frac{1}{N}$.

▷ A more formally correct notation for (8.94) is

$$(x * y)[n] = \sum_{j=0}^{N-1} x[j]y[n-j] \tag{8.95}$$

The notation (8.95) is understood in the following sense. Let $u[n]$ be the function obtained by the convolution of x and y:

$$u[n] = \sum_{j=0}^{N-1} x[j]y[n-j]$$

Then the notation $x * y$ is formally the same as u, that is it denotes the obtained function itself. On the other hand $(x * y)[n]$ is formally the same as $u[n]$, that is it can be understood as the function itself, but also as specifically the value of the function u at the point n.

In this text we will often omit the explicitly written argument after the convolution operation (as in $(x * y)[n]$) and go with the somewhat handwavy notation $x * y$ as used in (8.94), or sometimes even $x[n] * y[n]$, all these notations meaning the same.

Thus the product of Fourier transforms corresponds to the convolution of the original signals:

(DFT convolution property)

$$U[k] = X[k]Y[k] \iff u[n] = \frac{1}{N}(x * y) \tag{8.96}$$

Notice how, similarly to Parseval's theorem, the placement of the $1/N$ normalization coefficient in the property (8.96) is the same as its placement in the transform formulas (8.21).

▷ By the convolution property (8.96) of the discrete Fourier transform it immediately follows that N-periodic convolution is commutative ($x * y = y * x$) since the product $X[k]Y[k]$ commutes. This property of the convolution operation could also be demonstrated explicitly, which the readers might take as an exercise, if they wish.

To obtain the dual of (8.96) we can do transformations similar to (8.93), just this time applying (8.21b) to the product $x[n]y[n]$. The only differences are that this time the signs in the exponents are positive and there are no $1/N$ factors. Thus we can already see that the result should be

$$x[n]y[n] = \sum_{k=0}^{N-1} U[k]e^{2\pi ikn/N}$$

where

$$U[k] = \sum_{j=0}^{N-1} X[j]Y[k-j] = X * Y$$

Thus the dual counterpart of (8.96) is

(DFT dual convolution property)

$$u[n] = x[n]y[n] \iff U[k] = X * Y \tag{8.97}$$

▷ We could have obtained (8.97) from (8.96) by a direct application of the duality property. However the process is going to be a bit involved, as we are going to need to apply the duality twice, for each of the signals in turn.

This process however can be largely simplified, if we ignore the appearance of the normalization coefficients, simply replacing $x[n]$ with $X[n]$ and $X[k]$ with $x[-k]$ and do the same with y (so that the left-hand side of (8.96) turns into $u[-k] = x[-k]y[-k]$). The correct placement of the normalization coefficients in the result can be obtained by substituting some example signals for x and y and choosing the normalization coefficients in such a way that the equalities hold. A particularly good signal for this purpose is often the delta function discussed in Section 8.13. Using the delta function to find the normalization coefficient value is discussed in Subsection 8.13.11.

▷ Properties (8.96) and (8.97) were obtained for the amplitude version (8.21) of the discrete Fourier transform. The readers may obtain these properties for the other versions of discrete Fourier transform as an exercise, if they wish. Notice that the only differences to (8.96) and (8.97) should be in the normalization.

Discrete-time Fourier transform

In case of the discrete-time Fourier transform the spectra of the non-periodic signals $x[n]$ and $y[n]$ turn into continuous amplitude densities $X(f)$ and $Y(f)$. Doing the steps similar to (8.93), we respectively obtain (by (8.58a))

$$X(f)Y(f) = \left(\sum_{n=-\infty}^{\infty} x[n]e^{-2\pi ifn}\right)\left(\sum_{m=-\infty}^{\infty} y[m]e^{-2\pi ifm}\right)$$

$$= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} x[n]y[m]e^{-2\pi i f(n+m)}$$

$$= \sum_{l=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x[n]y[l-n]e^{-2\pi i f l} = \sum_{l=-\infty}^{\infty} u[l]e^{-2\pi i f l} \tag{8.98}$$

where $l = n + m$ and

$$u[n] = \sum_{j=-\infty}^{\infty} x[j]y[n-j]$$

▷ More specifically, what we did in (8.98) was

$$\sum_{n=-\infty}^{\infty} \left(x[n]e^{-2\pi i f n} \left(\sum_{m=-\infty}^{\infty} y[m]e^{-2\pi i f m} \right) \right)$$

$$= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} x[n]y[m]e^{-2\pi i f(n+m)} = \sum_{n=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} x[n]y[l-n]e^{-2\pi i f l}$$

$$= \sum_{l=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x[n]y[n-l]e^{-2\pi i f l}$$

thus we assumed that the convergence conditions of the sums are "sufficiently good" to justify the change of the summation order.

The expression

(discrete-argument convolution)

$$x * y = \sum_{j=-\infty}^{\infty} x[j]y[n-j] \tag{8.99}$$

is referred to as *convolution* of two nonperiodic discrete-time signals $x[\cdot]$ and $y[\cdot]$.

▷ There is no difference in notation between periodic and nonperiodic convolution (except for the optional subscript N on the periodic convolution sign, which we agreed not to use here). Thus, the type of the convolution implied by the expression $x * y$ needs to be derived from the kind of functions involved.

The last expression in (8.98) is discrete-time Fourier transform of $u[\cdot]$. The property (8.96) thereby becomes

(DTFT convolution property)

$$U(f) = X(f)Y(f) \iff u[n] = x * y \tag{8.100}$$

▷ The property (8.100) implies that (nonperiodic) convolution of two functions of discrete argument is commutative: $x * y = y * x$. The commutativity property can also be obtained directly from (8.99).

We are going to postpone the obtaining of the dual of (8.100) until the discussion of convolution in the context of Fourier series, where we are going

to find that (8.100) has two duals: the formal one (8.106), which is a property of the Fourier series, and the actual one (8.105), which is a property of the discrete-time Fourier transform.

Fourier series

Given T-periodic $x(t)$ and $y(t)$ and their spectra X_k and Y_k, we perform the steps similar to (8.93). By (8.64a)

$$\begin{aligned} X_k Y_k &= \left(\frac{1}{T}\int_{-T/2}^{T/2} x(t)e^{-2\pi ikFt}\,\mathrm{d}t\right)\left(\frac{1}{T}\int_{-T/2}^{T/2} y(\tau)e^{-2\pi ikF\tau}\,\mathrm{d}\tau\right) \\ &= \frac{1}{T}\int_{-T/2}^{T/2}\left(\frac{1}{T}\int_{-T/2}^{T/2} x(t)y(\tau)e^{-2\pi ikF(t+\tau)}\,\mathrm{d}\tau\right)\mathrm{d}t \\ &= \frac{1}{T}\int_{-T/2}^{T/2}\left(\frac{1}{T}\int_{-T/2}^{T/2} x(t)y(t'-t)e^{-2\pi ikFt'}\,\mathrm{d}t\right)\mathrm{d}t' \qquad (8.101) \\ &= \frac{1}{T}\int_{-T/2}^{T/2} u(t')e^{-2\pi ikFt'}\,\mathrm{d}t' \qquad (8.102) \end{aligned}$$

where $t' = t + \tau$ and where we used the periodicity to justify a change of the integration range to another range of length T, and where

$$u(t) = \frac{1}{T}\int_{-T/2}^{T/2} x(\tau)y(t-\tau)\,\mathrm{d}\tau$$

The expression

(periodic convolution)

$$x * y = \int_{-T/2}^{T/2} x(\tau)y(t-\tau)\,\mathrm{d}t \qquad (8.103)$$

is referred to as *convolution* of two continuous-time T-periodic signals $x(\cdot)$ and $y(\cdot)$, or, more precisely, as *continuous-time periodic convolution* of these two signals.

The last expression in (8.102) gives the k-th Fourier series coefficient for $u(t)$. Thus

(Fourier series convolution property)

$$U_k = X_k Y_k \iff u(t) = \frac{1}{T}(x * y) \qquad (8.104)$$

▷ The property (8.104) implies that convolution of two periodic functions of a continuous argument is commutative: $x * y = y * x$. The commutativity property can also be obtained directly from (8.103).

Now let's obtain the dual of (8.104). Actually, there are two different duals. One is a formal dual, which can be obtained by replacing $x(t)$ with $X(t)$ and

X_k with $x[-k]$ and doing the same for y and u, yielding

$$u[-k] = x[-k]y[-k] \iff U(t) = \frac{1}{T}(X * Y)$$

Thereby the obtained dual is actually a property of discrete-time Fourier transform, since we have gotten infinite nonperiodic time-domain signals and periodic frequency-domain signals. This should have been expected, since the discrete-time Fourier transform is the dual of the Fourier series. Replacing t with f and $-k$ with n and noticing that, by convention, the frequency-domain period with discrete-time Fourier transform is unity, we let $T = 1$, obtaining

(DTFT dual convolution property)

$$u[n] = x[n]y[n] \iff U(f) = X * Y \tag{8.105}$$

as the "actual dual" of (8.100).

The "actual dual" of (8.104) can be obtained as the formal dual of the discrete-time Fourier transform's property (8.100), where we get

$$u(-f) = x(-f)y(-f) \iff U_n = X * Y$$

Switching the argument notation yields

(Fourier series dual convolution property)

$$u(t) = x(t)y(t) \iff U_k = X * Y \tag{8.106}$$

but, formally, only under the assumption of a unit period $T = 1$ of x and y. It's not difficult to realize, however, that if (8.106) holds for $T = 1$, then it holds for any other T, since we can always shrink $x(t)$ and $y(t)$ to a unit period without affecting their Fourier series coefficients.

▷ The argument about normalization of (8.105) and (8.106) based on the duality transformation of the period T might feel somewhat handwavy. The readers, if they wish, may obtain these properties explicitly, or wait for Subsection 8.13.11 to check the normalization using the delta function.

Fourier transform

Given non-periodic $x(t)$ and $y(t)$ and their spectra $X(f)$ and $Y(f)$, we perform the steps following the ideas of (8.98) and (8.102). By (8.70a)

$$\begin{aligned} X(f)Y(f) &= \left(\int_{-\infty}^{\infty} x(t)e^{-2\pi i f t}\,\mathrm{d}t\right)\left(\int_{-\infty}^{\infty} x(\tau)e^{-2\pi i f \tau}\,\mathrm{d}\tau\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(t)y(\tau)e^{-2\pi i f(t+\tau)}\,\mathrm{d}\tau\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(t)y(t'-t)e^{-2\pi i f t'}\,\mathrm{d}t\,\mathrm{d}t' \\ &= \int_{-\infty}^{\infty} u(t')e^{-2\pi i f t'}\,\mathrm{d}t' \end{aligned} \tag{8.107}$$

where $t' = t + \tau$, and where we assumed that the improper integral convergence is "sufficiently good", and where

$$u(t) = \int_{-\infty}^{\infty} x(\tau)y(t-\tau)\,\mathrm{d}\tau$$

The expression

(convolution)

$$x * y = \int_{-\infty}^{\infty} x(\tau)y(t-\tau)\,\mathrm{d}\tau \tag{8.108}$$

is referred to as *convolution* of two nonperiodic continuous signals $x(\cdot)$ and $y(\cdot)$.

The last expression in (8.107) is Fourier transform of $u(t)$. Thus

(Fourier transform convolution property)

$$U(f) = X(f)Y(f) \iff u(t) = x * y \tag{8.109}$$

▷ The property (8.109) implies that (nonperiodic) convolution of two functions of a continuous argument is commutative: $x * y = y * x$. The commutativity property can also be obtained directly from (8.108).

It's not difficult to guess that the dual of (8.109) is

(Fourier transform dual convolution property)

$$u(t) = x(t)y(t) \iff U(f) = X * Y \tag{8.110}$$

since Fourier transform (8.70a) is practically symmetric to its inverse.

▷ For ω-based Fourier transform we have 3 different normalization placements. Respectively, we are going to have 3 different normalization placements in (8.109) and (8.110).

Goal 8.12 completed. ■

8.12 Shift and modulation

Consider the convolution (8.99) of two discrete-argument functions. We can interpret the right-hand side of (8.99) as

$$x * y = \sum_{j=-\infty}^{\infty} x_j y_j[n] \tag{8.111}$$

where $x_j = x[j]$ and where $y_j[n] = y[n-j]$ is the function $y[n]$ shifted by j in its argument's domain. Thus, the sum in the right-hand side of (8.111) can be interpreted as a linear combination of functions $y_j[\cdot]$, the values x_j being the coefficients of that linear combination.

To be able to analyze the interpretation suggested by (8.111) in more detail, we might want to find the Fourier transform of $y_j[\cdot]$. Generally, given some discrete-argument function $x[n]$ or continuous-argument function $x(t)$, we will be interested in the Fourier transform of its shifted version $x'[n] = x[n-d]$ or $x'(t) = x(t-d)$. This is what we're going to concentrate on in this section. We shall continue with the analysis of (8.111) in Section 8.13.

Goal 8.13. Find the Fourier transform dual-domain counterparts of shifting the function's argument in the original domain (for all 4 Fourier transform types). *Hint:* use the transform formulas.

Let $x(t)$ be an infinitely long nonperiodic function of a continuous argument t and let $x'(t) = x(t-d)$ be its shifted version, where d is the shift amount. By (8.70b)

$$\begin{aligned} x'(t) = x(t-d) &= \int_{-\infty}^{\infty} X(f) e^{2\pi i f(t-d)}\,\mathrm{d}f = \int_{-\infty}^{\infty} X(f) e^{2\pi i f t} e^{-2\pi i f d}\,\mathrm{d}f \\ &= \int_{-\infty}^{\infty} \big(X(f) e^{-2\pi i f d}\big) e^{2\pi i f t}\,\mathrm{d}f = \int_{-\infty}^{\infty} X'(f) e^{2\pi i f t}\,\mathrm{d}f \end{aligned}$$

where $X'(f) = X(f)e^{-2\pi i f d}$. That is, Fourier transform of the function $x(t-d)$ is $X(f)e^{-2\pi i f d}$:

(Fourier transform shift property, a.k.a. time-shift property)

$$x'(t) = x(t-d) \iff X'(f) = X(f)e^{-2\pi i f d} \tag{8.112}$$

In the time/frequency semantics, the formula (8.112) is referred to as the *time-shift property* of Fourier transform. It relates the spectrum $X'(f)$ of a time-shifted signal $x'(t)$ to the spectrum $X(f)$ of the unshifted signal $x(t)$. By (8.112), translation in the time domain (time-shifting) corresponds to multiplication (modulation) by a complex sinusoid in the frequency domain. More abstractly, (8.112) can be referred to as *translation property* of Fourier transform, thereby not suggesting any specific domain semantics.

The dual of (8.112) is obtained for the shifted function $X'(f) = X(f-d)$. By (8.70a)

$$\begin{aligned} X'(f) = X(f-d) &= \int_{-\infty}^{\infty} x(t) e^{-2\pi i (f-d) t}\,\mathrm{d}t = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} e^{2\pi i d t}\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} \big(x(t) e^{2\pi i f d}\big) e^{2\pi i f t}\,\mathrm{d}t = \int_{-\infty}^{\infty} x'(t) e^{2\pi i d t}\,\mathrm{d}t \end{aligned}$$

where $x'(t) = x(t)e^{2\pi i d t}$. That is, shifted Fourier transform $X(f-d)$ corresponds to the function $x(t)e^{2\pi i d t}$:

(Fourier transform dual shift property, a.k.a. frequency-shift property, a.k.a. modulation property)

$$X'(f) = X(f-d) \iff x'(t) = x(t)e^{2\pi i d t} \tag{8.113}$$

In the time/frequency domain semantics, the formula (8.113) is referred to as the *frequency-shift property* of Fourier transform. It givens the time-domain interpretation of shifting the spectrum $X(f)$ of a signal. By (8.113), translation in the frequency domain (frequency-shifting) corresponds to multiplication (modulation) by a complex sinusoid in the time domain. More abstractly, (8.113) can be also referred to as *modulation property* of Fourier transform.

In the same fashion, for Fourier series we obtain

(Fourier series shift and modulation properties)

$$x'(t) = x(t-d) \iff X'_k = X_k e^{-2\pi i k F d} \qquad (d \in \mathbb{R}) \tag{8.114a}$$

$$X'_k = X_{k-d} \iff x'(t) = x(t) e^{2\pi i d F t} \qquad (d \in \mathbb{Z}) \tag{8.114b}$$

For discrete-time Fourier transform we have

(DTFT shift and modulation properties)

$$x'[n] = x[n-d] \iff X'(f) = X(f) e^{-2\pi i f d} \qquad (d \in \mathbb{Z}) \tag{8.115a}$$

$$X'(f) = X(f-d) \iff x'[n] = x[n] e^{2\pi i d n} \qquad (d \in \mathbb{R}) \tag{8.115b}$$

and for discrete Fourier series

(DFT shift and modulation properties)

$$x'[n] = x[n-d] \iff X'[k] = X(f) e^{-2\pi i k d/N} \qquad (d \in \mathbb{Z}) \tag{8.116a}$$

$$X'[k] = X[k-d] \iff x'[n] = x[n] e^{2\pi i d n/N} \qquad (d \in \mathbb{Z}) \tag{8.116b}$$

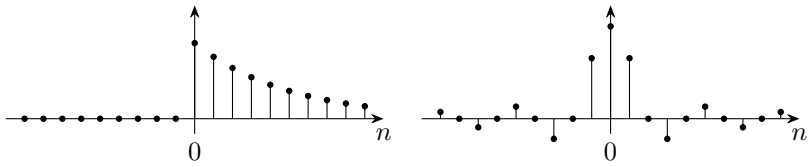
Goal 8.13 completed. ■

8.13 Delta function

8.13.1 Nonperiodic Kronecker delta

If time-domain convolution is understood as time-domain representation of frequency-domain filtering, then by (8.111) we can interpret discrete-time filtering as replacing each sample x_j by a signal $x_j y_j[\cdot]$. Indeed, the original signal $x[n]$ can be seen as a combination of individual samples, each sample occurring at its respective location $n = j$, the amplitude of this sample being equal to x_j. After the convolution, the resulting signal $u[n] = x * y$ can be seen as a combination (sum) of signals $y_j[\cdot]$, each signal $y_j[\cdot]$ "occurring at position $n = j$", the amplitude of $y_j[\cdot]$ being scaled by x_j.

▷ In the above, when talking of the position of an entire signal as a whole, we consider the unshifted signal $y[\cdot]$ to be occurring at zero position, respectively $y_j[\cdot]$ (where $y_j[n] = y[n-j]$) occurs at position j. This idea of the signal's position becomes more intuitive for the specific forms that the signal $y[n]$ assumes in signal processing practice. Such signals commonly happen to be of two types, shown below:



That is either the signal "starts" at $n = 0$ and decays at $n \to +\infty$, or the signal is an even function, often with a peak at $n = 0$, decaying towards $n \to \pm\infty$. Hopefully, the readers are going to find it intuitive to refer to such signals as "having zero position", while respectively e.g. the signal $y'[n] = y[n - 10]$ "has position 10", since its "most prominent point" is now at $n = 10$, etc.

To make our wording less handwavy, let's introduce a formal way to represent "a sample occurring at $n = j$" as a function of a discrete time argument. We can introduce the function

(Kronecker delta function)

$$\delta[n] = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{8.117}$$

representing a sample of a unit amplitude occurring at $n = 0$. This function is referred to as *Kronecker delta*.

▷ The Kronecker delta function $\delta[n]$ is related to the Kronecker delta symbol δ_{mn} as $\delta[m - n] = \delta_{mn}$.

Then we can represent the signal $x[\cdot]$ as

$$x[n] = \sum_{j=-\infty}^{\infty} x_j \delta_j[n] = \sum_{j=-\infty}^{\infty} x[j]\delta[n - j] = x * \delta \tag{8.118}$$

where $\delta_j[n] = \delta[n - j]$, similarly to how $y_j[\cdot]$ were introduced in (8.111). That is $x[n]$ is a sum of signals $x_j\delta_j[n]$. Each signal $x_j\delta_j[n]$ is simply a Kronecker delta function, which has been "shifted to the position j" and scaled by the amplitude factor x_j. Respectively the filtering in (8.111) simply replaces each $\delta_j[n]$ by $y_j[n]$.

▷ Effectively, $y[n]$ is what the filter produces in response to the input signal $\delta[n]$. Hence $y[n]$ is referred to as the *impulse response* of the filter. If the filter is time-invariant, then the input signal $\delta_j[n] = \delta[n-j]$ produces the output $y_j[n] = y[n-j]$. If the filter is also linear, then filtering a linear combination of $\delta_j[\cdot]$ produces a linear combination of $y_j[\cdot]$ with the same linear combination coefficients, and (8.111) follows. Thus linear time-invariant processing of functions of time is representable as time-domain convolution, and, by the convolution property of Fourier transform, it acts as a filter in the frequency domain.

▷ Dropping the intermediate expressions in (8.118) yields

(convolution with δ is transparent)

$$x[n] = x * \delta \tag{8.119}$$

which is one of the key properties of Kronecker delta.

Equation (8.119) can also be demonstrated by formal transformations, where we obtain

$$(x * \delta)[n] = \sum_{j=-\infty}^{\infty} x[j]\delta[n-j] = \sum_{j=n}^{n} x[j] = x[n]$$

Now let's study (8.118) from the frequency domain perspective. We already "know" the frequency domain counterpart of $x[\cdot]$ (it's simply $X(f)$), but what is the frequency domain counterpart of $\delta[\cdot]$?

Goal 8.14. Find the spectrum of $\delta[n]$.

Direct computation by (8.58a) yields

$$\Delta(f) = \sum_{n=-\infty}^{\infty} \delta[n]e^{-2\pi i f n} = \sum_{n=0}^{0} \delta[n]e^{-2\pi i f n} = \delta[0]e^{0} = 1 \tag{8.120}$$

That is, the frequency-domain counterpart of $\delta[\cdot]$ is $\Delta(f) \equiv 1$. Notice that this is in direct agreement with (8.119) and the convolution property of Fourier transform, since in the frequency domain we obtain $X(f) = X(f) \cdot 1 = X(f)\Delta(f)$.

Conversely, by taking the inverse transform (8.58b) of $\Delta(f) \equiv 1$ we obtain

$$\int_{-1/2}^{1/2} \Delta(f)e^{2\pi i f n}\,\mathrm{d}f = \int_{-1/2}^{1/2} e^{2\pi i f n}\,\mathrm{d}f$$

which at $n = 0$ yields

$$\int_{-1/2}^{1/2} e^{0}\,\mathrm{d}f = 1$$

and at $n \neq 0$ yields

$$\int_{-1/2}^{1/2} e^{2\pi i f n}\,\mathrm{d}f = \left.\frac{e^{2\pi i f n}}{2\pi i n}\right|_{n=-1/2}^{1/2} = \frac{1-1}{2\pi i n} = 0$$

and we obtain

$$\int_{-1/2}^{1/2} \Delta(f)e^{2\pi i f n}\,\mathrm{d}f = \delta[n]$$

as the inverse transform of $\Delta(f) \equiv 1$.

Thus

(DTFT spectrum of Kronecker delta function)

$$x[n] = \delta[n] \iff X(f) \equiv 1 \tag{8.121}$$

The formal dual of (8.121) is a property of the Fourier series, where we have

(Fourier series spectrum of a unit constant)

$$x(t) \equiv 1 \iff X_k = \delta[k] \tag{8.122}$$

which can be obtained/verified in a similar fashion to (8.121).

Goal 8.14 completed. ■

Equation (8.121) combined with (8.118) and the time-shift property of DTFT can be used to give a different interpretation to the discrete-time Fourier transform formula (8.58a). Obviously, DTFT is a linear transformation of $x[n]$, that is, DTFT of $\alpha x[n] + \beta y[n]$ is $\alpha X(f) + \beta Y(f)$ where $X(f)$ and $Y(f)$ are DTFTs of $x[n]$ and $y[n]$. By (8.118) we can represent $x[n]$ as a linear combination of $\delta_j[n]$. Then DTFT of $x[n]$ must be the same linear combination of $\Delta_j(f)$:

$$X(f) = \sum_{j=-\infty}^{\infty} x_j \Delta_j(f)$$

where $\Delta_j(f)$ are DTFTs of $\delta_j[n]$. But $\delta_j[n] = \delta[n-j]$ and by the time-shift property (8.115a) $\Delta_j(f) = \Delta(f)e^{-2\pi i f j} = 1 \cdot e^{-2\pi i f j} = e^{-2\pi i f j}$ and we obtain

$$X(f) = \sum_{j=-\infty}^{\infty} x_j \Delta_j(f) = \sum_{j=-\infty}^{\infty} x[j]e^{-2\pi i f j} = \sum_{n=-\infty}^{\infty} x[n]e^{-2\pi i f n} \tag{8.123}$$

which is the already familiar DTFT formula (8.58a).

▷ The dual of (8.123) can be written for the inverse transformation part of Fourier series by first representing the sequence X_k as a linear combination of Kronecker delta functions similarly to (8.118) and then applying the inverse transformation to each of these delta functions in turn.

8.13.2 Periodic Kronecker delta

If instead of discrete-time Fourier transform we are dealing with discrete Fourier transform, the relationship (8.121) is not applicable, since $x[n]$ in (8.121) is not periodic and the dual domain representation is not a periodic function of a continuous argument either. Let's see how the situations changes in case of DFT.

Goal 8.15. Construct a periodic counterpart of $\delta[n]$ and study its transformations under DFT.

Let $X[k] \equiv 1$, which is, obviously, a periodic function (where the period can be chosen arbitrarily) of a discrete argument k. Taking the inverse discrete Fourier transform, specifically the density version (8.23b), we have

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{2\pi i k n/N} = \frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi i k n/N}$$

where at n being a multiple of N, that is $n = mN$, we immediately obtain

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi i k\cdot(mN)/N} = \frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi i k m} = \frac{1}{N}N = 1$$

and at $n \neq mN$, by the geometric series formula,

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi i k n/N} = \frac{1}{N}\sum_{k=0}^{N-1} \left(e^{2\pi i n/N}\right)^k = \frac{1}{N}\cdot\frac{\left(e^{2\pi i n/N}\right)^N - 1}{e^{2\pi i n/N} - 1}$$

$$= \frac{1}{N} \cdot \frac{e^{2\pi in} - 1}{e^{2\pi in/N} - 1} = 0$$

Thus

$$x[n] = \begin{cases} 1 & \text{if } n \text{ is a multiple of } N \\ 0 & \text{if } n \text{ is not a multiple of } N \end{cases}$$

Apparently, $x[n]$ looks exactly like a periodic version of Kronecker delta. We can introduce

(periodic Kronecker delta function)

$$\delta_N[n] = \sum_{k=-\infty}^{\infty} \delta[n - kN] = \begin{cases} 1 & \text{if } n \text{ is a multiple of } N \\ 0 & \text{if } n \text{ is not a multiple of } N \end{cases} \tag{8.124}$$

as the *N-periodic Kronecker delta* function.

Conversely, taking the direct Fourier transform (8.23a) of $\delta_N[n]$ yields

$$\Delta_N[k] = \sum_{n=0}^{N-1} \delta_N[n] e^{-2\pi ikn/N} = \sum_{n=0}^{0} \delta_N[n] e^{-2\pi ikn/N} = \delta_N[0] e^0 = 1$$

Thus

(density-version DFT spectrum of periodic Kronecker delta)

$$x[n] = \delta_N[n] \iff X[k] \equiv 1 \tag{8.125a}$$

The amplitude version of Fourier transform yields N times smaller $X[k]$ compared to the density version, thus (8.125a) takes the form

(amplitude-version DFT spectrum of periodic Kronecker delta)

$$x[n] = \delta_N[n] \iff X[k] \equiv \frac{1}{N} \tag{8.125b}$$

The duals of (8.125a) and (8.125b) are respectively

(DFT spectrum of a unit constant)

$$x[n] \equiv 1 \iff X[k] = N\delta_N[k] \qquad \text{(density version)} \tag{8.126a}$$

$$x[n] \equiv 1 \iff X[k] = \delta_N[k] \qquad \text{(amplitude version)} \tag{8.126b}$$

Goal 8.15 completed. ■

It is easily noticed that the representation (8.118) of a nonperiodic signal as a linear combination of samples (via a convolution with the delta function) holds for periodic signals and periodic convolution as well:

$$x[n] = \sum_{j=0}^{N-1} x[j]\delta_{Nj}[n] = \sum_{j=0}^{N-1} x[j]\delta_N[n-j] = x * \delta_N \tag{8.127}$$

where $\delta_{Nj}[n] = \delta_N[n-j]$ is $\delta_N[\cdot]$ shifted to the position j. Each term $x[j]\delta_{Nj}[n]$ thereby represents a periodic signal, consisting of the samples $x[j]$, $x[j \pm N]$, $x[j \pm 2N]$, etc. The sum of N such periodic signals with nonzero samples occurring at mutually exclusive positions reconstructs the original signal $x[n]$.

▷ Dropping the intermediate expressions in (8.127) we obtain the periodic counterpart of the property (8.119):

$$x[n] = x * \delta \tag{8.128}$$

This property can also be obtained formally, obtaining

$$(x * \delta_N)[n] = \sum_{j=0}^{N-1} x[j]\delta_N[n-j] = \sum_{j=n}^{n+N-1} x[j]\delta_N[n-j] = \sum_{j=n}^{n} x[j] = x[n]$$

(where the summation range change is by the periodicity of the expression under the summation sign).

The representation (8.127) allows to reinterpret discrete Fourier transform as a linear combination of transformed periodic Kronecker delta functions, similarly to how we reinterpreted discrete-time Fourier transform in (8.123):

$$X[k] = \sum_{j=0}^{N-1} x_j \Delta_{Nj}[k] = \sum_{j=0}^{N-1} x[j] e^{-2\pi ikj/N} = \sum_{n=0}^{N-1} x[n] e^{-2\pi ikn/N} \tag{8.129}$$

where $\Delta_{Nj}[k] = e^{-2\pi ikj/N}$ are the DFTs of $\delta_{Nj}[n]$ obtained by the time-shift property (8.116a).

8.13.3 DFT roundtrip

Suppose we are given an N-periodic function $x[n]$ and we apply discrete Fourier transform to this function and then apply inverse discrete Fourier transform to the obtained result. We will refer to such sequence of direct and inverse DFT as a discrete Fourier transform *roundtrip*. We already know that such roundtrip should yield back the original function $x[n]$, due to the transformation matrices $\mathbf{H}^*/N$ and $\mathbf{H}$ in (8.15) being mutually inverse. Let's however re-verify that result not in terms of matrix notation used by (8.15) but in terms of explicit sums used by (8.21).

Let $x'[n]$ be the roundtrip result. We want to show that $x'[n] = x[n]$. Substituting (8.21a) into the right-hand side of (8.21b) we obtain

$$\begin{aligned}
x'[n] &= \sum_{k=0}^{N-1} X[k] e^{2\pi ikn/N} = \sum_{k=0}^{N-1} \left(\frac{1}{N} \sum_{\nu=0}^{N-1} x[\nu] e^{-2\pi ik\nu/N} \right) e^{2\pi ikn/N} \\
&= \frac{1}{N} \sum_{k=0}^{N-1} \sum_{\nu=0}^{N-1} x[\nu] e^{2\pi ik(n-\nu)/N} = \sum_{\nu=0}^{N-1} \left(x[\nu] \cdot \frac{1}{N} \sum_{k=0}^{N-1} e^{2\pi ik(n-\nu)/N} \right) \\
&= \sum_{\nu=0}^{N-1} \left(x[\nu] \cdot \frac{1}{N} \sum_{k=0}^{N-1} \Delta_N[k] e^{2\pi ik(n-\nu)/N} \right)
\end{aligned}$$

where $\Delta_N[k] \equiv 1\ \forall k$. Now inside the parentheses we have $x[\nu]$ multiplied by the inverse discrete Fourier transform (density version) of the function $\Delta_N[k] \equiv 1$ with respect to the "time-domain" variable $n - \nu$. As we already know, the inverse discrete Fourier transform of $\Delta_N[\cdot]$ is simply $\delta_N[\cdot]$, thus

$$\frac{1}{N}\sum_{k=0}^{N-1} \Delta_N[k] e^{2\pi i k(n-\nu)/N} = \delta_N[n-\nu]$$

and continuing the previous transformations we obtain

$$\begin{aligned} x'[n] &= \sum_{\nu=0}^{N-1}\left(x[\nu]\cdot\frac{1}{N}\sum_{k=0}^{N-1}\Delta_N[k]e^{2\pi i k(n-\nu)/N}\right) = \sum_{\nu=0}^{N-1}\left(x[\nu]\cdot\delta_N[n-\nu]\right) \\ &= x * \delta_N = x[n] \end{aligned} \tag{8.130}$$

We have thereby found that the discrete Fourier transform roundtrip is equivalent to convolving the original function $x[n]$ with the periodic Kronecker delta $\delta_N[n]$, and, as we know, this convolution yields back $x[n]$. This result seems trivial for discrete Fourier transform, since we already verified exactly the same by (8.15), but it's going to become more useful with other Fourier transform types, which we obtained as the limiting versions of discrete Fourier transform without paying proper attention to convergence details, and thus we cannot really be sure, under which conditions the roundtrip property continues to hold.

8.13.4 DTFT roundtrip

Goal 8.16. Study the DTFT roundtrip and construct a sufficient condition for the roundtrip to deliver the original function. *Hint:* follow the DFT roundtrip study as a reference.

Let $x[n]$ be a nonperiodic infinitely long function and let $x'[n]$ be the DTFT roundtrip result for this function. We expect that $x'[n] = x[n]$. Let's verify that. Substituting (8.58a) into the right-hand side of (8.58b) yields

$$\begin{aligned} x'[n] &= \int_{-1/2}^{1/2} X(f)e^{2\pi i f n}\,\mathrm{d}f = \int_{-1/2}^{1/2}\left(\sum_{\nu=-\infty}^{\infty} x[\nu]e^{-2\pi i f \nu}\right)e^{2\pi i f n}\,\mathrm{d}f \\ &= \int_{-1/2}^{1/2}\left(\sum_{\nu=-\infty}^{\infty} x[\nu]e^{2\pi i f(n-\nu)}\right)\mathrm{d}f \end{aligned}$$

We would like to swap the order of the integration and summation, similarly to how we swapped the order of summations in the analysis of the DFT roundtrip:

$$x'[n] \stackrel{?}{=} \sum_{\nu=-\infty}^{\infty}\int_{-1/2}^{1/2} x[\nu]e^{2\pi i f(n-\nu)}\,\mathrm{d}f$$

This could be made possible, if $\sum x[\nu]e^{2\pi i f(n-\nu)}$ converged uniformly (with respect to f). The latter, in turn, could be ensured, if the sum $\sum x[\nu]$ converged absolutely. Indeed, then we would have

$$\sum_{\nu=-\infty}^{\infty}\left|x[\nu]e^{2\pi i f(n-\nu)}\right| \le \sum_{\nu=-\infty}^{\infty}|x[\nu]| < \infty$$

and the absolute and uniform convergence of $\sum x[\nu]e^{2\pi i f(n-\nu)}$ would follow. Therefore, let's assume absolute convergence of the latter sum. Then we have

$$\begin{aligned} x'[n] &= \sum_{\nu=-\infty}^{\infty} \int_{-1/2}^{1/2} x[\nu]e^{2\pi i f(n-\nu)}\,\mathrm{d}f = \sum_{\nu=-\infty}^{\infty} x[\nu] \int_{-1/2}^{1/2} e^{2\pi i f(n-\nu)}\,\mathrm{d}f \\ &= \sum_{\nu=-\infty}^{\infty} x[\nu] \int_{-1/2}^{1/2} \Delta(f)e^{2\pi i f(n-\nu)}\,\mathrm{d}f \end{aligned}$$

where $\Delta(f) \equiv 1$. The integral in the latter expression represents the inverse DTFT of $\Delta(f)$ with respect to the "time-domain" variable $n-\nu$ and is therefore equal to $\delta[n-\nu]$ and we obtain

$$x'[n] = \sum_{\nu=-\infty}^{\infty} x[\nu]\delta[n-\nu] = x * \delta = x[n] \tag{8.131}$$

We have just obtained that, under the assumption of the absolute summability of $x[n]$, the DTFT roundtrip yields back the original function.

▷ We have obtained that absolute summability of $x[n]$ is a sufficient condition to ensure that the DTFT roundtrip is transparent (doesn't change the transformed function). This condition is not necessary, however. E.g. the DTFT roundtrip for

$$x[n] = \begin{cases} 1/n & \text{if } n \neq 0 \\ 0 & \text{if } n = 0 \end{cases}$$

(which is, essentially, a dual of the Fourier series of a sawtooth function) is transparent as well, even though $x[n]$ is not absolutely summable. The question of a sufficient and necessary condition is, however, pretty complicated.

Goal 8.16 completed. ■

One might now expect that, since Fourier series is a dual of discrete-time Fourier transform, we can reuse the just obtained results for the Fourier series roundtrip. This, however, is not exactly true. While the DTFT roundtrip starts with a discrete-argument function $x[n]$, the Fourier series roundtrip would normally start with a continuous-argument function $x(t)$. So, while the two roundtrip paths are effectively the same, they have opposite starting positions. Before we analyze the Fourier series roundtrip, we shall introduce the continuous-argument counterpart of the periodic Kronecker delta function $\delta_N[n]$.

8.13.5 Periodic Dirac delta

With Kronecker delta function we first introduced the nonperiodic version, which is somewhat more intuitive, and then introduced the periodic one, where the latter could be also expressed as an infinite sum of the former as in (8.124). We would have liked to follow the same approach with the continuous-argument counterpart of Kronecker delta, but the respective nonperiodic function is somewhat more difficult to analyze, and therefore we are going to start with the periodic version instead.

Goal 8.17. Construct a continuous-argument counterpart of $\delta_N[n]$. *Hint:* follow the construction steps of $\delta_N[n]$ as a reference.

Following our usual approach taken in previous subsections, we compute the inverse transform part of the Fourier series given by $X_k \equiv 1\ \forall k$. Substituting $X_k \equiv 1$ into (8.64b) gives

$$x(t) = \sum_{k=-\infty}^{\infty} e^{2\pi i k F t} = \lim_{N\to\infty} \sum_{k=-N}^{N} e^{2\pi i k F t} \tag{8.132a}$$

where we recalled that the infinite sum in (8.64b) is understood in the Cauchy principal value sense. Obviously, the limit diverges at $t = 0$, but let's continue nevertheless:

$$x(t) = \lim_{N\to\infty} \sum_{k=-N}^{N} e^{2\pi i k F t} = \lim_{N\to\infty} \left(e^{-2\pi i N F t} \cdot \frac{e^{2\pi i \cdot (2N+1) F t} - 1}{e^{2\pi i F t} - 1} \right) \tag{8.132b}$$

where we used the geometric series formula. Strictly speaking the geometric series formula is applicable only for $t \neq 0$, but we can notice that it can also work at $t = 0$ if we understand the formula in the limiting sense, that is instead of evaluating the expression at $t = 0$ we take its limit at $t \to 0$. Continuing,

$$\begin{aligned} x(t) &= \lim_{N\to\infty} \frac{e^{2\pi i (N+1) F t} - e^{-2\pi i N F t}}{e^{2\pi i F t} - 1} = \lim_{N\to\infty} \frac{e^{2\pi i (N+\frac{1}{2}) F t} - e^{-2\pi i (N+\frac{1}{2}) F t}}{e^{2\pi i \cdot \frac{1}{2} F t} - e^{-2\pi i \cdot \frac{1}{2} F t}} \\ &= \lim_{N\to\infty} \frac{e^{(2N+1)\pi i F t} - e^{-(2N+1)\pi i F t}}{e^{\pi i F t} - e^{-\pi i F t}} = \lim_{N\to\infty} \frac{\sin\big((2N+1)\pi F t\big)}{\sin \pi F t} \\ &= \lim_{N\to\infty} \frac{\sin\big((N+\frac{1}{2}) \cdot 2\pi F t\big)}{\sin(\frac{1}{2} \cdot 2\pi F t)} = \lim_{N\to\infty} D_N(2\pi F t) \end{aligned} \tag{8.132c}$$

where the function

$$D_N(t) = \sum_{k=-N}^{N} e^{ikt} = \frac{\sin\big((N+\frac{1}{2})t\big)}{\sin \frac{1}{2} t} \tag{8.133}$$

is referred to as the N-th *Dirichlet kernel* function. Notice that for $t = 0$ we continue to understand the expressions in the limiting sense to avoid the division by zero. Fig. 8.2 plots two different functions $D_N(t)$ for two different values of N.

▷ Dirichlet kernel is a collection of functions $D_N(t)$, which instead can be notated as a two-argument function $D[N, t)$ or $D(t, N]$ (depending on the convention), where one argument is continuous and the other is discrete. In this notation it looks like a kernel of an integral transform, where the transform's "output" is a function of a discrete argument.

Looking at the first of the two expressions in (8.133) we realize that each $D_N(t)$ is a 2π-periodic function and that the area underneath one period is equal to 2π, since this is what the integration of the term at $k = 0$ gives, while all other terms give zero upon integrating them. Let's actually switch back to $D_N(2\pi t)$:

$$D_N(2\pi t) = \sum_{k=-N}^{N} e^{2\pi i k t} = \frac{\sin\big((2N+1)\pi t\big)}{\sin \pi t} \tag{8.134}$$

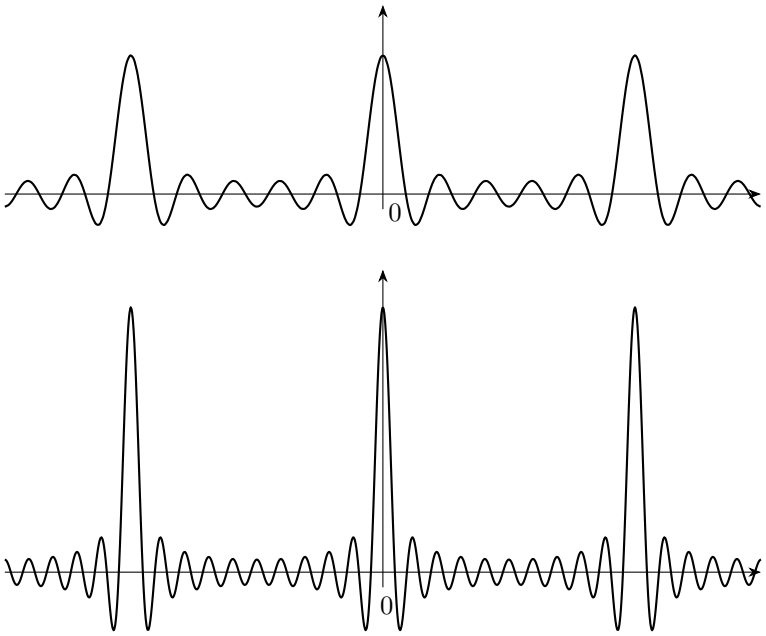


Figure 8.2: Dirichlet kernel function $D_N(t)$ for $N = 5$ (top) and $N = 10$ (bottom).

in which case the function is 1-periodic and the area underneath one period is unity:

$$\int_{t_0}^{t_0+1} D_N(2\pi t)\,\mathrm{d}t = 1$$

Now, let's for a while concentrate on the numerator of $D_N(2\pi t)$ which is a sinusoid

$$\eta_N(t) = \sin\big((2N+1)\pi t\big)$$

Notice that the period of η_0 is 2, and the period of η_N is thereby $\frac{2}{2N+1}$. As $N \to \infty$, this sinusoid oscillates faster and faster. Therefore an integral of this function over any fixed interval is going to approach zero, since at larger N, even if this interval is not a multiple of 2, the positive and negative areas of the sinusoid are going to cancel almost exactly anyway:

$$\lim_{N\to\infty} \int_a^b \eta_N(t)\,\mathrm{d}t = 0$$

The same property will however also hold for a product of $\eta_N(t)$ with almost any other function $g(t)$:

$$\lim_{N\to\infty} \int_a^b g(t)\eta_N(t)\,\mathrm{d}t = 0$$

Indeed, as long as the speed of change of $g(t)$ is bounded,[6] we can always "zoom into $g(t)$" sufficiently far in, so that $g(t)$ on a given interval is approximately

[6] Intuitively, if $|g'(t)|$ is bounded, although strictly speaking $g'(t)$ doesn't even need to exist everywhere. More rigorous treatments use the idea of *Lipschitz continuity*.

constant: $g(t) \approx \tilde{g}$. Then $g(t)\eta_N(t)$ will be approximately equal to $\tilde{g}\eta_N(t)$ on that interval, hence the integral of $g(t)\eta_N(t)$ on that interval will converge to approximately the same value as the integral of $\tilde{g}\eta_N(t)$, that is approximately to zero. Then, given any large interval $[a, b]$ we can always split it into sufficiently small intervals, and the integral of $g(t)\eta_N(t)$ on $[a, b]$ will be the sum of the integrals on those sufficiently small integrals. Making these intervals smaller and smaller we find that the integral of $g(t)\eta_N(t)$ on $[a, b]$ must be indefinitely small, that is equal to zero.

▷ We have effectively arrived at the statement of the *Riemann–Lebesgue lemma*:

$$\lim_{\omega\to\infty} \int_a^b x(t) \sin \omega t \, \mathrm{d}t = 0 \tag{8.135a}$$

or, more generally

$$\lim_{\omega\to\infty} \int_a^b x(t) e^{j\omega t} \, \mathrm{d}t = 0 \tag{8.135b}$$

(applicable to functions $x(t)$ with a bounded speed of change). A corollary of this is that Fourier series coefficients X_k of such $x(t)$ decay to zero at $k \to \infty$.

A jump discontinuity in $x(t)$ at any point $t = t_0 \in [a, b]$ doesn't affect the statement of (8.135), since we can represent the integral over $[a, b]$ as a sum of two integrals over $[a, t_0]$ and $[t_0, b]$, each of the two integrals approaching zero by (8.135). By induction we can extend this result to an arbitrary finite number of jump discontinuities on $[a, b]$. Under an additional assumption that $x(t)$ is bounded on $[a, b]$ the result can be extended to a countable number of jump discontinuities, by effectively ignoring the contributions of the integrals over smaller-size intervals.

By letting $a = -\infty$, $b = +\infty$ and additionally assuming an absolutely integrable $x(t)$, Riemann–Lebesgue lemma can be applied in the context of Fourier transform, obtaining $X(\omega) \to 0$ at $\omega \to \pm\infty$. Notice that with Fourier series it would be easy to infer $X_k \to 0$ from the series convergence (if the convergence is given or assumed), since the terms of a converging series must decay to zero. However the same statement doesn't work for converging improper integrals (e.g. the function under the integral can consist of regularly spaced constant-height peaks of indefinitely decreasing widths), so we wouldn't be able to apply the same reasoning to infer $X(\omega) \to 0$ from convergence of inverse Fourier transform.

A more rigorous treatment of this topic falls outside the scope of this book and can be found in other texts.

Now the same idea of the equality of the integral to zero at $N \to \infty$ can be applied to $g(t)D_N(2\pi t)$, except in close neighborhoods of integer values of t. Indeed

$$g(t)D_N(2\pi t) = g(t)\frac{\sin\big((2N+1)\pi t\big)}{\sin \pi t} = \frac{g(t)}{\sin \pi t}\eta_N(t) = h(t)\eta_N(t)$$

where $h(t) = \frac{g(t)}{\sin \pi t}$ is simply another function, and the integral of $h(t)\eta_N(t)$ also must be converging to zero as $N \to \infty$, unless t is close to an integer value where $\sin \pi t$ is close to zero (in which case the speed of change of $h(t)$ is not bounded). Thus an integral of $g(t)D_N(2\pi t)$ on any interval $[a, b]$ is also approaching zero at $N \to \infty$, provided $[a, b]$ doesn't contain any integer values.

Let's summarize what we found out about $D_N(2\pi t)$:

$$D_N(2\pi(t+1)) = D_N(2\pi t) \qquad (D_N(2\pi t) \text{ is 1-periodic}) \tag{8.136a}$$

$$\int_{t_0}^{t_0+1} D_N(2\pi t)\,\mathrm{d}t = 1 \qquad \text{for any } t_0 \text{ and any } N \tag{8.136b}$$

$$\lim_{N\to\infty}\int_a^b g(t)D_N(2\pi t)\,\mathrm{d}t = 0 \qquad \text{for } [a,b] \text{ not containing integers} \tag{8.136c}$$

From (8.136b) it follows that

$$\int_{-1/2}^{1/2} g(0)D_N(2\pi t)\,\mathrm{d}t = g(0)$$

and combining this with (8.136c) we should conclude that

$$\lim_{N\to\infty}\int_{-1/2}^{1/2} g(t)D_N(2\pi t)\,\mathrm{d}t = g(0) \tag{8.137a}$$

since, by (8.136c), the values of $g(t)$ at any points except $t = 0$ do not contribute to the integral. Effectively, $D_N(\cdot)$ behaves as being equal to zero (in the limit $N \to \infty$) everywhere on $[-\frac{1}{2}, \frac{1}{2}]$ except $t = 0$, where is behaves as becoming infinitely large (so that a single value at $t = 0$ makes a finite contribution to the integral). This behavior matches what we see in Fig. 8.2: the peaks become infinitely high and infinitely narrow, while everywhere else the function turns into a sinusoid of indefinitely growing frequency (and a slowly changing amplitude).

▷ If we allow $g(t)$ to be jump-discontinuous at $t = 0$, we obtain

$$\lim_{N\to\infty}\int_{-1/2}^{1/2} g(t)D_N(2\pi t)\,\mathrm{d}t = \frac{g(+0)+g(-0)}{2} \tag{8.137b}$$

since $D_N(2\pi t)$ are even functions, and the areas of $g(t)$ immediately to the right and to the left of $t = 0$ contribute equally to the integral of $g(t)D_N(2\pi t)$ for each N.

A jump discontinuity at any other point $t = a \neq 0$ doesn't contribute to the integral, which can be seen by simply representing the integral over $[-\frac{1}{2}, \frac{1}{2}]$ as sum of two integrals over $[-\frac{1}{2}, a]$ and $[a, \frac{1}{2}]$ respectively.

We will now imagine as if $D_N(2\pi t)$ at $N \to \infty$ converge to a 1-periodic function $\delta_1(t)$, known as *1-periodic Dirac delta*:

(periodic Dirac delta of a unit period)

$$\delta_1(t) = \lim_{N\to\infty} D_N(2\pi t) \tag{8.138}$$

which retains the properties (8.136) and (8.137) of $D_N(2\pi t)$, which we reorganize as

(properties of 1-periodic Dirac delta)

$$\delta_1(t+1) = \delta_1(t) \qquad (\delta_1(t) \text{ is 1-periodic}) \tag{8.139a}$$

$$\int_{-1/2}^{1/2} g(t)\delta_1(t)\,\mathrm{d}t = g(0) \qquad \text{if } g(t) \text{ is continuous at } t=0 \tag{8.139b}$$

$$\int_{-1/2}^{1/2} g(t)\delta_1(t)\,\mathrm{d}t = \frac{g(+0)+g(-0)}{2} \qquad \text{generally} \tag{8.139c}$$

Thereby $\delta_1(t)$ is not a real function, but some kind of concise notational summary of the limiting properties of $D_N(2\pi t)$ at $N \to \infty$.

▷ Formally, $\delta_1(t)$ is *generalized function* (a.k.a. *distribution*). It doesn't have defined values at any point, but is defined in terms of specifying the values of the integrals $\int_a^b \delta_1(t)g(t)\,\mathrm{d}t$ for any $g(t)$ and any interval. E.g. (8.139) serve as such specification, as they allow to compute the value of any integral of this form. Indeed, one can infer from (8.139b) that the contribution of all values of $g(t)$ except at $t=0$ to the integral is zero and therefore

$$\int_a^b g(t)\delta_1(t)\,\mathrm{d}t = \sum_{a<n<b} g(n) \qquad (\text{where } n \in \mathbb{Z},\ a<b \text{ and } a,b \notin \mathbb{Z})$$

For integer a and/or b one can use (8.139c) to deduce that the respective contributions to the integral at those points are $g(a)/2$ and $g(b)/2$.

In practice there also usually need to be some restrictions on $g(t)$. E.g. we assumed that $g(t)$ has bounded speed of change, except possibly at a number of points, where it has jump discontinuities.

▷ Notice that (8.139b), by formally letting $g(t) \equiv 1$, automatically implies

(the area under 1-periodic Dirac delta on $[-\frac{1}{2}, \frac{1}{2}]$ is unity)

$$\int_{-1/2}^{1/2} \delta_1(t)\,\mathrm{d}t = 1 \tag{8.140}$$

which is a particular case of the more general property (8.136b).

We could also imagine a similar "function" which is T-periodic rather than 1-periodic. Since $\delta_1(t)$ is the limit of $D_N(2\pi t)$, it follows that the limit of $D_N(2\pi F t)$ is simply $\delta_1(Ft)$, which, since $F = 1/T$, can also be written as $\delta_1(t/T)$.

We will introduce the *T-periodic Dirac delta* function, also known as *Dirac comb*, as

(periodic Dirac delta of an arbitrary period)

$$\delta_T(t) = \frac{1}{T}\delta_1\left(\frac{t}{T}\right) \tag{8.141}$$

where, by convention, we multiply the result by $1/T$ to keep the area under the function equal to unity, so that

$$\int_{-T/2}^{T/2} g(t)\delta_T(t)\,\mathrm{d}t = g(0)$$

rather than yielding a T times larger value. Thereby

$$\lim_{N\to\infty} D_N(2\pi F t) = T\delta_T(t) \qquad \text{where } F = 1/T \tag{8.142}$$

Returning to the beginning of our discussion and (8.132), we obtain that by (8.132) and (8.142) the Fourier series with $X_k \equiv 1$ "converges" to $x(t) = T\delta_T(t)$. Let's now try the reverse, that is apply the Fourier series coefficient formula (8.64a) to $x(t) = T\delta_T(t)$ to see which X_k we obtain. Substituting $x(t) = T\delta_T(t)$ into (8.64a) yields

$$\begin{aligned} X_k &= \frac{1}{T}\int_{-T/2}^{T/2} T\delta_T(t)e^{-2\pi i k F t}\,\mathrm{d}t = \int_{-T/2}^{T/2} \delta_T(t)e^{-2\pi i k F t}\,\mathrm{d}t \\ &= \int_{-T/2}^{T/2} \delta_T(t) g_k(t)\,\mathrm{d}t \end{aligned}$$

where $g_k(t) = e^{-2\pi i k F t}$. By (8.139b) the latter integral is equal to $g_k(0)$, that is simply to 1, and we obtain $X_k = 1\ \forall k$. Thus we obtain the "actual" dual counterpart of (8.122):

(Fourier series spectrum of periodic Dirac delta)

$$X_k \equiv 1 \iff x(t) = T\delta_T(t) \tag{8.143a}$$

$$x(t) = \delta_T(t) \iff X_k \equiv \frac{1}{T} \tag{8.143b}$$

where (8.143b) has been obtained by dividing (8.143a) by T. Notice the strong similarity between (8.143b) and (8.125b).

A formal dual of (8.143), which is simultaneously the "actual" dual of (8.121), is

(DTFT spectrum of a unit constant)

$$x[n] \equiv 1 \iff X(f) = \delta_1(f) \tag{8.144}$$

Goal 8.17 completed. ■

The periodic Dirac delta function can also be used to represent a continuous periodic $x(t)$ as a "linear combination of samples", along the lines of (8.127):

(periodic convolution with δ_T is transparent)

$$x(t) = \int_{-T/2}^{T/2} x_\tau \delta_{T\tau}(t)\,\mathrm{d}\tau = \int_{-T/2}^{T/2} x(\tau)\delta_T(t-\tau)\,\mathrm{d}\tau = x * \delta_T \tag{8.145}$$

where $x_\tau\,\mathrm{d}\tau$ are the coefficients of the linear combination of the functions $\delta_{T\tau}(t) = \delta_T(t-\tau)$. This in turn allows an alternative interpretation of the direct transformation part of the Fourier series, similarly to (8.123) and (8.129):

$$X_k = \int_{-T/2}^{T/2} x_\tau \Delta_{T\tau}[k]\,\mathrm{d}\tau = \int_{-T/2}^{T/2} \left(x(\tau)\cdot \frac{1}{T} e^{-2\pi i k F \tau} \right) \mathrm{d}\tau$$

$$= \frac{1}{T}\int_{-T/2}^{T/2} x(\tau)e^{-2\pi ikF\tau}\mathrm{d}\tau \tag{8.146}$$

where $\Delta_{T\tau}[k] = \frac{1}{T}e^{-2\pi ikF\tau}$ are Fourier series coefficients of $\delta_{T\tau}(t) = \delta_T(t-\tau)$.

▷ Formula (8.145) can be obtained by a formal application of (8.139b):

$$x * \delta_T = \int_{-T/2}^{T/2} x(\tau)\delta_T(t-\tau)\,\mathrm{d}\tau = \int_{-T/2}^{T/2} x(t-\tau)\delta_T(\tau)\,\mathrm{d}\tau \overset{(8.139b)}{=} x(t)$$

where we can replace $x(\tau)\delta_T(t-\tau)$ with $x(t-\tau)\delta_T(\tau)$ under the integral sign by the commutativity of periodic convolution, and where (8.139b) is applied with respect to $x(t-\tau)$ as a function of τ, yielding $x(t)$ at $\tau = 0$.

More generally, by using (8.139c), the formula (8.145) takes the form

(periodic convolution with δ_T at a jump discontinuity)

$$x * \delta_T = \frac{x(t+0) + x(t-0)}{2} \tag{8.147}$$

which accommodates possible discontinuities of $x(t)$.

8.13.6 Fourier series roundtrip

Goal 8.18. Study the Fourier series roundtrip and construct a sufficient condition for the roundtrip to deliver the original function. *Hint:* refer to the previous roundtrip studies and use the convergence of Dirichlet kernel functions to the periodic Dirac delta.

Dirichlet kernel functions $D_N(\cdot)$ and $\delta_T(\cdot)$ as their limit are extremely useful for analyzing the Fourier series roundtrip. Indeed, let $x(t)$ be the original function and $x'(t)$ be the roundtrip result. We are going to rewrite (8.64b) in terms of the Cauchy principal value limit and substitute (8.64a) for X_k:

$$\begin{aligned}
x'(t) &= \lim_{N\to\infty}\sum_{k=-N}^{N} X_k e^{2\pi ikFt} = \lim_{N\to\infty}\sum_{k=-N}^{N}\left(\frac{1}{T}\int_{-T/2}^{T/2} x(\tau)e^{-2\pi ikF\tau}\,\mathrm{d}\tau\right)e^{2\pi ikFt} \\
&= \frac{1}{T}\lim_{N\to\infty}\int_{-T/2}^{T/2}\left(\sum_{k=-N}^{N} x(\tau)e^{2\pi ikF(t-\tau)}\right)\mathrm{d}\tau \\
&= \frac{1}{T}\lim_{N\to\infty}\int_{-T/2}^{T/2}\left(x(\tau)\sum_{k=-N}^{N} e^{2\pi ikF(t-\tau)}\right)\mathrm{d}\tau \\
&= \frac{1}{T}\lim_{N\to\infty}\int_{-T/2}^{T/2} x(\tau)D_N\big(2\pi kF(t-\tau)\big)\mathrm{d}\tau = \frac{1}{T}\lim_{N\to\infty}\big(x * D_N(2\pi kF\cdot)\big)
\end{aligned} \tag{8.148}$$

We have obtained that the partial sums of Fourier series (8.64b) are equal to periodic convolutions of $x(t)$ with Dirichlet kernel functions $D_N(2\pi kFt)$. In the limit, $D_N(2\pi kFt)$ turn into $T\delta_T(t)$ and respectively the limit in (8.148) turns into

$$x'(t) = \frac{1}{T}\big(x * T\delta_T\big) = x * \delta_T = x(t) \tag{8.149}$$

where the last equality is by (8.145). Notice that (8.149) is the Fourier series counterpart of (8.130) and (8.131).

▷ For the points of jump discontinuities we obtain by (8.147)

$$x'(t) = \frac{1}{T}\big(x * T\delta_T\big) = x * \delta_T = \frac{x(t+0) + x(t-0)}{2} \tag{8.150}$$

The result of (8.149) thereby looks like a confirmation of Fourier series roundtrip transparency, however convolution with δ_T expects the speed of change of $x(\cdot)$ to be bounded. If the speed of change of $x(t)$ is not bounded, we would not be so sure, whether we can switch from the limit of $D_N(2\pi kFt)$ to $T\delta_T(t)$, and respectively whether Fourier series will yield back the original $x(t)$. Thus, we have obtained a sufficient condition for the Fourier series roundtrip transparency: the bounded speed of change of $x(\cdot)$ (with the reservation of (8.150) for the points of discontinuity).

Thus, while Fourier series roundtrip is pretty much the same as the DTFT roundtrip, due to a different starting point we have obtained a very different sufficient condition for the roundtrip transparency. While the DTFT roundtrip condition was specifying the expectations towards a nonperiodic $x[n]$, with Fourier series we specify the expectations towards a periodic $x(t)$.

Also, while with DTFT the inverse transform was a proper integral taken over a unit frequency range, with Fourier series the inverse transform is an infinite sum, which may therefore converge or diverge. We can therefore express the roundtrip transparency of Fourier series by saying that Fourier series converges to the original function $x(t)$. Therefore the question of roundtrip transparency of Fourier series is often referred to as the question of Fourier series convergence.

Goal 8.18 completed. ■

8.13.7 Nonperiodic Dirac delta

Goal 8.19. Construct a continuous-argument counterpart of $\delta[n]$. *Hint:* follow the construction steps of $\delta_T(t)$ as a reference.

Supplying $X(f) \equiv 1$ into the inverse Fourier transform (8.70b) and recalling that the improper integrals in Fourier transform are understood in the the Cauchy principal value sense yields

$$\begin{aligned}
x(t) &= \int_{-\infty}^{\infty} X(f)e^{2\pi ift}\,\mathrm{d}f = \int_{-\infty}^{\infty} e^{2\pi ift}\,\mathrm{d}f = \lim_{F\to\infty}\int_{-F}^{F} e^{2\pi ift}\,\mathrm{d}f \\
&= \lim_{F\to\infty}\left(\frac{e^{2\pi ift}}{2\pi it}\bigg|_{f=-F}^{F}\right) = \lim_{F\to\infty}\frac{e^{2\pi iFt} - e^{-2\pi iFt}}{2\pi it} \\
&= \lim_{F\to\infty}\left(F\cdot\frac{e^{2\pi iFt} - e^{-2\pi iFt}}{2\pi iFt}\right) = \lim_{F\to\infty}\left(2F\cdot\frac{e^{i\cdot 2\pi Ft} - e^{-i\cdot 2\pi Ft}}{2i\cdot 2\pi Ft}\right) \\
&= \lim_{F\to\infty}\left(2F\cdot\frac{\sin 2\pi Ft}{2\pi Ft}\right) = \lim_{F\to\infty}(2F\operatorname{sinc} 2\pi Ft)
\end{aligned} \tag{8.151}$$

where $[-F, F]$ is the finite integration range (approaching $(-\infty, +\infty)$ as $F \to \infty$) and where

$$\operatorname{sinc} t = \frac{\sin t}{t}$$

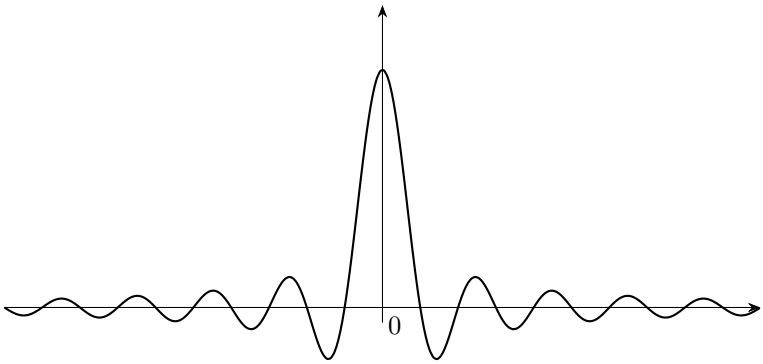


Figure 8.3: Cardinal sine function $\operatorname{sinc} t$.

is the *cardinal sine* function (Fig. 8.3).

▷ As previously, at $t = 0$ we compute the limits at $t \to 0$ instead of evaluating the expressions directly. Thus, in particular, a more correct definition of cardinal sine is

(definition of $\operatorname{sinc} t$***)***

$$\operatorname{sinc} t = \lim_{\tau \to t} \frac{\sin \tau}{\tau}$$

(thereby $\operatorname{sinc} 0 = 1$).

At this point we could introduce the functions $D_F(t)$ such that

$$D_F(2\pi t) = \int_{-F}^{F} e^{2\pi i f t} \, \mathrm{d}f = 2F \operatorname{sinc} 2\pi F t \tag{8.152a}$$

that is, $D_F(t) = 2F \operatorname{sinc} Ft$. Fig. 8.4 illustrates.

▷ In terms of ω we can introduce the functions

$$D_\Omega(t) = \int_{-\Omega}^{\Omega} e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} = \frac{\Omega}{\pi} \operatorname{sinc} \Omega t \tag{8.152b}$$

The functions (8.152) would be the counterparts of Dirichlet kernel functions $D_N(t)$ defined by (8.133). Therefore, while in the discussion of Fourier transform we were obtaining $x(t)$ as a limit of $D_N(2\pi F t)$ at $N \to \infty$, here we would be obtaining $x(t)$ as a limit of $D_F(2\pi t)$ at $F \to \infty$.

▷ From a common mathematical notation standpoint, the notations D_N and D_F look like notations of one and the same entity, instantiated for two different subscript values N and F. We have been a bit nonchalant in using D_N and D_F to denote two different entities, being able to tell between the two by the subscript letter used, or by the fact that N in D_N is an integer value whereas F in D_F is real.

We could also have notated $D_F(t)$ as $D(F, t)$ or $D(t, F)$ to make the notation compliant with the standard notation for integral transform kernels, but we choose to highlight the analogy to the Dirichlet kernel functions $D_N(t)$ instead.

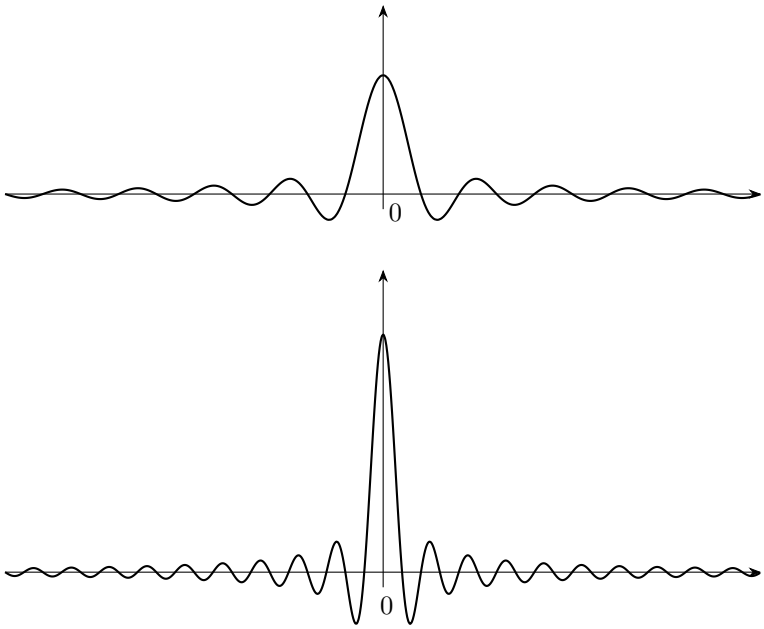


Figure 8.4: Functions $D_F(2\pi t)$ at two different values of F.

Differently from Fourier series case, where we were dealing with periodic functions, and respectively $D_N(2\pi t)$ were 1-periodic, $D_F(2\pi t)$ are not periodic. So the property (8.136a) of $D_N(2\pi t)$ doesn't have a counterpart with $D_F(2\pi t)$. What about (8.136b)? The integral of $D_N(2\pi t)$ over its unit period was equal to 1. What is the integral of $D_F(2\pi t)$ over $(-\infty, +\infty)$?

Goal 8.19a. Find the value of the integral of $D_F(2\pi t)$ over $(-\infty, +\infty)$. *Hint:* use the similarities between $D_N(2\pi t)$ and $D_F(2\pi t)$.

The functions $D_N(2\pi t)$ and $D_F(2\pi t)$ are quite similar, and we can even try to obtain the latter as the limiting case of the former. Specifically, let's consider the the "primary period" $[-\frac{1}{2}, \frac{1}{2}]$ of $D_N(2\pi t)$. We can notice that at large N the non-zero part of the area under of $D_N(2\pi t)$ on this primary period is concentrated around $t = 0$ (as manifested by the properties (8.136b) and (8.136c)). Simultaneously, in the neighborhood of $t = 0$ the denominator of $D_N(2\pi t)$ is well approximated by πt, which essentially has the same form as the denominator of $D_F(2\pi t)$.

To simplify the work a little bit, we can also notice from (8.152) that the area under $D_F(2\pi t)$ on any given interval is independent of F, since there is a multiplier of F applied both to the function's argument and to the function's value. We can therefore choose a particularly convenient value $F = \frac{1}{2}$, in which case

$$D_{\frac{1}{2}}(2\pi t) = D_F(2\pi t)\Big|_{F=\frac{1}{2}} = \operatorname{sinc} \pi t = \frac{\sin \pi t}{\pi t}$$

Let's try bringing $D_N(2\pi t)$ (at large N) to a similar form.

As the first step, let's multiply the argument of $D_N(2\pi t)$ by $F = 1/(2N+1)$:

$$\begin{aligned}
D_N(2\pi F t) &= D_N\left(\frac{2\pi t}{2N+1}\right) = \frac{\sin \pi t}{\sin \frac{\pi t}{2N+1}} = \frac{\sin \pi t}{\pi t} \cdot \frac{\pi t}{\sin \frac{\pi t}{2N+1}} \\
&= (2N+1)\frac{\sin \pi t}{\pi t} \cdot \frac{\frac{\pi t}{2N+1}}{\sin \frac{\pi t}{2N+1}} = (2N+1)\frac{\sin \pi t}{\pi t} \cdot \frac{1}{1+o(t/N)} \\
&= (2N+1)\frac{\sin \pi t}{\pi t} \cdot \big(1+o(t/N)\big) \\
&= (2N+1)D_{\frac{1}{2}}(2\pi t) \cdot \big(1+o(t/N)\big) \qquad (\text{at } t/N \to 0)
\end{aligned}$$

Thereby at small t/N the function $D_N(\frac{2\pi t}{2N+1})$ is approaching $(2N+1)D_{\frac{1}{2}}(2\pi t)$. Respectively, on the intervals where t/N is small, integrals of $D_{\frac{1}{2}}(2\pi t) = \frac{\sin \pi t}{\pi t}$ are $2N+1$ times smaller than integrals of $D_N(\frac{2\pi t}{2N+1})$. On the other hand, integrals of $D_N(\frac{2\pi t}{2N+1})$ are $2N+1$ times larger than the integrals of $D_N(2\pi t)$ taken over a $2N+1$ times smaller interval:

$$\begin{aligned}
\int_{-T}^{T} D_{\frac{1}{2}}(2\pi t)\,\mathrm{d}t &= \frac{1}{2N+1}\int_{-T}^{T}\left(D_N\left(\frac{2\pi t}{2N+1}\right) + o(t/N)\right)\mathrm{d}t \\
&= \int_{-T/(2N+1)}^{T/(2N+1)} \big(D_N(2\pi t) + o(t)\big)\,\mathrm{d}t \\
&= \int_{-T/(2N+1)}^{T/(2N+1)} D_N(2\pi t)\,\mathrm{d}t + o\left((T/N)^2\right) \qquad (\text{at } T/N \to 0)
\end{aligned}$$

Now we want to make the integration interval of $D_N(2\pi t)$ in the latter expression constant and then let $N \to \infty$. Letting $\alpha = T/(2N+1)$,

$$\int_{-(2N+1)\alpha}^{(2N+1)\alpha} D_{\frac{1}{2}}(2\pi t)\,\mathrm{d}t = \int_{-\alpha}^{\alpha} D_N(2\pi t)\,\mathrm{d}t + o(\alpha^2) \qquad (\text{at } \alpha \to 0)$$

For each particular value of α, we can let $N \to \infty$, obtaining

$$\int_{-\infty}^{\infty} D_{\frac{1}{2}}(2\pi t)\,\mathrm{d}t = \int_{-\alpha}^{\alpha} \delta_1(t)\,\mathrm{d}t + o(\alpha^2) = 1 + o(\alpha^2) \qquad (\text{at } \alpha \to 0)$$

where we assumed $0 < \alpha < 1$. Letting $\alpha \to 0$ we obtain

$$\int_{-\infty}^{\infty} D_{\frac{1}{2}}(2\pi t)\,\mathrm{d}t = \int_{-\infty}^{\infty} \frac{\sin \pi t}{\pi t}\,\mathrm{d}t = \int_{-\infty}^{\infty} \operatorname{sinc} \pi t\,\mathrm{d}t = 1$$

Recalling that the integrals of $D_F(2\pi t)$ are independent of F, we have

$$\int_{-\infty}^{\infty} D_F(2\pi t)\,\mathrm{d}t = 1 \tag{8.153a}$$

which is the counterpart of (8.136b).

Goal 8.19a completed. ■

By essentially the same considerations as with D_N, we also obtain the counterpart of (8.136c).:

$$\lim_{F\to\infty} \int_a^b g(t) D_F(2\pi t)\,\mathrm{d}t = 0 \qquad \text{for } 0 \notin [a,b] \tag{8.153b}$$

(for $g(t)$ of bounded speed).

▷ By (8.153a)

$$\int_{-\infty}^{\infty} D_F(2\pi t)\,\mathrm{d}t = 2F\int_{-\infty}^{\infty} \operatorname{sinc} 2\pi F t\,\mathrm{d}t = 1$$

and substituting $F = 1/2\pi$ we obtain

$$\int_{-\infty}^{\infty} \operatorname{sinc} t\,\mathrm{d}t = \pi \tag{8.154}$$

By (8.153a) and (8.153b) we obtain the counterpart of (8.137a)

$$\lim_{F\to\infty}\int_{-\infty}^{\infty} g(t)D_F(2\pi t)\,\mathrm{d}t = g(0) \tag{8.155a}$$

with its refinement

$$\lim_{F\to\infty}\int_{-\infty}^{\infty} g(t)D_F(2\pi t)\,\mathrm{d}t = \frac{g(+0)+g(-0)}{2} \tag{8.155b}$$

which is a counterpart of (8.137b)

We thereby introduce the *Dirac delta* function $\delta(t)$ (the Fourier transform counterpart of $\delta_1(t)$) as an imagined limit of $D_F(2\pi t)$ at $F\to\infty$:

(Dirac delta)

$$\delta(t) = \lim_{F\to\infty} D_F(2\pi t) \tag{8.156}$$

which retains the properties (8.153) and (8.155) of $D_F(2\pi t)$:

(properties of Dirac delta)

$$\int_{-\infty}^{\infty} g(t)\delta(t)\,\mathrm{d}t = g(0) \qquad \text{if } g(t) \text{ is continuous at } t=0 \tag{8.157a}$$

$$\int_{-\infty}^{\infty} g(t)\delta(t)\,\mathrm{d}t = \frac{g(+0)+g(-0)}{2} \qquad \text{generally} \tag{8.157b}$$

Thereby $\delta(t)$ is not a real function, but some kind of concise notational summary of the limiting properties of $D_F(2\pi t) = 2F\operatorname{sinc} 2\pi F t$ at $F\to\infty$.

▷ It follows from (8.157a) that

$$\int_a^b g(t)\delta(t)\,\mathrm{d}t = 0 \qquad \text{if } 0\notin[a,b]$$

and it further follows that

$$\int_a^b g(t)\delta(t)\,\mathrm{d}t = g(0) \qquad \text{if } 0\in(a,b)$$

while for $a=0$ or $b=0$ (and $a<b$) we obtain respectively $g(+0)/2$ or $g(-0)/2$ by (8.157b).

Letting $g(t) \equiv 1$ in (8.157a) yields

(the area under the Dirac delta is unity)

$$\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1$$

which is the limiting case of (8.153a).

▷ Similarly to how $\delta_N[n]$ is representable via $\delta[n]$ by (8.124), we can represent $\delta_1(t)$ via $\delta(t)$ as

$$\delta_1(t) = \sum_{n=-\infty}^{\infty} \delta(t-n) \tag{8.158a}$$

and respectively

(periodic Dirac delta in terms of nonperiodic Dirac delta)

$$\delta_T(t) = \sum_{n=-\infty}^{\infty} \delta(t-nT) = \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta\left(\frac{t}{T}-n\right) \tag{8.158b}$$

Returning to the beginning of our discussion and (8.151), we obtain that by (8.151) and (8.156) the inverse Fourier transform with $X(f) \equiv 1$ "converges" to $x(t) = \delta(t)$. Let's now try the reverse, that is apply the direct Fourier transform formula (8.70a) to $x(t) = \delta(t)$ to see which $X(f)$ we obtain. Substituting $x(t) = \delta(t)$ into (8.70a) yields

$$X(f) = \int_{-\infty}^{\infty} \delta(t) e^{-2\pi i f t}\,\mathrm{d}t = \int_{-\infty}^{\infty} \delta(t) g_f(t)\,\mathrm{d}t$$

where $g_f(t) = e^{-2\pi i k f t}$. By (8.157a) the latter integral is equal to $g_f(0)$, that is simply to 1, and we obtain $X(f) = 1\ \forall f$. Thus we obtain the counterpart of the Fourier series property (8.143):

(spectrum of Dirac delta)

$$X(f) \equiv 1 \iff x(t) = \delta(t) \tag{8.159a}$$

and it's not difficult to realize that the dual of (8.159a) is

(spectrum of a unit constant)

$$x(t) \equiv 1 \iff X(f) = \delta(f) \tag{8.159b}$$

In terms of ω we need to be sensitive about which of the ω-related versions of Fourier transform we pick. E.g. in (8.71) we obviously obtain $X(\omega) = 1$ from $x(t) = \delta(t)$ by (8.71a), but (8.71b) gives $x(t) = \frac{1}{2\pi}\delta(t)$ for $X(\omega) = 1$.

Goal 8.19 completed. ■

Like with the periodic Dirac delta function in (8.145), the nonperiodic Dirac delta function can be used to represent a continuous nonperiodic $x(t)$ as a "linear combination of samples":

(convolution with Dirac delta is transparent)

$$x(t) = \int_{-\infty}^{\infty} x_\tau \delta_\tau(t)\,\mathrm{d}\tau = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)\,\mathrm{d}\tau = x * \delta \tag{8.160}$$

where $\delta_\tau(t) = \delta(t-\tau)$ are "time-shifted" $\delta(t)$. This in turn allows an alternative interpretation of the direct Fourier transform, similarly to (8.146):

$$X(f) = \int_{-\infty}^{\infty} x_\tau \Delta_\tau(f)\,\mathrm{d}\tau = \int_{-\infty}^{\infty} x(\tau)e^{-2\pi i f\tau}\,\mathrm{d}\tau \tag{8.161}$$

where $\Delta_\tau(f) = e^{-2\pi i f\tau}$ are Fourier transforms of $\delta_\tau(t) = \delta(t-\tau)$.

▷ Formula (8.160) can be obtained by a formal application of (8.157a):

$$x * \delta = \int_{-T/2}^{T/2} x(\tau)\delta(t-\tau)\,\mathrm{d}\tau = \int_{-T/2}^{T/2} x(t-\tau)\delta(\tau)\,\mathrm{d}\tau \overset{(8.157a)}{=} x(t)$$

where we can replace $x(\tau)\delta(t-\tau)$ with $x(t-\tau)\delta(\tau)$ under the integral sign by the commutativity of convolution, and where (8.157a) is applied with respect to $x(t-\tau)$ as a function of τ, yielding $x(t)$ at $\tau = 0$.

More generally, by using (8.157b), the formula (8.160) takes the form

(convolution with Dirac delta at a jump discontinuity)

$$x * \delta = \frac{x(+0) + x(-0)}{2} \tag{8.162}$$

which accommodates possible discontinuities of $x(t)$.

8.13.8 Fourier transform roundtrip

Goal 8.20. Study the Fourier transform roundtrip and construct a sufficient condition for it to converge to the original function. *Hint:* use the Fourier series convergence study as a reference.

Similarly to how we used Dirichlet kernel functions $D_N(\cdot)$ to analyze the Fourier series roundtrip in Subsection 8.13.6, the functions $D_F(\cdot)$ can be used to analyze the Fourier transform roundtrip transparency. Let $x(t)$ be the original function and let $x'(t)$ be the roundtrip result. Let's rewrite (8.70b) in terms of Cauchy principal value limit and substitute (8.70a) for $X(f)$, like we did in (8.148):

$$x'(t) = \lim_{F\to\infty} \int_{-F}^{F} X(f)e^{2\pi i f t}\,\mathrm{d}f = \lim_{F\to\infty} \int_{-F}^{F} \left(\int_{-\infty}^{\infty} x(\tau)e^{-2\pi i f\tau}\,\mathrm{d}\tau\right) e^{2\pi i f t}\,\mathrm{d}f \tag{8.163a}$$

In (8.148) the next thing we did was swapping the order of summation and integration, where both the sum and the integral were taken over finite ranges.

Here, however, one of the integrals is over $(-\infty, \infty)$. We still can swap the order of the integrals, but only assuming that $x(\tau)$ is a "nice enough" function. We can require $x(\tau)$ to be absolutely integrable, in which case the improper integral of $x(\tau)e^{-2\pi i f\tau}$ will also converge uniformly. Then

$$\begin{aligned} x'(t) &= \lim_{F\to\infty} \int_{-\infty}^{\infty} \left(\int_{-F}^{F} x(\tau) e^{2\pi i f(t-\tau)} \, \mathrm{d}f \right) \mathrm{d}\tau \\ &= \lim_{F\to\infty} \int_{-\infty}^{\infty} \left(x(\tau) \int_{-F}^{F} e^{2\pi i f(t-\tau)} \, \mathrm{d}f \right) \mathrm{d}\tau \\ &= \lim_{F\to\infty} \int_{-\infty}^{\infty} x(\tau) D_F\big(2\pi(t-\tau)\big) \mathrm{d}\tau = \lim_{F\to\infty} \big(x * D_F(2\pi\cdot)\big) \end{aligned} \tag{8.163b}$$

We have obtained that the "partial integrals" of the inverse Fourier transform (8.70b) are equal to convolutions of $x(t)$ with the functions $D_F(2\pi t)$. In the limit, $D_F(2\pi t)$ turn into $\delta(t)$ and respectively the limit in (8.163b) turns into

$$x'(t) = x * \delta = x(t) \tag{8.164}$$

where the last equality is by (8.160). Notice that (8.164) is the Fourier transform counterpart of (8.149).

▷ For the points of discontinuity we obtain by (8.162)

$$x'(t) = x * \delta = \frac{x(t+0) + x(t-0)}{2} \tag{8.165}$$

The result of (8.164) looks like a confirmation of Fourier transform roundtrip transparency, however we have collected some restrictions of $x(\cdot)$. These were the absolute integrability of $x(\cdot)$ as well as the bounded speed of $x(\cdot)$ These restrictions act as sufficient conditions for the transparency of Fourier transform roundtrip.

▷ Similarly to Fourier series case, we can refer to the question of Fourier transform roundtrip transparency as the question of convergence of Fourier transform.

Goal 8.20 completed. ■

8.13.9 Convergence as filtering

In the transformations (8.151) we have obtained the cardinal sine function $\operatorname{sinc} t$ as the underlying element of the functions $D_F(2\pi t)$ introduced by (8.152). We can notice that the integral in (8.152a) can be interpreted as an inverse Fourier transform of a function which is equal to 1 for $f \in [-F, F]$ and to zero otherwise. We can notate such function as $\operatorname{rect}(f/2F)$ where the *rectangular function* $\operatorname{rect}(t)$ is defined as

$$\operatorname{rect}(t) = \begin{cases} 1 & \text{if } |t| < \frac{1}{2} \\ \frac{1}{2} & \text{if } |t| = \frac{1}{2} \\ 0 & \text{if } |t| > \frac{1}{2} \end{cases} \tag{8.166}$$

Fig. 8.5 illustrates.

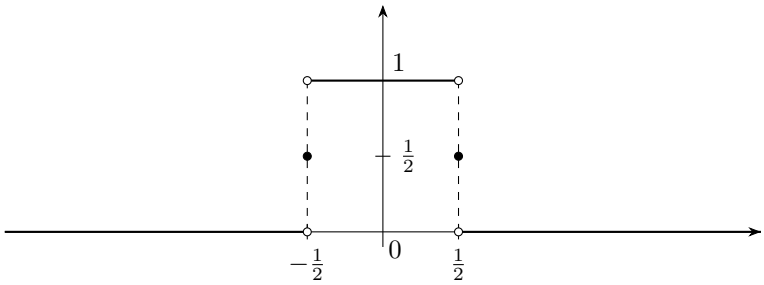


Figure 8.5: Unit rectangular function $\operatorname{rect} t$.

▷ The readers might already have guessed that we let $\operatorname{rect}(\pm\frac{1}{2}) = \frac{1}{2}$ to make sure that the function's value at the points of discontinuity is unchanged by the Fourier transform roundtrip. Indeed, as we shall learn a bit later, Fourier transform, similarly to Fourier series, converges to the average of the left- and right-side values at the points of discontinuity.

Thus, Fourier transform of $D_F(2\pi t)$ is $\operatorname{rect}(f/2F)$:

$$\int_{-\infty}^{\infty} D_F(2\pi t)\, e^{2\pi f t}\, \mathrm{d}t = \operatorname{rect}(f/2F) \tag{8.167}$$

▷ The roundtrip convergence of (8.167) can be justified by noticing that $\operatorname{rect}(f/2F)$ is absolutely integrable and has a zero derivative everywhere except at the two jump points.

▷ By writing $D_F(2\pi t)$ by (8.152a) explicitly in terms of the sinc function, (8.167) turns into

$$\int_{-\infty}^{\infty} (2F \operatorname{sinc} 2\pi F t)\, e^{2\pi f t}\, \mathrm{d}t = \operatorname{rect}(f/2F) \tag{8.168a}$$

which at $F = \frac{1}{2}$ yields

(Fourier transform of sinc is rectangular function)

$$\int_{-\infty}^{\infty} (\operatorname{sinc} \pi t)\, e^{2\pi f t}\, \mathrm{d}t = \operatorname{rect} f \tag{8.168b}$$

Now we can use (8.167) to give another interpretation to the convergence of the inverse Fourier transform integral by (8.163). Comparing (8.163b) to (8.163a) we can notice that convolutions of $x(t)$ with $D_F(2\pi t)$ in the time domain (8.163b) correspond to partial inverse Fourier transform integrals (integration going over $[-F, F]$ instead of $(-\infty, +\infty)$) in (8.163a). By the convolution property of Fourier transform, time-domain convolution should correspond to frequency-domain multiplication, and the latter is exactly what is happening in (8.163a), as we can interpret taking the partial integral over $[-F, F]$ as taking a full integral of the same function multiplied by $\operatorname{rect}(f/2F)$:

$$x'(t) = \lim_{F\to\infty} \int_{-F}^{F} X(f) e^{2\pi i f t}\, \mathrm{d}f = \lim_{F\to\infty} \int_{-\infty}^{\infty} X(f) \operatorname{rect}\left(\frac{f}{2F}\right) e^{2\pi i f t}\, \mathrm{d}f$$

$$= \lim_{F\to\infty} \int_{-\infty}^{\infty} X(f)Y(f)e^{2\pi i f t}\,\mathrm{d}f$$

where

$$Y(f) = \operatorname{rect}\left(\frac{f}{2F}\right)$$

acts as a spectral filter applied to $X(f)$. Respectively, the convolution with $D_F(2\pi t)$ in (8.163b) is just the time-domain representation of filtering by $Y(f)$.

Thus, we can interpret the partial inverse Fourier transform integral over $[-F, F]$ in (8.163a) as applying a *brickwall bandlimiting filter* to the signal $x(t)$, clipping the signal's spectrum to the $[-F, F]$ frequency band, and thereby producing the bandlimited version of $x(t)$. Respectively the limit at $F \to \infty$ is interpreted as letting the band's limiting frequency F grow to infinity, where we expect to obtain the original signal $x(t)$.

The same interpretation can be made for Fourier series, where the partial Fourier series sums can be interpreted as bandlimited versions of the original signal, clipped just to first N harmonics. Equation (8.148) simply gives the time-domain interpretation of that bandlimiting as convolution. Respectively, by letting $N \to \infty$ we expect to obtain the original signal.

8.13.10 Continuous representation of discrete functions

Dirac delta can be used to write sums as integrals. Suppose we are given a sum

$$\sum_n g(t_n) \tag{8.169}$$

where $g(t)$ is some continuous function and the summation goes over some implied finite or infinite range of n. By (8.157a)

$$g(t_n) = \int_{-\infty}^{\infty} g(t)\delta(t - t_n)\,\mathrm{d}t$$

therefore

$$\begin{aligned}\sum_n g(t_n) &= \sum_n \int_{-\infty}^{\infty} g(t)\delta(t - t_n)\,\mathrm{d}t = \int_{-\infty}^{\infty} g(t)\left(\sum_n \delta(t - t_n)\right)\mathrm{d}t \\ &= \int_{-\infty}^{\infty} g(t)\widehat{\delta}(t)\,\mathrm{d}t \end{aligned} \tag{8.170a}$$

where

$$\widehat{\delta}(t) = \sum_n \delta(t - t_n) \tag{8.170b}$$

We have thus rewritten the discrete sum (8.169) as the integral (8.170a).

Furthermore, if t_n is are regularly spaced points $t_n = nT$ where n varies from $-\infty$ to ∞, then the sum in (8.170b) simply turns into a T-periodic Dirac delta by (8.158b):

$$\sum_{n=-\infty}^{\infty} g(nT) = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} g(t)\delta(t - nT)\,\mathrm{d}t = \int_{-\infty}^{\infty} g(t)\left(\sum_{n=-\infty}^{\infty} \delta(t - nT)\right)\mathrm{d}t$$

$$= \int_{-\infty}^{\infty} g(t)\delta_T(t)\,\mathrm{d}t \tag{8.171}$$

▷ We have been a bit nonchalant about swapping the integration and summation in (8.171). The readers are encouraged to meditate on the details, convincing themselves that this transformation is justified. In particular, one needs to decide how to understand the infinite sum of generalized function terms $\delta(t-nT)$ in both (8.158b) and (8.171).

Using the idea expressed by (8.171) we could rewrite the Fourier series (8.64b) as an inverse Fourier transform integral. The terms of the Fourier series in (8.64b) have the form $g_k = X_k e^{2\pi i k F t}$. We can replace g_k by $g(f_k)$ where $f_k = kF$, which is achieved by $g(f) = \tilde{X}(f)e^{2\pi i f t}$ where $\tilde{X}(kF) = \tilde{X}(f_k) = X_k$. Applying (8.171) yields

$$\begin{aligned}
x(t) &= \sum_{k=-\infty}^{\infty} X_k e^{2\pi i k F t} = \sum_{k=-\infty}^{\infty} g_k = \sum_{k=-\infty}^{\infty} g(f_k) = \sum_{k=-\infty}^{\infty} g(kF) \\
&= \int_{-\infty}^{\infty} g(f)\delta_F(f)\,\mathrm{d}f = \int_{-\infty}^{\infty} \left(\tilde{X}(f)e^{2\pi i f t}\right)\delta_F(f)\,\mathrm{d}f \\
&= \int_{-\infty}^{\infty} \left(\tilde{X}(f)\delta_F(f)\right)e^{2\pi i f t}\,\mathrm{d}f = \int_{-\infty}^{\infty} X(f)e^{2\pi i f t}\,\mathrm{d}f
\end{aligned}$$

where $X(f) = \tilde{X}(f)\delta_F(f)$. So $X(f) = \tilde{X}(f)\delta_F(f)$ is Fourier transform of $x(t)$. We can rewrite it as

$$\begin{aligned}
X(f) = \tilde{X}(f)\delta_F(f) &= \tilde{X}(f)\sum_{k=-\infty}^{\infty}\delta(f-kF) = \sum_{k=-\infty}^{\infty}\tilde{X}(f)\delta(f-kF) \\
&= \sum_{k=-\infty}^{\infty}\tilde{X}(kF)\delta(f-kF) = \sum_{k=-\infty}^{\infty} X_k\delta(f-kF)
\end{aligned}$$

Thus, Fourier transform of a periodic function $x(t)$ is

(Fourier series rewritten as Fourier transform)

$$X(f) = \sum_{k=-\infty}^{\infty} X_k\delta(f-kF) \tag{8.172}$$

where X_k are the Fourier series coefficients for $x(t)$.

8.13.11 Normalization checks

Suppose we want to write the convolution property formula for the amplitude version of DFT. It's easy to remember that the property itself has the form

$$u[n] = x * y \iff U[k] = K \cdot X[k]Y[k] \tag{8.173}$$

but we might be not immediately remember the correct value of the normalization coefficient K. We also might be not 100% sure, whether the rule about the normalization coefficient placement matching the one in the transform formulas holds for (8.173).

Delta functions can be quite convenient as test functions for the purposes of finding the correct normalization. E.g. in (8.173) we can let $x[n] = y[n] = \delta_N[n]$, where we also notice that $u[n] = \delta_N * \delta_N = \delta_N[n]$. Applying the amplitude-version direct DFT (8.21a) yields

$$U[k] = X[k] = Y[k] = \frac{1}{N}\sum_{n=0}^{N-1} \delta[n]e^{-2\pi ikn/N} = \frac{1}{N}$$

and substituting into the right-hand side of (8.173) yields

$$\frac{1}{N} = K \cdot \left(\frac{1}{N} \cdot \frac{1}{N}\right)$$

and it follows that $K = N$, which is in agreement with (8.96).

In the same fashion we can try to obtain the normalization for the convolution property of the integral Fourier transform. We again assume

$$u(t) = x * y \iff U(f) = K \cdot X(f)Y(f) \tag{8.174}$$

This time if we try to let $x(t) = y(t) = \delta(t)$ we notice that we're computing a convolution of two Dirac delta functions, but we can only integrate a product of a Dirac delta function with a function of bounded speed, whereas the speed of Dirac delta function is hardly bounded. To get around this problem we can let $x(t) = \delta(t)$, while $y(t)$ can be some other "nice" function. In this case we have $u(t) = y(t)$ and $U(f) = Y(f)$, while by (8.70a)

$$X(f) = \int_{-\infty}^{\infty} \delta(t)e^{-2\pi ift} = 1$$

Substituting into the right-hand side of (8.174) yields

$$Y(f) = K \cdot (1 \cdot Y(f))$$

and it follows that $K = 1$, which is in agreement with (8.106).

Actually, if we ignore the fact that we shouldn't convolve Dirac delta with itself and insist on letting $x(t) = y(t) = \delta(t)$, then we obtain $X(f) = Y(f) \equiv 1$ and by the convolution property (8.106) (or by (8.174), recalling that we just found that $K = 1$) we have $U(f) \equiv 1$. Thereby $u(t) = \delta(t)$ suggesting that

$$\delta * \delta = \delta(t) \tag{8.175}$$

where we notice that (8.175) is in formal agreement with (8.160), which effectively states that $x * \delta = x \ \forall x$.

To be more precise, we can recall that $\delta(t)$ is essentially a limiting version of $D_F(2\pi t)$ given by (8.152). Understanding (8.175) in the limiting sense suggests

$$D_F(2\pi t) * D_F(2\pi t) = D_F(2\pi t) \tag{8.176a}$$

with its frequency-domain counterpart

$$\text{rect}(f/2F) \cdot \text{rect}(f/2F) = \text{rect}(f/2F) \tag{8.176b}$$

But (8.176b) is true by (8.167),[7] therefore (8.176a) is also true, and we obtain (8.175) in the limit. A similar result is obtained if we take separate limits for each of the two functions $D_F(2\pi\cdot)$ involved in the convolution, where we effectively need to take the limit of $D_F(2\pi\cdot) * \delta = D_F(2\pi\cdot)$. Thus, even though $\delta(t)$ doesn't have bounded speed, it seems that we can convolve $\delta(t)$ with itself, again obtaining $\delta(t)$.

Another property where one can use delta functions to find the correct normalization is Parseval's theorem. Let's e.g. reobtain the correct normalization for Parseval's theorem specified for the Fourier transform, where this time let's pick the formula (8.92a) written for the density-like version with ω as the frequency argument. Pretending we do not know correct normalization, we write

$$\int_{-\infty}^{\infty} x^*(t)y(t)\,\mathrm{d}t = K \cdot \int_{-\infty}^{\infty} X^*(\omega)Y(\omega)\,\mathrm{d}\omega \tag{8.177}$$

for an unknown K.

Letting $x(t) = \delta(t)$ and taking an arbitrary "nice" $y(t)$, we have $X(t) \equiv 1$. We want to substitute these into (8.177), where we notice that we need to use the complex conjugate of $\delta(t)$. Intuitively, $\delta^*(t) = \delta(t)$, but, if we're not completely sure, we can swap x and y. Let therefore $x(t)$ be some "nice function", while $y(t) = \delta(t)$ and $Y(t) \equiv 1$. Substituting into (8.177) yields

$$x^*(0) = K \cdot \int_{-\infty}^{\infty} X^*(\omega)\,\mathrm{d}\omega$$

or

$$x(0) = K^* \cdot \int_{-\infty}^{\infty} X(\omega)\,\mathrm{d}\omega$$

Comparing to (8.71b) (where we let $t = 0$) we find that $K = 1/2\pi$, which is in agreement with (8.92a).

8.14 Correlation

The dot product of two vectors allows to compute the cosine of the angle between them (in the literal or in a generalized sense, depending on the linear space):

$$\cos\theta = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\cdot|\vec{y}|} = \frac{\langle\vec{x},\vec{y}\rangle}{\|\vec{x}\|\cdot\|\vec{y}\|} \tag{8.178}$$

We could use the latter expression to measure the "amount of similarity" between two functions x and y:

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\|\cdot\|y\|} \tag{8.179}$$

where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ are as in Parseval's theorem (8.91). We shall consider the situation $\theta = 0$ as the maximum possible similarity between the two functions.

[7] At $f = \pm F$ (8.176b) formally turns into an incorrect equality $\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2}$, but a different value at a single point doesn't affect (inverse) Fourier transform integrals, thus it is not a problem.

It's not difficult to realize that at $\theta = 0$ the functions are simply identical, up to a scaling factor.

We now shall consider the situation when two functions are pretty similar, except that one of the functions is horizontally shifted with respect to the other, therefore evaluating $\langle x, y\rangle$ gives a smaller result than the actual amount of similarity between the two functions. We want therefore to compute $\langle x, y\rangle$ for all possible shifts, which we apply to one of the two functions, say x, so instead of computing a dot product with $x(t)$ we compute it with $x_{\Delta t}(t) = x(t - \Delta t)$. This makes the value $\langle x, y\rangle$ depend on Δt, thus defining a function

$$u(\Delta t) = \langle x_{\Delta t}, y\rangle = \int_{-\infty}^{\infty} x^*(t - \Delta t)y(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} x^*(t)y(t + \Delta t)\,\mathrm{d}t$$

The function $u(\Delta t)$ is referred to as *cross-correlation* of functions x and y. Cross-correlation has a special symbol, similar to the convolution:

$$u(\Delta t) = x \star y = \int_{-\infty}^{\infty} x^*(t - \Delta t)y(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} x^*(t)y(t + \Delta t)\,\mathrm{d}t \tag{8.180}$$

The expression $x \star x$ is referred to as *autocorrelation* of $x(t)$.

> ▷ Obviously, "$\star$" denoting cross-correlation shouldn't be confused with the Hodge star introduced in Chapter 5.

Goal 8.21. Obtain the Fourier transform dual-domain representation of cross-correlation and autocorrelation. *Hint:* reduce the problem to the dual-domain representation of convolution.

The integral expressions in (8.180) look remarkably similar to convolution, in particular we can interpret the first of the two integrals as a convolution with $x^*(-t)$:

$$u(\Delta t) = \int_{-\infty}^{\infty} x^*(t - \Delta t)y(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} x^*(-(\Delta t - t))y(t)\,\mathrm{d}t = x^*(-t) * y(t)$$

That is

$$x \star y = x^*(-t) * y(t) \tag{8.181}$$

By (8.54c), Fourier transform of $x^*(-t)$ is $X^*(f)$, therefore, by the convolution property (8.109), we obtain

(Fourier transform correlation property)

$$u(t) = x \star y \iff U(f) = X^*(f)Y(f) \tag{8.182a}$$

It follows from (8.182a) that the Fourier transform of the autocorrelation $x \star x$ is $|X(f)|^2$:

(Fourier transform autocorrelation property)

$$u(t) = x \star x \iff U(f) = X^*(f)X(f) = |X(f)|^2 \tag{8.182b}$$

Apparently, the property (8.182) has a dual. There are also versions of this property and its duals for all other types of Fourier transform.

▷ Consider (8.182a) at $t = 0$. By definition, at zero shift, crosscorrelation yields the inner product of the two functions: $u(0) = (x \star y)(0) = \langle x, y\rangle$. On the other hand, by the constant property, $u(0) = \int_{-\infty}^{\infty} U(f)\,\mathrm{d}f = \langle X, Y\rangle$, and thus we obtain Parseval's theorem $\langle x, y\rangle = \langle X, Y\rangle$.

Goal 8.21 completed. ■

8.15 Differentiation and integration

Goal 8.22. Obtain the dual-domain representations of differentiation and integration for those Fourier transform types where it makes sense. *Hint:* differentiate or integrate the linear combinations of sinusoids term by term.

Fourier series

Differentiating both sides of the Fourier series formula (8.65b) we obtain

(differentiation of Fourier series)

$$y(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t) = \frac{\mathrm{d}}{\mathrm{d}t}\sum_{k=-\infty}^{\infty} X_k e^{ik\Omega t} = \sum_{k=-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}t}X_k e^{ik\Omega t} = \sum_{k=-\infty}^{\infty}(ik\Omega X_k)e^{ik\Omega t} \tag{8.183}$$

Thereby differentiation multiplies all partials by $ik\Omega$.

▷ In principle, in order to be able to differentiate the sum term by term, we need to require uniform convergence of the series after the differentiation, which is a sufficient condition allowing term by term differentiation. But we do not get into too many rigorous details in this book, concentrating on the general ideas. Also, in practice, the differentiation property (8.183) (as well as its opposite, the integration property (8.185), which we are about to discuss) can hold even in the absence of uniform convergence. We shall see examples of that in Section 8.16.

Respectively, integration produces the opposite effect:

$$\begin{aligned} y(t) &= \int_{t_0}^{t} x(\tau)\,\mathrm{d}\tau = \int_{t_0}^{t}\left(\sum_{k=-\infty}^{\infty} X_k e^{ik\Omega\tau}\right)\mathrm{d}\tau = \sum_{k=-\infty}^{\infty}\int_{t_0}^{t} X_k e^{ik\Omega\tau}\,\mathrm{d}\tau \\ &= \sum_{k=-\infty}^{\infty}\left(\frac{X_k}{ik\Omega}e^{ik\Omega\tau} + C_k\right) \end{aligned} \tag{8.184}$$

however, at this point we notice that there is zero division at $k = 0$. Therefore we can integrate only if $X_0 = 0$ and there is simply no term at $k = 0$. On the other hand, we can realize that the requirement $X_0 = 0$ is necessary anyway in order to ensure the T-periodicity of $y(t)$. Indeed, by (8.64a)

$$X_0 = \frac{1}{T}\int_0^T x(t)\,\mathrm{d}t = y(T) - y(0)$$

Therefore $X_0 \neq 0$ would mean that $y(T) \neq y(0)$ and $y(t)$ is not T-periodic.

We can therefore continue under the assumption $X_0 = 0$ (which also yields $C_0 = 0$):

(integration of zero-constant-term Fourier series)

$$y(t) = \int_{t_0}^{t} \left(\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} X_k e^{ik\Omega\tau} \right) \mathrm{d}\tau = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \left(\frac{X_k}{ik\Omega} e^{ik\Omega t} + C_k \right)$$

$$= \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{X_k}{ik\Omega} e^{ik\Omega t} + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} C_k = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{X_k}{ik\Omega} e^{ik\Omega t} + C \tag{8.185}$$

Thereby integration divides all partials by $ik\Omega$.

▷ The sum of C_k in (8.185) is generally expected to converge because

$$C_k = -\frac{X_k}{ik\Omega} e^{ik\Omega t_0} = O(X_k/k)$$

and X_k are already generally expected to be decaying as $k \to \pm\infty$.

Practically the value of $C = \sum C_k$ can be computed directly by (8.64a), as the arithmetic average of $y(t)$:

$$C = Y_0 = \frac{1}{T} \int_0^T y(t)\,\mathrm{d}t = \frac{1}{T} \int_0^T \int_{t_0}^{t} x(\tau)\,\mathrm{d}\tau\,\mathrm{d}t$$

while the remaining terms of Fourier series of $y(t)$ can be obtained by (8.185).

Fourier transform

Fourier transform has properties similar to (8.183) and (8.185), which can be obtained by differentiating/integrating the integral in the inverse Fourier transform (e.g. in (8.71b)) "term by term", or by taking the limit of (8.183) and (8.185) at $T \to \infty$. This yields

(differentiation and integration of inverse Fourier transform)

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} = \int_{-\infty}^{\infty} i\omega X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} \tag{8.186a}$$

$$\int_{t_0}^{t} x(\tau)\,\mathrm{d}\tau = \int_{t_0}^{t} \int_{-\infty}^{\infty} X(\omega) e^{i\omega\tau} \frac{\mathrm{d}\omega}{2\pi}\,\mathrm{d}\tau = \int_{-\infty}^{\infty} \frac{X(\omega)}{i\omega} e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} + C \tag{8.186b}$$

In (8.186b), instead of $X_0 = 0$ we have the requirement that $X(\omega)$ must approach zero sufficiently quickly as $\omega \to 0$.

Since with Fourier transform both domains are continuous, it is also possible to differentiate and integrate in the other domain:

(differentiation and integration of direct Fourier transform)

$$\frac{\mathrm{d}}{\mathrm{d}\omega} X(\omega) = \frac{\mathrm{d}}{\mathrm{d}\omega} \int_{-\infty}^{\infty} x(t) e^{-i\omega t}\,\mathrm{d}t = \int_{-\infty}^{\infty} (-it) x(t) e^{-i\omega t}\,\mathrm{d}t \tag{8.187a}$$

$$\int_{\omega_0}^{\omega} X(\eta)\,\mathrm{d}\eta = \int_{\omega_0}^{\omega} \int_{-\infty}^{\infty} x(t) e^{-i\eta t}\,\mathrm{d}t\,\mathrm{d}\eta = \int_{-\infty}^{\infty} \frac{x(t)}{-it} e^{-i\omega t}\,\mathrm{d}t + C \tag{8.187b}$$

which are the duals of (8.186).

For f-based Fourier transform the formulas (8.186) and (8.187) attain an additional 2π multiplier. Specifically, (8.186) become

(differentiation and integration of inverse Fourier transform, ordinary frequency version)

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty} X(f)e^{2\pi ift}\,\mathrm{d}f = \int_{-\infty}^{\infty} 2\pi if X(f)e^{2\pi ift}\,\mathrm{d}f \tag{8.188a}$$

$$\int_{t_0}^{t} x(\tau)\,\mathrm{d}\tau = \int_{t_0}^{t}\int_{-\infty}^{\infty} X(f)e^{2\pi if\tau}\,\mathrm{d}f\,\mathrm{d}\tau = \int_{-\infty}^{\infty} \frac{X(f)}{2\pi if}e^{2\pi ift}\,\mathrm{d}f + C \tag{8.188b}$$

and (8.187) become

(differentiation and integration of direct Fourier transform, ordinary frequency version)

$$\frac{\mathrm{d}}{\mathrm{d}f}X(f) = \frac{\mathrm{d}}{\mathrm{d}f}\int_{-\infty}^{\infty} x(t)e^{-2\pi ift}\,\mathrm{d}t = \int_{-\infty}^{\infty} (-2\pi it)x(t)e^{-2\pi ift}\,\mathrm{d}t \tag{8.189a}$$

$$\int_{f_0}^{f} X(\nu)\,\mathrm{d}\nu = \int_{f_0}^{f}\int_{-\infty}^{\infty} x(t)e^{-2\pi i\nu t}\,\mathrm{d}t\,\mathrm{d}\nu = \int_{-\infty}^{\infty} \frac{x(t)}{-2\pi it}e^{-2\pi ift}\,\mathrm{d}t + C \tag{8.189b}$$

One can notice that, if we introduce explicitly distinct notations $X_f(f)$ and $X_\omega(\omega)$, we'd have $X'_f(f) = \frac{\mathrm{d}}{\mathrm{d}f}X_f(f) = \frac{\mathrm{d}}{\mathrm{d}f}X_\omega(2\pi f) = 2\pi X'_\omega(2\pi f) = 2\pi X'_\omega(\omega)$, which explains the extra 2π factor in (8.188a). The remaining occurrences of the extra 2π factor are explained in a similar way.

Discrete-domain differentiation and integration

Discrete-time Fourier transform, having one of the domains continuous, also has counterparts of (8.183) and (8.185) which are their duals, relating the differentiation and integration in the ω domain to the multiplication and division by in in the n domain. We cannot differentiate in the discrete n domain, therefore there is no respective property of the DTFT for multiplication/division by $i\omega$ in the ω domain.

Discrete Fourier transform, with both domains discrete, doesn't have such properties at all.

We could still attempt building approximations of differentiation in a discrete domain, resulting in approximations of the differentiation and integration properties. E.g. we could use $\frac{x[n+1]-x[n-1]}{2}$ as such approximate derivative, yielding in case of complex exponentials

$$\frac{e^{i\omega(n+1)} - e^{i\omega(n-1)}}{2} = \frac{e^{i\omega} - e^{-i\omega}}{2}e^{i\omega n} = i\sin\omega \cdot e^{i\omega n}$$

Respectively the DTFT of $\frac{x[n+1]-x[n-1]}{2}$ is $i\sin\omega\cdot X(\omega)$ which at small ω turns into $i\omega X(\omega)$. Better discrete approximations of differentiation can be built by interpolating the discrete-domain $x[n]$ (by a sufficiently high-order interpolation scheme) and differentiating such interpolations, resulting in closer approximations to $i\omega X(\omega)$ in the other domain.

We could also try to address the problem from the opposite side. Namely, we could ask ourselves, what is the inverse DTFT of $i\omega X(\omega)$. To do that we could treat the inverse DTFT integral (8.59b) as if it is an inverse Fourier transform integral, the variable n being real rather than integer and the function $X(\omega)$ vanishing outside of $[-\pi,\pi]$. This effectively defines a continuous function $\tilde{x}(t)$ (which interpolates the discrete values $x[n]$) as the inverse Fourier transform of such bandlimited $X(\omega)$:

$$\tilde{x}(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{X}(\omega)e^{i\omega t}\,\mathrm{d}\omega$$

where

$$\tilde{X}(\omega)=\operatorname{rect}\left(\frac{\omega}{2\pi}\right)X(\omega)$$

Respectively $i\omega\tilde{X}(\omega)$ is Fourier transform of $\tilde{x}'(t)$ and $i\omega X(\omega)$ is DTFT of $\tilde{x}'(n)$, which can be seen as some kind of "discrete-argument derivative" of $x[n]$.

Goal 8.22 completed. ■

8.16 Discontinuities

In this section we are going to look at Fourier transform and Fourier series of discontinuous functions and study the effect of these discontinuities on the dual domain representation of the same function.

8.16.1 Sawtooth function

We shall begin the analysis of discontinuous functions with piecewise polynomial functions, the simplest case occurring with piecewise linear polynomials. In turn, the simplest periodic function that is simultaneously a piecewise linear polynomial is the sawtooth function, which we now shall introduce and analyze.

Let $x_0(t)=\delta_T(t)$ and consider its Fourier series given by (8.143b):

(Fourier series for Dirac comb)

$$x_0(t)=\delta_T(t)=\frac{1}{T}\sum_{k=-\infty}^{\infty}e^{ik\Omega t}=\frac{1}{T}+\frac{2}{T}\sum_{k=1}^{\infty}\cos k\Omega t$$

Integrating from 0 to t and using (8.184) we obtain

$$x_1(t)=\int_0^t x_0(\tau)\,\mathrm{d}\tau=\frac{1}{T}\sum_{k=-\infty}^{\infty}\frac{1}{ik\Omega}e^{ik\Omega t}+C=\sum_{k=-\infty}^{\infty}\frac{1}{2\pi ik}e^{ik\Omega t}+C$$

where we used $T\Omega=2\pi$. We can notice that we have obtained an infinite term at $k=0$, due to a nonzero constant term $1/T$ in $x_0(t)$.

Let's remove this term from $x_0(t)$, redefining $x_0(t)$ as

(Fourier series for "centered" Dirac comb)

$$x_0(t) = \delta_T(t) - \frac{1}{T} = \frac{1}{T} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} e^{ik\Omega t} \tag{8.190}$$

Now we get

$$x_1(t) = \int_0^t x_0(\tau)\,\mathrm{d}\tau = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{1}{2\pi ik} e^{ik\Omega t} + C = \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{1}{k} \sin k\Omega t + C \tag{8.191}$$

where

$$C = X_1[0] = \frac{1}{T} \int_0^T x_1(t)\,\mathrm{d}t$$

We might be not sure if Fourier series for $x_1(t)$ converges, since the terms only decay as $1/k$ and the series is at best conditionally converging. Let's try to find out what the integral of $\delta_T(t)$ looks like. We know that the integral of $\delta_T(t)$ over the entire period T is 1:

$$\int_0^T \delta_T(t)\,\mathrm{d}t = 1$$

One half of that value must come from the area to the right of $t = 0$:

$$\int_0^\varepsilon \delta_T(t)\,\mathrm{d}t = \frac{1}{2} \quad \forall 0 < \varepsilon < T$$

and the other half from the area from the area to the left of $t = T$:

$$\int_{T-\varepsilon}^T \delta_T(t)\,\mathrm{d}t = \frac{1}{2} \quad \forall 0 < \varepsilon < T$$

and it follows that

$$\int_0^t \delta_T(t)\,\mathrm{d}t = \frac{1}{2} \quad \forall 0 < t < T$$

On the other hand, integration of the term $-1/T$ in $x_0(t)$ in (8.190) contributes the term t/T to the same integral and we obtain

$$x_1(t) = \int_0^t x_0(t)\,\mathrm{d}t = \int_0^t \left(\delta_T(t) - \frac{1}{T}\right) \mathrm{d}t = \frac{1}{2} - \frac{t}{T} \quad \forall 0 < t < T$$

The function $x_1(t)$ looks as shown in Fig. 8.6. It's referred to as the (falling) "sawtooth" function.

We can immediately notice from Fig. 8.6 that the average value of the sawtooth over one period is zero, therefore $C = X_1[0] = 0$ in (8.191) and we can rewrite (8.191) as

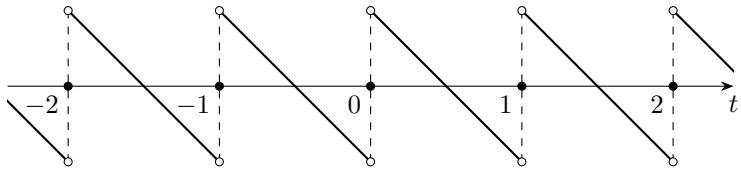


Figure 8.6: Falling sawtooth at $T = 1$.

(Fourier series for falling sawtooth function)

$$x_1(t) = \int_0^t x_0(\tau)\,\mathrm{d}\tau = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{1}{2\pi i k} e^{ik\Omega t} = \frac{1}{\pi}\sum_{k=1}^{\infty} \frac{1}{k}\sin k\Omega t \tag{8.192}$$

From Fig. 8.6 we also see that $x_1(t)$ has limited speed and a single discontinuous jump on each period. Therefore upon convolution with $\delta_T(t)$ in (8.149) it should be unchanged, respectively its Fourier series (8.192) should converge back to $x_1(t)$. At the discontinuous jumps, by (8.165) it converges back to the average of the values on either side of the jump, that's why we let $x_1(t) = 0$ $\forall t = nT$, where $n \in \mathbb{Z}$. As mentioned, a series, whose terms are decaying as $1/k$, cannot converge absolutely, therefore it converges conditionally.

▷ We have just established that Fourier series for the sawtooth function converges, but due to the lack of established uniform convergence we might not be certain if the integration property applies and whether the sawtooth's Fourier series terms obtained by the integration property are correct. We can reobtain them explicitly. By (8.65a) and assuming $k \neq 0$ we have

$$\begin{aligned}
X_1[k] &= \frac{1}{T}\int_0^T x_1(t)e^{-ik\Omega t}\,\mathrm{d}t = \frac{1}{T}\int_0^T \left(\frac{1}{2} - \frac{t}{T}\right) e^{-ik\Omega t}\,\mathrm{d}t \\
&= 0 - \frac{1}{T}\int_0^T \frac{t}{T}\cdot e^{-ik\Omega t}\,\mathrm{d}t = -\frac{1}{T^2}\int_{t=0}^T t\frac{\mathrm{d}(e^{-ik\Omega t})}{-ik\Omega} \\
&= \frac{1}{2\pi i k T}\int_{t=0}^T t\cdot \mathrm{d}(e^{-ik\Omega t}) = \frac{1}{2\pi i k T}\left(te^{-ik\Omega t}\Big|_{t=0}^T - \int_{t=0}^T e^{-ik\Omega t}\,\mathrm{d}t\right) \\
&= \frac{1}{2\pi i k T}te^{-ik\Omega t}\Big|_{t=0}^T - 0 = \frac{1}{2\pi i k T}T = \frac{1}{2\pi i k}
\end{aligned}$$

while for $k = 0$ we have already found out that $X_1[0] = 0$. Thus, despite the lack of established uniform convergence, we obtained correct Fourier series coefficients by integration.

The series (8.192) converges conditionally, thus we don't have absolute convergence, which would have allowed us to infer uniform convergence of the series. It turns out that, indeed, the series (8.192) doesn't converge uniformly, but only pointwise. Visually this manifests itself in the *Gibbs phenomenon*, which is demonstrated by Fig. 8.7.

As we sum up more and more terms of the infinite Fourier series for the sawtooth, the resulting curve still stays somewhat away from the true sawtooth shape. One can see in Fig. 8.7 that the ripples only "get horizontally

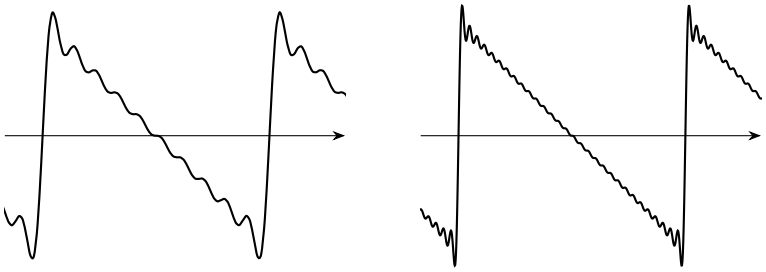


Figure 8.7: Gibbs phenomenon. Partial sums of Fourier series for sawtooth with first 10 and first 30 terms respectively.

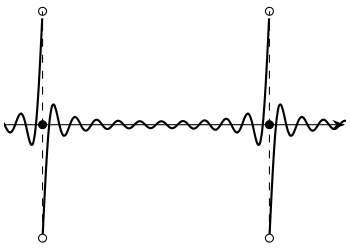


Figure 8.8: Gibbs phenomenon residual. Deviation of the partial sum (with 10 first terms) of the Fourier series (8.192) from the true function.

compressed towards the jump" but their maximum vertical amplitude doesn't really decrease. The series still does converge to the true sawtooth, because at any given point the ripple amplitude does decay towards zero, which means the series *converges pointwise*. However the maximum ripple amplitude doesn't decrease to zero, thus the series doesn't converge uniformly.

Actually the maximum approximation error occurs not at the highest ripple, but close to the points where the sawtooth abruptly changes sign (Fig. 8.8). The value of the partial sum is approximately zero in some neighborhood of any such point, while the true function changes from maximally large positive value to maximally large negative value. Thus the error there is basically as large as the sawtooth's amplitude.

8.16.2 Periodic piecewise parabolic function

Integrating the sawtooth Fourier series (8.192) once again term by term we obtain

$$x_2(t) = \int_0^t x_1(\tau)\,\mathrm{d}\tau = \sum_{k\neq 0} \frac{1}{-2\pi k^2\Omega} e^{ik\Omega t} + C = \sum_{k\neq 0} \frac{T}{-4\pi^2 k^2} e^{ik\Omega t} + C$$

where

$$C = X_2[0] = \frac{1}{T}\int_0^T x_2(t)\,\mathrm{d}t = \frac{1}{T}\int_0^T \left(\int_0^t x_1(\tau)\,\mathrm{d}\tau\right)\,\mathrm{d}t$$

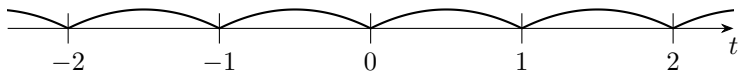


Figure 8.9: Piecewise parabolic function (8.193) at $T = 1$.

$$= \frac{1}{T}\int_0^T \int_0^t \left(\frac{1}{2} - \frac{t}{T}\right) \mathrm{d}\tau\, \mathrm{d}t = \frac{1}{T}\int_0^T \left(\frac{t}{2} - \frac{t^2}{2T}\right) \mathrm{d}t = \frac{1}{T}\left(\frac{t^2}{4} - \frac{t^3}{6T}\right)\Bigg|_{t=0}^{T}$$
$$= \frac{1}{T}\left(\frac{T^2}{4} - \frac{T^3}{6T}\right) = \frac{T}{12}$$

and

(Fourier series for periodic piecewise parabolic function)

$$x_2(t) = \int_0^t x_1(\tau)\,\mathrm{d}\tau = T \cdot \left(\frac{1}{12} + \sum_{k\neq 0} \frac{1}{-4\pi^2 k^2} e^{ik\Omega t}\right)$$
$$= T \cdot \left(\frac{1}{12} - \sum_{k=1}^{\infty} \frac{1}{2\pi^2 k^2} \cos k\Omega t\right) \tag{8.193}$$

Fig. 8.9 illustrates.

Apparently, linear segments of the sawtooth are turned into parabolic segments upon integration, while the jump discontinuity in (8.192) has been converted to a derivative discontinuity (a breakpoint) in (8.193). At the same time the terms of (8.193) decay as $1/k^2$. We might be getting an idea that the decay rate of the terms of Fourier series is related to the continuity of the function.

Another consideration supporting that conjecture is the following. Fourier series whose coefficients X_k decay as $1/k^2$ is converging absolutely and uniformly (with respect to t). On the other hand, each Fourier series term is a continuous function and so is any partial sum of the Fourier series. By the uniform convergence, the full sum of the Fourier series should be continuous as well. The only possibility to have a discontinuous full sum is nonuniform convergence. This is what is happening e.g. in (8.192), where the convergence is conditional, but cannot happen in (8.193) where the convergence is absolute.

On the Gibbs phenomenon side we also notice that this time the ripple amplitude decays to zero as the number of the terms of the series grows (Fig. 8.10), which is the visual counterpart of the uniform convergence.

8.16.3 Periodic piecewise polynomial functions

Goal 8.23. Generalize the previous results to arbitrary periodic piecewise polynomial functions.

By repeated integration of (8.193) (each time throwing out the constant term of the series, so that the integration delivers a periodic function) we can

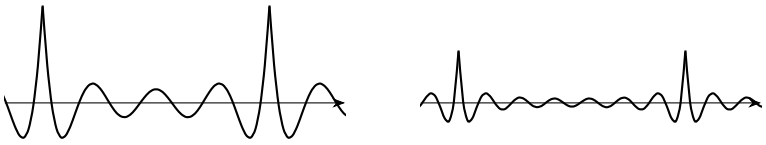


Figure 8.10: Gibbs phenomenon residual for breakpoints. Deviation of the partial sums (with 3 and 6 first terms) of the Fourier series (8.193) from the true function.

obtain the functions $x_3(t) = \int x_2(t)\,\mathrm{d}t$, $x_4(t) = \int x_3(t)\,\mathrm{d}t$, etc. More precisely,

$$x_{n+1}(t) = \int_{t_0}^{t} \big(x_n(t) - X_n[0]\big)\mathrm{d}t$$

Each function $x_n(t)$ turns out to be an n-th order polynomial function with a discontinuity in the $(n-1)$-th derivative at $t = nT$ ($n \in \mathbb{Z}$), while the Fourier series coefficients decay as $1/k^n$.

Now suppose we are having some arbitrary T-periodic piecewise polynomial function $y(t)$ (of a finite order) with some jump discontinuity. We can pick the sawtooth function $x_1(t)$ of the same period and adjust the amplitude and horizontal position of that sawtooth to make sure the jump in the sawtooth is aligned with the jump in $y(t)$ and that both jumps have the same amplitude. If we then subtract this sawtooth from $y(t)$, we thereby eliminate the respective jump in $y(t)$:

$$y_1(t) = y(t) - a_1 x_1(t - t_1)$$

where a_1 and t_1 specify the amplitude and the alignment of $x_1(t)$. If there is another jump in $y(t)$, this jump will be still present in $y_1(t)$ and we can eliminate it in a similar fashion:

$$y_2(t) = y_1(t) - a_2 x_1(t - t_2)$$

Suppose there are no more jumps in $y_2(t)$, but there is a breakpoint. We can eliminate this breakpoint by subtracting the piecewise parabolic function $x_2(t)$ (given by (8.193)):

$$y_3(t) = y_2(t) - a_3 x_2(t - t_3)$$

This way we can eliminate all breakpoints. If there are still remaining discontinuities in higher-order derivatives, we can eliminate those by subtracting the functions $x_3(t)$, $x_4(t)$, etc. At some point all discontinuities have been eliminated (assuming there is a finite number of those) and we obtain a function

$$\bar{y}(t) = y(t) - \sum_n a_n x_1(t - t_n) - \sum_n a_n x_2(t - t_n) - \ldots - \sum_n a_n x_M(t - t_n)$$

where the total number of terms if finite, where each summation symbol runs over a different range of n corresponding to the points of discontinuity of the respective derivative, and where M is defined by the highest order of the discontinuous derivative of $y(t)$. Now $\bar{y}(t)$ is a periodic polynomial function of a finite order without any discontinuities in the function itself or in any of its

derivatives. It follows that this function is constant, since any non-constant polynomial $P(t)$ (of a finite order) cannot be periodic. Denoting this constant by the same symbol $\bar{y}$ (so that $\bar{y}(t) = \bar{y}$) we obtain

$$y(t) = \bar{y} + \sum_n a_n x_1(t - t_n) - \sum_n a_n x_2(t - t_n) - \ldots - \sum_n a_n x_M(t - t_n)$$

or, more formally written

$$y(t) = \bar{y} + \sum_{m=1}^{M} \sum_{n=1}^{N_m} a_{mn} x_m(t - t_{mn}) \tag{8.194}$$

where we relabelled a_n and t_n as a_{mn} and t_{mn} and N_m is the number of discontinuities corresponding to $x_m(\cdot)$.

Thus $y(t)$ can be represented as a linear combination of functions $x_m(t-t_{mn})$ plus the constant term. Respectively the Fourier series terms of $y(t)$ are linear combinations of the respective Fourier terms of $x_m(t - t_{mn})$. Now recall that Fourier series terms of $x_m(t)$ decay as $1/k^m$. By the time-shift property (8.114a), Fourier series coefficients of $x_m(t - t_{mn})$ also decay as $1/k^m$.

It follows that the Fourier series coefficients of $y(t)$ decay as $1/k$ (since $1/k$ is the slowest order of decay among $1/k^m$), provided there is at least one component of the form $a_{1n}x_1(t-t_{1n})$, that is provided $N_1 \geq 1$, that is provided $y(t)$ has at least one jump discontinuity. If there are no jump discontinuities but only breakpoints, then the slowest order of decay is $1/k^2$, occurring due to the terms of the form $a_{2n}x_2(t - t_{2n})$. Etc. Thus *Fourier series coefficients of any piecewise polynomial periodic function decay as* $1/k^m$, where m is defined by the strongest type of discontinuity present in the function.

We have also seen that the decay rate of Fourier series coefficients is related to the Fourier series convergence and to the Gibbs phenomenon. We have seen that with the jump type discontinuity we are having $1/k$ decay rate, respectively there is no uniform but only pointwise convergence, which manifests itself as a non-decaying amplitude of the Gibbs ripples. With the breakpoint-type discontinuity (more precisely, with a jump discontinuity in the 1st derivative) we had $1/k^2$ decay rate, uniform convergence and decaying amplitude of the Gibbs ripples. With a jump discontinuity in the 2nd derivative we are going to have $1/k^3$ decay rate, as well as the uniform convergence and the Gibbs ripple decay. Etc.

Conversely, if we go into the opposite direction from the jump type discontinuity, we should expect opposite effects. Indeed, by differentiating the sawtooth function we obtain back the periodic Dirac delta $\delta_T(t)$ (plus a constant offset). Expanding $\delta_T(t)$ into Fourier series, we recall that by (8.143b) the Fourier series coefficients are constant, that is they decay as $1/k^0$. As the partial sums of Fourier series for $\delta_T(t)$ go, we can realize that these partial sums are simply given by the Dirichlet kernels $D_N(2\pi t)$. Respectively we notice that the "Gibbs ripples" in the Dirichlet kernels $D_N(2\pi t)$ not only do not decay, but actually become higher with N, as illustrated by Fig. 8.2. Specifically, the main peak itself, as well as the ripples immediately next to the main peak become higher in amplitude as N grows. We also do not get even pointwise convergence this time (as the readers might recall, the convergence result $\delta_T(t)$ is even not a function in the normal sense).

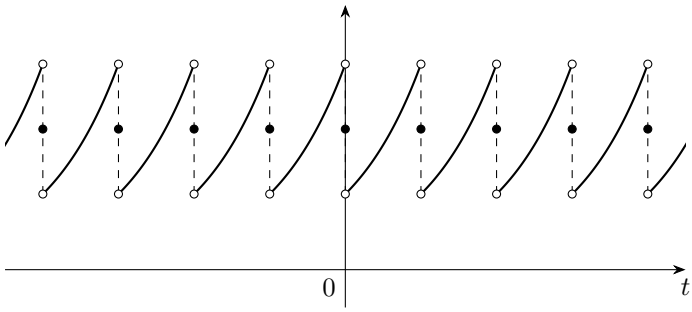


Figure 8.11: Periodic piecewise exponential function.

Goal 8.23 completed. ■

8.16.4 Periodic arbitrary piecewise functions

Goal 8.24. Generalize the previous results to nonpolynomial periodic piecewise functions.

We have seen that the decay rate of Fourier series coefficients of piecewise polynomial periodic functions is defined solely by the discontinuities present in the function. Similar observation tends to hold for many other periodic functions. E.g. let's consider a periodic piecewise exponential defined by

$$y(t) = \begin{cases} e^{st} & \text{if } t \in (0,1) \\ \dfrac{e^s - 1}{2} & \text{if } t = 0 \end{cases} \tag{8.195}$$

(where $s \in \mathbb{C}$ is some constant) and being periodically extended outside of $[0,1)$. Fig. 8.11 illustrates for $s \in \mathbb{R}$. Notice that the function (8.195) is discontinuous in all its derivatives at integer values of t, not just jump-discontinuous in its value. In that sense it has an infinite number of discontinuities (one discontinuity in each derivative) at each integer point $t \in \mathbb{Z}$.

By (8.65a)

$$\begin{aligned} Y_k &= \int_0^1 e^{st} e^{-ik\Omega t}\,\mathrm{d}t = \int_0^1 e^{st} e^{-2\pi ikt}\,\mathrm{d}t = \int_0^1 e^{(s-2\pi ik)t}\,\mathrm{d}t = \left.\frac{e^{(s-2\pi ik)t}}{s-2\pi ik}\right|_{t=0}^{1} \\ &= \frac{e^{s-2\pi ik}-1}{s-2\pi ik} = \frac{e^s e^{-2\pi ik}-1}{s-2\pi ik} = \frac{e^s-1}{s-2\pi ik} \end{aligned}$$

Apparently, Y_k decay as $1/k$ at $k \to \infty$ (unless $s = 2\pi in$ for some $n \in \mathbb{Z}$, which corresponds to a $1/n$-periodic complex sinusoid, which thereby doesn't have any discontinuities).

▷ The obtained result can be reused for piecewise real sinusoids (with a single discontinuous jump on a period) by representing a sinusoid as a sum of two exponentials of an imaginary argument, yielding the same decay rate of $1/k$.

Let's now try to generalize the just obtained result to other periodic functions with jump discontinuities. That is, we conjecture that a wide class of periodic

functions with jump discontinuities exhibits the same property: the decay rate of Fourier series coefficients of such functions is $1/k$.

In order to support the above conjecture, one could consider expanding the given function into Taylor series (between each pair of neighboring points of discontinuity), which suggests that the function is being seen as a limiting case of a finite-order polynomial, as the polynomial order grows to infinity. In this case (8.194) potentially becomes an infinite sum at $M \to \infty$, since the total number of derivative discontinuities can be infinite, as it was the case with (8.195). Respectively we never obtain the remainder function $\bar{y}(t)$. Even if the total number of derivative discontinuities is finite, we cannot argue that after removing all discontinuities we obtain a constant function $\bar{y}(t)$, this function can be some non-constant periodic function instead. Conversely, with an infinite number of derivative discontinuities, there can be questions whether the infinite sum in (8.194) converges at all.

We could however consider a limited version of (8.194) where only the jump discontinuities of the function itself (but not of its derivatives) are separated from the function:

$$y(t) = \bar{y}(t) + \sum_{n=1}^{N_1} a_{1n} x_1(t - t_{1n}) \tag{8.196}$$

In this case $\bar{y}(t)$ is not a constant, but a "remainder" function. Since we eliminated all jump discontinuities of the function by taking them into a separate sum, $\bar{y}(t)$ is continuous. Let's take one additional assumption: Fourier series of $\bar{y}'(t)$ converges (this is quite a reasonable assumption, e.g. it would suffice that $\bar{y}''(t)$ is bounded except at a finite set of points of jump discontinuity of $\bar{y}'(t)$). Then the Fourier series coefficients $\bar{Y}_k^{(\prime)}$ of $\bar{y}'(t)$ must decay to zero, effectively being the coefficients of a converging series. By the integration property of Fourier series, the coefficients $\bar{Y}_k$ of $\bar{y}(t)$ satisfy $\bar{Y}_k = \bar{Y}_k^{(\prime)}/ik\omega$ (notice that the periodicity of $\bar{y}(t)$ implies that the constant term of the Fourier series of $\bar{y}'(t)$ is zero, therefore the integration property is directly applicable). Combining $\bar{Y}_k = \bar{Y}_k^{(\prime)}/ik\omega$ with $\bar{Y}_k^{(\prime)} \to 0$ yields $\bar{Y}_k = o(1/k)$. Therefore, converting (8.196) to Fourier series yields

$$Y_k = \bar{Y}_k + \frac{C}{k} = o(1/k) + \frac{C}{k}$$

where C is a constant dependent on a_{1n} and t_{1n}. Thus Y_k decay as $1/k$.

If the function $y(t)$ is continuous, but its derivative $y'(t)$ has jump discontinuities, we can apply the same argument to $y'(t)$, yielding $1/k$ decay rate for the Fourier series coefficients of $y'(t)$, and the $1/k^2$ decay rate for the Fourier series coefficients of $y(t)$ follows by the integration property of the Fourier series (notice that the periodicity of $y(t)$ implies that the constant term of the Fourier series of $y'(t)$ is zero, therefore the integration property is directly applicable). Repeating the same argument over and over for higher-order derivatives of $y(t)$ we obtain that the Fourier series coefficient decay rate property of piecewise polynomials holds for most of the other functions as well: it is determined by the strongest type of discontinuity.

▷ Apparently, there is a dual observation for discrete-time Fourier transform, relating the the decay rate of $x[n]$ to the continuity properties of its transform.

Goal 8.24 completed. ■

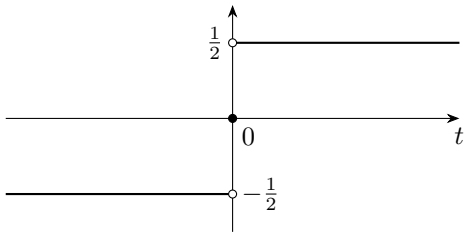


Figure 8.12: (Centered) Heaviside step function $\frac{1}{2}\operatorname{sgn} t = \int_0^t \delta(\tau)\,\mathrm{d}\tau$.

8.16.5 Discontinuities in nonperiodic functions

Goal 8.25. Generalize the previous results to nonperiodic functions.

Fourier transform exhibits similar behavior in regards to continuity properties in one domain and the rolloff speed[8] in the other domain. It's a bit more difficult to construct demonstrative examples in this case, since integration of $\delta(t)$ produces a step function with a jump discontinuity at $t = 0$:

$$x(t) = \int_0^t \delta(\tau)\,\mathrm{d}\tau = \frac{1}{2}\operatorname{sgn} t \tag{8.197}$$

where

$$\operatorname{sgn} t = \begin{cases} 1 & \text{if } t \geq 0 \\ -1 & \text{if } t < 0 \\ 0 & \text{if } t = 0 \end{cases}$$

Fig. 8.12 illustrates.

Computing direct Fourier transform (8.71a) of (8.197) is going to be a challenge, as it doesn't look like the integral in (8.71a) is going to converge in a normal sense, since the function doesn't even decay to zero at $t \to \pm\infty$. Apparently, adding or subtracting a constant term to $\delta(t)$, similarly to what we did in (8.190) is not going to help with that. We could however "bend" the function (8.197), making it absolutely integrable, by replacing it with

$$x(t) = \begin{cases} e^{-t} & \text{if } t \geq 0 \\ -e^{t} & \text{if } t < 0 \\ 0 & \text{if } t = 0 \end{cases} \tag{8.198}$$

where we also, for the sake of simplicity, doubled the amplitude of the jump, since we're anyway only doing qualitative analysis. Fig. 8.13 shows the graph of the function (8.198).

Taking the direct Fourier transform of (8.198) gives by (8.71a)

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}\,\mathrm{d}t = \int_0^{\infty} e^{-t}e^{-i\omega t}\,\mathrm{d}t - \int_{-\infty}^{0} e^{t}e^{-i\omega t}\,\mathrm{d}t$$

[8] In regards to the spectra, especially continuous ones, the decaying at infinity is often referred to as *rolloff*, since the spectrum defines the amplitudes of the sinusoidal components that make up the signal.

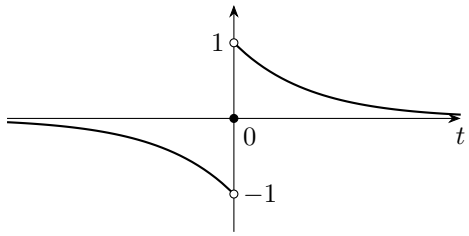


Figure 8.13: Exponentially bent step function (of step amplitude 2).

$$
\begin{aligned}
&= \int_0^\infty e^{-(1+i\omega)t}\,\mathrm{d}t - \int_{-\infty}^0 e^{(1-i\omega)t}\,\mathrm{d}t \\
&= \frac{e^{-(1+i\omega)t}}{-(1+i\omega)}\bigg|_{t=0}^{\infty} - \frac{e^{(1-i\omega)t}}{(1-i\omega)}\bigg|_{t=-\infty}^{0} \\
&= \left(0 - \frac{1}{-(1+i\omega)}\right) - \left(\frac{1}{1-i\omega} - 0\right) = \frac{1}{1+i\omega} - \frac{1}{1-i\omega} \\
&= \frac{(1-i\omega)-(1+i\omega)}{1+\omega^2} = \frac{-2i\omega}{1+\omega^2}
\end{aligned}
$$

which clearly rolls off as $1/\omega$ at $\omega \to \pm\infty$.

It seems we might have a similar correspondence between the continuity and the rolloff like we had with Fourier series. Let's try a higher speed rolloff. Integrating (8.198) from $-\infty$ gives, for $t \le 0$

$$
y(t) = \int_{-\infty}^t x(\tau)\,\mathrm{d}\tau = \int_{-\infty}^t (-e^\tau)\,\mathrm{d}\tau = (-e^\tau)\bigg|_{\tau=-\infty}^{t} = -e^t - 0 = -e^t
$$

while for $t > 0$ we obtain

$$
\begin{aligned}
y(t) &= y(0) + \int_0^t x(\tau)\,\mathrm{d}\tau = -e^0 + \int_0^t e^{-\tau}\,\mathrm{d}\tau = -1 + (-e^{-\tau})\bigg|_{\tau=0}^{t} \\
&= -1 + (-e^{-t} - (-1)) = -e^{-t}
\end{aligned}
$$

Thus

$$
y(t) = \int_0^t x(\tau)\,\mathrm{d}\tau = -e^{-|t|}
$$

Fig. 8.14 illustrates the graph of $y(t)$, where we can see that the jump discontinuity has turned into a breakpoint.

Noticing that $X(\omega)$ is asymptotically proportional to ω at $\omega \to 0$, we can apply the integration property (8.186b) yielding

$$
Y(\omega) = \frac{X(\omega)}{i\omega} = \frac{-2i\omega}{i\omega(1+\omega^2)} = \frac{-2}{1+\omega^2}
$$

where we notice, that, as expected, $Y(\omega)$ rolls off as $1/\omega^2$ at $\omega \to \pm\infty$.

Thus we observe that a jump discontinuity results in the spectrum rolloff of order $1/\omega$, a breakpoint discontinuity results in the rolloff of order $1/\omega^2$, etc. The Dirac delta function's spectrum is constant, that is it rolls off as $1/\omega^0$.

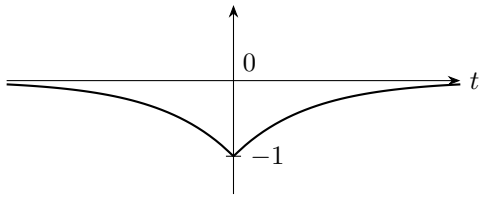


Figure 8.14: Exponentially bent breakpoint function (of "breakpoint amplitude" 2).

▷ This relationship doesn't 100% work in the opposite direction (rolloff speed defining the type of discontinuities). E.g. the inverse Fourier transform of

$$X(\omega) = \frac{1}{1+|\omega|}$$

diverges at $t = 0$ instead of having a jump discontinuity.

▷ Since both domains are continuous now, there is a dual relationship between the continuity of $X(\omega)$ and the rolloff speed of $x(t)$.

Goal 8.25 completed. ■

8.16.6 Integral sine function

While the direct Fourier transform integral (8.71a) doesn't converge for the centered Heaviside step function (8.197) (Fig. 8.12), there are generalizations of the integral convergence which allow to compute direct Fourier transforms of functions like that. The details of these generalizations fall outside the scope of this book, but we still can obtain Fourier transform of (8.197) by using the integration property of Fourier transform.

Fourier transform of $\delta(t)$ is 1, therefore Fourier transform of the step function (8.197) (the latter being the integral of $\delta(t)$) must be $1/i\omega$. Of course there are problems with the applicability of the integration property, e.g. the area under $\delta(t)$ being nonzero, but let's ignore it for a moment. Taking the inverse Fourier transform of $1/i\omega$ by (8.71b) and recalling that the integral is taken in the Cauchy principal value sense yields

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{e^{i\omega t}}{i\omega}\frac{\mathrm{d}\omega}{2\pi} &= \int_0^{\infty}\left(\frac{e^{i\omega t}}{i\omega} + \frac{e^{-i\omega t}}{-i\omega}\right)\frac{\mathrm{d}\omega}{2\pi} = \int_0^{\infty}\frac{e^{i\omega t} - e^{-i\omega t}}{i\omega}\frac{\mathrm{d}\omega}{2\pi} \\ &= \frac{1}{\pi}\int_0^{\infty}\frac{\sin\omega t}{\omega}\,\mathrm{d}\omega = \frac{1}{\pi}\int_{\omega=0}^{\infty}\frac{\sin\omega t}{\omega t}\,\mathrm{d}(\omega t) = \frac{1}{\pi}\int_{\omega=0}^{\infty}\operatorname{sinc}\omega t\,\mathrm{d}(\omega t) \\ &= \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\operatorname{sinc}\omega t\,\mathrm{d}(\omega t) = \frac{\operatorname{sgn} t}{2\pi}\int_{-\infty}^{\infty}\operatorname{sinc}\tau\,\mathrm{d}\tau \overset{(8.154)}{=} \frac{\operatorname{sgn} t}{2\pi}\cdot\pi = \frac{1}{2}\operatorname{sgn} t\end{aligned} \tag{8.199}$$

where at $t = 0$ we simply take the average of the values at $t = +0$ and $t = -0$.

Thus, the inverse Fourier transform of $1/i\omega$ is $\frac{1}{2}\operatorname{sgn} t$. Respectively we can consider $1/i\omega$ to be the direct Fourier transform of $\frac{1}{2}\operatorname{sgn} t$. Even though the direct transform formula (8.71a) doesn't converge, the main idea of Fourier

transform (at least from the point of view presented in this book) is to expand a function into a (finite or infinite, countable or uncountable) linear combination of sinusoids, and this expansion is represented by the inverse Fourier transform formula. It is in this sense that we can understand $1/i\omega$ as the Fourier transform of $\frac{1}{2}\operatorname{sgn} t$.

So the spectrum of $\frac{1}{2}\operatorname{sgn} t$ still rolls off as $1/\omega$, similarly to (8.198), which is in agreement with the presence of a jump discontinuity in $\frac{1}{2}\operatorname{sgn} t$. What about the Gibbs phenomenon? Until now we have been considering Gibbs phenomenon only in connection with Fourier series, where we have been taking partial sums of Fourier series with progressively increasing number of terms. The Fourier transform counterpart of this is computing "partial improper integrals" of the inverse Fourier transform:

$$\int_{-\Omega}^{\Omega} X(\omega)e^{i\omega t}\,\frac{\mathrm{d}\omega}{2\pi} \tag{8.200}$$

Letting $\Omega \to \infty$ corresponds to letting $N \to \infty$ in the partial sums of Fourier series.

▷ This is pretty much the same what we have been doing in Subsection 8.13.9, where we were considering the convergence of Fourier series and Fourier transform as bandlimiting of the signal to progressively wider and wider frequency band.

Letting $X(\omega) = 1/i\omega$ in the partial inverse Fourier transform integral (8.200) yields

$$\begin{aligned} x_\Omega(t) &= \int_{-\Omega}^{\Omega} \frac{e^{i\omega t}}{i\omega}\,\frac{\mathrm{d}\omega}{2\pi} = \int_0^{\Omega}\left(\frac{e^{i\omega t}}{i\omega} + \frac{e^{-i\omega t}}{-i\omega}\right)\frac{\mathrm{d}\omega}{2\pi} = \int_0^{\Omega} \frac{e^{i\omega t} - e^{-i\omega t}}{i\omega}\,\frac{\mathrm{d}\omega}{2\pi} \\ &= \frac{1}{\pi}\int_0^{\Omega} \frac{\sin\omega t}{\omega}\,\mathrm{d}\omega = \frac{1}{\pi}\int_{\omega=0}^{\Omega} \frac{\sin\omega t}{\omega t}\,\mathrm{d}(\omega t) = \frac{1}{\pi}\int_{\omega=0}^{\Omega} \operatorname{sinc}\omega t\,\mathrm{d}(\omega t) \\ &= \frac{1}{\pi}\int_0^{\Omega t} \operatorname{sinc}\tau\,\mathrm{d}\tau = \frac{1}{\pi}\operatorname{Si}\Omega t \end{aligned} \tag{8.201}$$

where

(integral sine function)

$$\operatorname{Si} t = \int_0^t \operatorname{sinc}\tau\,\mathrm{d}\tau \tag{8.202}$$

is the *integral sine* function. Fig. 8.15 illustrates.

Since $x_\Omega(t) = \frac{1}{\pi}\operatorname{Si}\Omega t$, it follows that, as Ω grows, the sine integral ripples in $x_\Omega(t)$ only become denser, but do not change their amplitude (Ω in this case working similarly to the frequency of a sinusoid $\sin\Omega t$). In this respect they behave pretty much like the ripples of the Gibbs phenomenon in the sawtooth signal shown in Fig. 8.7.

▷ In fact the Gibbs phenomenon ripples in the sawtooth are the sum of the ripples of infinitely many shifted integral sine functions, each integral sine occurring at a sawtooth's jump point.

Respectively, the inverse Fourier transform integral (8.199) converges only pointwise but not uniformly. One can see from Fig. 8.15 that

$$\lim_{\Omega\to+\infty} x_\Omega(t) = \lim_{\Omega\to+\infty} \frac{1}{\pi}\operatorname{Si}\Omega t = \frac{1}{\pi}\lim_{\tau\to+\infty}\operatorname{Si}(\tau\operatorname{sgn} t)$$

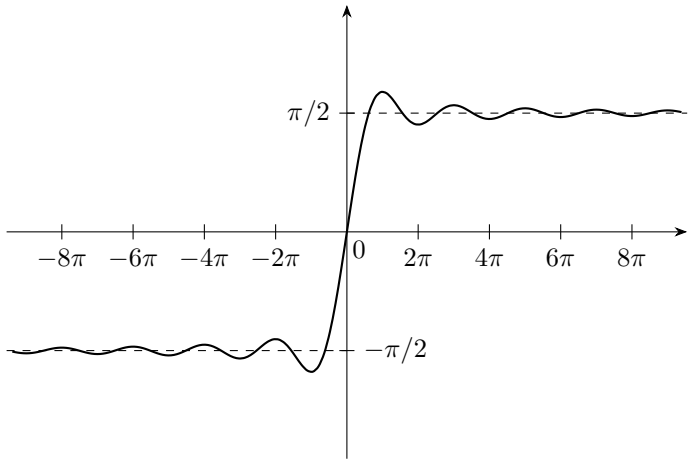


Figure 8.15: Integral sine function $\operatorname{Si} t$. The asymptotes are at $\pm\pi/2$ by (8.154). Note that, since $\operatorname{sinc} 0 = 1$, the graph actually goes through the origin at 45°. The angle is larger in the picture due to different scalings of the two axes.

$$= \frac{\operatorname{sgn} t}{\pi} \lim_{\tau\to+\infty} \operatorname{Si}(\tau) = \frac{1}{2}\operatorname{sgn} t \quad \forall t$$

(where "$\forall t$" means that the convergence is pointwise). This once again confirms (this time in a more detailed manner) that the inverse Fourier transform of $1/i\omega$ is $\frac{1}{2}\operatorname{sgn} t$.

For $\delta(t)$ the corresponding spectrum is simply the unit constant and the partial inverse Fourier transform integrals yield the functions $D_\Omega(t)$ given by (8.152b). We can see in Fig. 8.4 that the ripple amplitude grows with Ω.

Conversely, integrating $\frac{1}{2}\operatorname{sgn} t$ yields $|t|/2$, and the corresponding spectrum, by the integration property, should be $(1/i\omega)^2 = -1/\omega^2$. There should be even more severe Fourier transform convergence problems than with $\frac{1}{2}\operatorname{sgn} t$, as $|t|/2$ grows indefinitely at $t \to \infty$, but if these problems are somehow circumvented, we should also expect a decay of Gibbs ripple amplitude to zero as $\Omega \to \infty$.

▷ The integral (8.202) cannot be taken analytically, which means that $\operatorname{Si} t$ is not expressible in elementary functions, thus $\operatorname{Si} t$ is an "essentially new function". On the other hand, further repeated integration of $\operatorname{Si} t$ doesn't require introduction of further new functions. E.g.

$$\begin{aligned}\int_0^t \operatorname{Si}\tau\,\mathrm{d}\tau &= \tau\operatorname{Si}\tau\Big|_{\tau=0}^t - \int_{\tau=0}^t \tau\,\mathrm{d}(\operatorname{Si}\tau) = t\operatorname{Si} t - \int_0^t \tau\operatorname{sinc}\tau\,\mathrm{d}\tau \\ &= t\operatorname{Si} t - \int_0^t \sin\tau\,\mathrm{d}\tau = t\operatorname{Si} t + \cos\tau\Big|_{\tau=0}^t = t\operatorname{Si} t + \cos t - 1\end{aligned}$$

$$\begin{aligned}\int_0^t \tau\operatorname{Si}\tau\,\mathrm{d}\tau &= \int_{\tau=0}^t \operatorname{Si}\tau\cdot\mathrm{d}\left(\frac{\tau^2}{2}\right) = \frac{\tau^2}{2}\operatorname{Si}\tau\Big|_{\tau=0}^t - \int_{\tau=0}^t \frac{\tau^2}{2}\,\mathrm{d}(\operatorname{Si}\tau) \\ &= \frac{t^2}{2}\operatorname{Si} t - \int_0^t \frac{\tau^2}{2}\operatorname{sinc}\tau\,\mathrm{d}\tau = \frac{t^2}{2}\operatorname{Si} t - \int_0^t \frac{\tau}{2}\sin\tau\,\mathrm{d}\tau \\ &= \frac{t^2}{2}\operatorname{Si} t + \int_{\tau=0}^t \frac{\tau}{2}\mathrm{d}(\cos\tau) = \frac{t^2}{2}\operatorname{Si} t + \left(\frac{\tau}{2}\cos\tau\Big|_{\tau=0}^t - \frac{1}{2}\int_0^t \cos\tau\,\mathrm{d}\tau\right)\end{aligned}$$

$$= \frac{t^2}{2}\operatorname{Si} t + \frac{t}{2}\cos t - \frac{1}{2}\sin t$$

etc.

8.16.7 Compactly supported spectra

We have found that, if $x(t)$ is differentiable all the way down to the $(n-1)$-th derivative while having a jump in the n-th derivative, its spectrum generally rolls off as $1/\omega^{n+1}$. One can imagine then that the spectrum of an infinitely differentiable $x(t)$ will roll off as $1/\omega^\infty$.

This statement is not very precise, as $1/\omega^\infty$ rolloff is not a very precisely defined concept. However it probably can be understood as rolling off faster than $1/\omega^n$ $\forall n \in \mathbb{N}$. In Subsection 8.18.5 we will compute Fourier transform of the Gaussian distribution function e^{-t^2}, which is infinitely differentiable, and find out its Fourier transform to be just another Gaussian distribution function (with different parameters), which clearly rolls off faster than $1/\omega^n$ $\forall n \in \mathbb{N}$. However currently we'd like to analyze a simpler case of a "$1/\omega^\infty$" rolloff. Specifically we'd like to consider spectra which are simply equal to zero outside of a certain interval $[-\Omega, \Omega]$.

That is, let $X(\omega) = 0$ $\forall \omega \notin [-\Omega, \Omega]$. Taking the inverse Fourier transform (8.71b) yields

$$x(t) = \int_{-\infty}^{\infty} X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} = \int_{-\Omega}^{\Omega} X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi}$$

That is we are having an uncountable linear combination of complex sinusoids $e^{i\omega t}$ with the linear combination coefficients $X(\omega)\,\mathrm{d}\omega/2\pi$. Then the n-th derivative of $x(t)$ can be represented as the same linear combination of the n-th derivatives of the sinusoids $e^{i\omega t}$, the latter being equal to $(i\omega)^n e^{i\omega t}$:

$$x^{(n)}(t) = \int_{-\Omega}^{\Omega} (i\omega)^n X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} \tag{8.203}$$

So far this is nothing new, we just repeated the steps of deriving the differentiation property (8.186a), just this time integrating only over the interval $[-\Omega, \Omega]$, where $X(\omega)$ is nonzero. However this integration over the finite interval instead of integration over the entire $\mathbb{R}$ ensures the convergence of the integral (8.203).

In the general case, upon the application of the differentiation property (8.186a) we cannot be sure if $i\omega X(\omega)$ still decays sufficiently quickly at $\omega \to \infty$ to ensure the convergence of the inverse Fourier transform of $i\omega X(\omega)$. E.g. if $x(t)$ had a jump discontinuity and $X(\omega)$ decays as $1/\omega$, then Fourier transform of $x'(t)$ would "decay" as $1/\omega^0$, that is it won't decay, and the inverse Fourier transform would not be going to converge. By integrating over a finite range, we do not care anymore about the behavior of $X(\omega)$ at $\omega \to \infty$, this behavior not affecting the convergence of the inverse Fourier transform integral. This suggests that $x(t)$ with a *compactly supported spectrum* $X(\omega)$ (the latter being a fancy word to describe $X(\omega) = 0$ $\forall |\omega| > \Omega$) is infinitely differentiable and therefore all its derivatives are continuous.

Furthermore, we can notice that at large n the linear combination in (8.203) is "dominated" by the terms at $\omega \approx \Omega$:

$$|x^{(n)}(t)| \approx \left| \int_{|\omega|\approx\Omega} (i\omega)^n X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} \right| \approx \left| \int_{|\omega|\approx\Omega} \Omega^n X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} \right|$$
$$= \Omega^n \left| \int_{|\omega|\approx\Omega} X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} \right|$$

where $\int_{|\omega|\approx\Omega}$ is intuitive notation for integration over $[-\Omega, -\Omega+\varepsilon] \cup [\Omega-\varepsilon, \Omega]$ for some small $\varepsilon > 0$. It follows that the derivatives $x^{(n)}(t)$ at $n \to \infty$ are proportional to Ω^n.

There is a somewhat different angle at which we can look at the just derived property of compactly supported spectra. We can expand the elements $e^{i\omega t}$ of the linear combination in (8.203) into Taylor series:

$$x(t) = \int_{-\Omega}^{\Omega} X(\omega) e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} = \int_{-\Omega}^{\Omega} X(\omega) \left(\sum_{n=0}^{\infty} \frac{(i\omega t)^n}{n!} \right) \frac{\mathrm{d}\omega}{2\pi}$$

Taylor series for e^z is absolutely converging. Also, since $|z| = |i\omega t|$ is bounded by Ω, this convergence is uniform on $\omega \in [-\Omega, \Omega]$. Therefore we can integrate term-by-term:

$$x(t) = \sum_{n=0}^{\infty} \left(\int_{-\Omega}^{\Omega} (i\omega)^n X(\omega) \frac{\mathrm{d}\omega}{2\pi} \right) \frac{t^n}{n!} = \sum_{n=0}^{\infty} x^{(n)}(0) \frac{t^n}{n!} \tag{8.204}$$

where we can recognize that the integral in parentheses is exactly the one in (8.203) evaluated at $t = 0$, therefore it is equal to $x^{(n)}(0)$. Alternatively, we can simply notice that the latter expression in (8.204) is the Taylor series for $x(t)$. This suggests that $x(t)$ is an analytic function with derivatives decaying as Ω^n (if we formally allow t to take complex values).

An analytic function which is zero on some interval of a real line is zero everywhere (all derivatives at some point within this interval must be zero and so must be the Taylor expansion of the function at this point). Therefore this also suggests that a function with a compactly supported spectrum cannot be simultaneously compactly supported in its original domain (unless it's zero everywhere). In other words, the function cannot have compact support in both Fourier transform domain simultaneously.

▷ The latter statement is one manifestation of the "reciprocal duality" of the two domains of Fourier transform, the other manifestations being the domain scaling property, discussed later in Section 8.17 and the Heisenberg uncertainty relationship, discussed later in Subsection 8.18.4.

8.16.8 Spectral discontinuities

Until now we have been concentrating on discontinuities in the original t-domain function. Let's now take a look at the dual picture, where the discontinuities are present in the spectrum $X(\omega)$. While formally this dual picture should be

pretty much identical to the previously discussed situation, the switched domain semantics may result in a different point of view and we might be able to obtain new insights.

Let therefore $X(\omega)$ contain a (finite) number of discontinuities. We can expand $X(\omega)$ into a sum similar to (8.194)

$$X(\omega) = \bar{X}(\omega) + \sum_{m=1}^{M} \sum_{n=1}^{N_m} a_{mn} X_m(\omega - \omega_{mn}) \tag{8.205}$$

where $\bar{X}(\omega)$ doesn't have any discontinuities. As the "basis functions" $X_m(\omega)$ we can choose the exponentially bent step function (8.198) (Fig. 8.13) and its integrals, so that $X_1(\omega)$ is given by (8.198):

$$X_1(\omega) = e^{-|\omega|} \operatorname{sgn} \omega$$

$X_2(\omega)$ is given by integration of $X_1(\omega)$:

$$X_2(\omega) = -e^{-|\omega|}$$

etc.

We have found earlier that Fourier transform of (8.198) ($e^{-|t|} \operatorname{sgn} t$) is equal to $-2i\omega/(1+\omega^2)$. By duality, the inverse Fourier transform $x_1(t)$ of $X_1(\omega) = e^{-|\omega|} \operatorname{sgn} \omega$ must be (qualitatively, up to some constant coefficients)

$$x_1(t) = \frac{2it}{1+t^2}$$

By the frequency-shift property, the inverse Fourier transform of the term $X_1(\omega - \omega_{1n})$ of (8.205) must be (qualitatively)

$$x_{1n}(t) = x_1(t) e^{i\omega_{1n} t} = \frac{2it}{1+t^2} e^{i\omega_{1n} t}$$

Similarly, the inverse Fourier transform of the term $X_2(\omega - \omega_{2n})$ of (8.205) must be (qualitatively)

$$x_{2n}(t) = x_2(t) e^{i\omega_{2n} t} = \frac{2}{1+t^2} e^{i\omega_{2n} t}$$

etc. It follows that generally the time-domain counterpart of (8.205) is given by

$$x(t) = \bar{x}(t) + \sum_{m=1}^{M} \sum_{n=1}^{N_m} a_{mn} x_m(t) e^{i\omega_{mn} t} = \bar{x}(t) + \sum_{m=1}^{M} \left(x_m(t) \sum_{n=1}^{N_m} a_{mn} e^{i\omega_{mn} t} \right)$$

That is it contains a linear combination of $x_m(t)$ modulated by the complex sinusoids $e^{i\omega_{mn} t}$. This in turn suggests that $x(t)$ manifests sinusoidal elements of frequencies ω_{mn}, which are exactly the frequencies where discontinuities are present in $X(\omega)$. How prominently visible these sinusoidal elements are, apparently depends on the specific form of $\bar{x}(t)$ and $x_m(t)$, since other prominent features of $\bar{x}(t)$ and $x_m(t)$ might be masking these sinusoids. But generally speaking, these sinusoids must be there.

The simplest example of this effect is found by letting $X(\omega) = \operatorname{rect}(\omega)$. This $X(\omega)$ has two jump discontinuities at $\omega = \pm\frac{1}{2}$. By (8.168) we know that the

respective $x(t)$ is a sinc function, which is essentially a sinusoid modulated by $1/t$. We can obtain the exact form of $x(t)$ from (8.152b) by letting $\Omega = \frac{1}{2}$, yielding

$$x(t) = \frac{1}{2\pi} \operatorname{sinc} \frac{t}{2} = \frac{1}{\pi t} \sin \frac{t}{2}$$

As expected, $x(t)$ has a sinusoidal component of frequency $\omega = \frac{1}{2}$ (corresponding to two complex sinusoidal components at $\omega = \pm\frac{1}{2}$). We may also recall from our previous discussion that a jump discontinuity in one domain should cause a $1/t$ decay rate in the other domain, and indeed $x(t)$ decays as $1/t$.

Another example, where spectral discontinuities result in t-domain sinusoids, is the previously discussed Gibbs phenomenon. Remember that, with Fourier transform, Gibbs phenomenon was appearing in partial inverse Fourier transform integrals, such as (8.201). This partial integral can be alternatively seen as two discontinuities in the spectrum occurring at $\omega = \pm\Omega$:

$$x_\Omega(t) = \int_{-\Omega}^{\Omega} \frac{e^{i\omega t}}{i\omega} \frac{\mathrm{d}\omega}{2\pi} = \int_{-\infty}^{\infty} \frac{\operatorname{rect}(\omega/2\Omega)}{i\omega} e^{i\omega t} \frac{\mathrm{d}\omega}{2\pi} = \frac{1}{\pi} \operatorname{Si} \Omega t$$

In Fig. 8.15 (page 477) one can see that the ripples in the $\operatorname{Si} t$ function appear approximately with the period of 2π, which is the period of the sinusoid $\sin t$, which has frequency $\omega = 1$. Even though $\operatorname{Si} t$ itself doesn't decay to zero, the ripples in $\operatorname{Si} t$ are, in fact, also decaying as $1/t$.

Thus the primary cause for the Gibbs phenomenon are not the discontinuities in the t domain, but the jump discontinuity in the spectrum occurring due to the computation of partial inverse Fourier transform integrals. In a sense, the t-domain discontinuities are just "making the Gibbs phenomenon's sinusoidal ripples more prominent".

8.16.9 Dual-domain effects of features

We can summarize the effects of discontinuities on the dual domain as follows:

- A discontinuity in the n-th derivative in the t or ω domain generally bounds the function's rolloff speed in the other domain to be not faster than $1/\omega^{n+1}$ or $1/t^{n+1}$ respectively.

- A discontinuity (in the function or in one of its derivatives) at $t = t_0$ or $\omega = \omega_0$ in one domain generally contributes sinusoidal ripples of frequency t_0 or ω_0 in the other domain.

Thus discontinuities in one domain cause sinusoidal ripples and potentially slower decay rate in the other domain. If the discontinuities are "smoothed" (instead of a discontinuous jump in the function's value or in one of its derivatives there is a quick but continuous transition), the function or its derivative is still "approximately discontinuous", therefore the usual dual domain manifestations of a discontinuity will occur in an "approximate form". E.g. we could consider the integral sine function $\operatorname{Si} \Omega t$ as a smoothed version of the discontinuous Heaviside step function (8.197). By (8.201), Fourier transform of $x(t) = \frac{1}{\pi} \operatorname{Si} \Omega t$ is $X(\omega) = \frac{1}{i\omega} \operatorname{rect}(\omega/2\Omega)$. That is Fourier transform of $\operatorname{Si} \Omega t$ still decays as $1/i\omega$, as it would in case of a true discontinuity, but does so only for a while, for

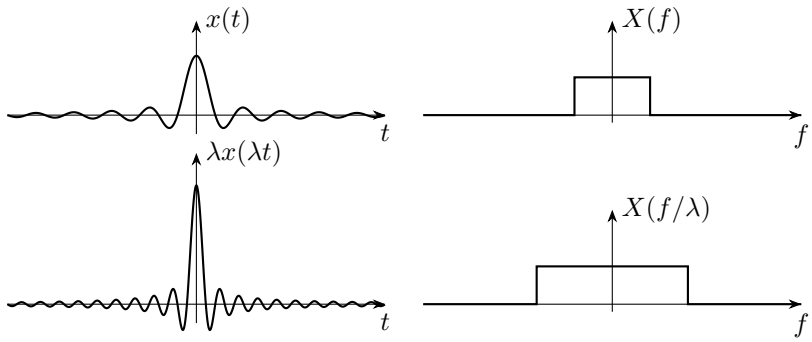


Figure 8.16: The effect of f-domain stretching on the t domain.

$|\omega| < \Omega$. Similarly a time-shifted integral sine $x(t) = \operatorname{Si}\Omega(t - t_0)$, by the time-shift property will have the spectrum $X(\omega) = \frac{1}{i\omega}\operatorname{rect}(\omega/2\Omega)e^{-i\omega t_0}$, thereby the function $X(\omega)$ obtaining a sinusoidal component of frequency t_0.

Thus, if the function $x(t)$ has a "feature" (something which at least approximately looks like a discontinuity in the function or one of its derivatives) at $t = t_0$, its spectrum $X(\omega)$ is likely to attain a sinusoidal component of frequency t_0, as well as decay as $1/\omega^{n+1}$ for a while (where n is the lowest order of an approximately discontinuous derivative). The range of this $1/\omega^{n+1}$ decay will be approximately inversely proportional to the "width of the discontinuity".

8.17 Domain scaling

In the derivation of nonperiodic Dirac delta function we have been dealing with functions $D_F(2\pi t)$ given by (8.152a) with their respective Fourier transform $\operatorname{rect}(f/2F)$ given by (8.168):

$$x(t) = 2F \operatorname{sinc} 2\pi F t \iff X(f) = \operatorname{rect}(f/2F)$$

or, replacing $2F$ by F

$$x(t) = F \operatorname{sinc} \pi F t \iff X(f) = \operatorname{rect}(f/F) \tag{8.206}$$

Suppose we double the value of F in (8.206). Thereby the function $X(f)$ doubles its horizontal length, while the "horizontal length" of $x(t)$ halves, but simultaneously the amplitude of $x(t)$ doubles. The doubling of the amplitude of $x(t)$ makes sure that the area under $x(t)$ stays constant. This area is supposed to be equal to $X(0)$ and, since $X(0)$ doesn't change by such transformation, neither should the area under $x(t)$.

Generally, for $x(t)$ and $X(f)$ given by (8.206),

$$X'(f) = X(f/\lambda) \iff x'(t) = \lambda x(\lambda t) \qquad \forall \lambda > 0 \tag{8.207}$$

where x' and X' denote the functions after the transformation (rather than differentiation). Fig. 8.16 illustrates. Let's find out, how generic is this effect.

Goal 8.26. Find out if (8.207) holds generally.

In (8.207) the t and f variables are scaled in a mutually reciprocal fashion. Apparently, this must be due to the way how t and f participate in the exponential $e^{2\pi ift}$. Given a λ times larger t, we need a λ times smaller f to obtain the same value of $e^{2\pi ift}$. We also should recall that in Fourier transform (and in integrals in general) the functions are having density semantics (compared to the total value of the integral). Therefore, when changing the domain's scale by the factor λ, we also need to change the function's value by the same factor. Let's try to express the same more formally.

By (8.70b), given $X'(f) = X(f/\lambda)$ (that is stretching by the factor $\lambda > 0$ in the f-domain) we obtain

$$x'(t) = \int_{-\infty}^{\infty} X'(f)e^{2\pi ift}\,\mathrm{d}f = \int_{-\infty}^{\infty} X(f/\lambda)e^{2\pi ift}\,\mathrm{d}f$$

Letting $\nu = f/\lambda$ as the new dual-domain variable we continue

$$x'(t) = \int_{-\infty}^{\infty} X(\nu)e^{2\pi i\lambda\nu t}\,\mathrm{d}(\lambda\nu) = \lambda\int_{-\infty}^{\infty} X(\nu)e^{2\pi i\lambda\nu t}\,\mathrm{d}\nu$$

Formally replacing ν with f

$$x'(t) = \lambda\int_{-\infty}^{\infty} X(f)e^{2\pi if\cdot\lambda t}\,\mathrm{d}f = \lambda x(\lambda t)$$

that is, we have obtained (8.207) as a general property of Fourier transform.

There is one small extension which we can make to (8.207). In the transformations, that we did, we assumed $\lambda > 0$, because for $\lambda < 0$ the integration limits will swap upon switching from f to ν and the integral will change sign. This sign change can be accounted for by simply rewriting $\lambda x(\lambda t)$ as $|\lambda| \cdot x(\lambda t)$ and we obtain

(Fourier transform f-domain scaling property, a.k.a frequency-scaling property)

$$X'(f) = X(f/\lambda) \iff x'(t) = |\lambda| \cdot x(\lambda t) \tag{8.208a}$$

(alternatively, to handle negative λ we can simply apply the property (8.54a), yielding the same result (8.208a)). By duality, we apparently get the same property for stretching in the other domain:

(Fourier transform t-domain scaling property, a.k.a. time-scaling property)

$$x'(t) = x(t/\lambda) \iff X'(f) = |\lambda| \cdot X(\lambda f) \tag{8.208b}$$

Thus, horizontal stretching in one domain by a factor λ results in simultaneous horizontal contraction by the same factor λ and vertical stretching by $|\lambda|$ in the other domain.

Goal 8.26 completed. ■

8.18 Power and energy

8.18.1 Power and energy in two domains

Recall the Parseval's theorem for the Fourier series, specifically its particular case given by (8.90b), which we for convenience write here once again:

$$\|x\|^2 = \frac{1}{T}\int_{-T/2}^{T/2} \left|x(t)\right|^2 \mathrm{d}t = \sum_{k=-\infty}^{\infty} \left|X_k\right|^2 = \|X\|^2 \tag{8.209}$$

In the time/frequency interpretation this identity has a special meaning. Suppose $x(t)$ denotes a voltage signal connected to a unit load. Then $|x(t)|^2$ is the instantaneous power which is being dissipated at this unit load at the time moment t. Respectively the left-hand side integral in (8.209) denotes the average power dissipated over one period of the signal $x(t)$.

Now consider a single complex partial $x_k(t) = X_k e^{2\pi ikFt}$ of the Fourier series. The power dissipated by this partial is

$$|x_k(t)|^2 = \left|X_k e^{2\pi ikFt}\right|^2 = \left|X_k\right|^2 \cdot \left|e^{2\pi ikFt}\right|^2 = \left|X_k\right|^2$$

The latter statement is a bit awkward, since $x_k(t)$ is a complex-valued signal, and it is not clear whether complex-valued voltage (being just a speculative mathematical construct) can dissipate any real power. We can however add two conjugate partials $x_k(t)$ and $x_{-k}(t)$ together and consider the power of their real sum, which is the respective real partial:

$$\begin{aligned}
\left|x_k(t) + x_{-k}(t)\right|^2 &= \left|X_k e^{2\pi ikFt} + X_{-k} e^{-2\pi ikFt}\right|^2 \\
&= \left|X_k e^{2\pi ikFt} + (X_k e^{2\pi ikFt})^*\right|^2 \\
&= \left|2\,\mathrm{Re}\left(X_k e^{2\pi ikFt}\right)\right|^2 \\
&= \Big(2 \cdot |X_k| \cdot \cos(2\pi kFt + \arg X_k)\Big)^2 \\
&= 4 \cdot |X_k|^2 \cdot \cos^2(2\pi kFt + \arg X_k)
\end{aligned}$$

Now the average of a squared cosine over one period is $1/2$, therefore the average of $|x_k(t) + x_{-k}(t)|^2$ over one period is equal to $2 \cdot |X_k|^2$. It is as if $x_k(t)$ and $x_{-k}(t)$ contributed the amounts $|X_k|^2$ and $|X_{-k}|^2$ respectively to the average power. In that sense we can consider each complex partial $x_k(t)$ contributing $|X_k|^2$ to the total average power and this suggests that we can interpret the sum in the right-hand side of (8.209) as the total average power contributed by all individual partials $x_k(t)$ of the Fourier series together.

From the perspective of Fourier series and signal processing specifically, the "power dissipated on a unit load" given by $|x(t)|^2$ is considered simply as the "power of the signal". This interpretation is strengthened by the fact that usually power emerges as "something squared", thus the signal doesn't necessarily need to have voltage semantics. The equation (8.209) thereby simply reads:

(power-based interpretation of Parseval's theorem for periodic signals)

"the average power of a signal is equal to the sum of average powers contributed by the individual partials".

▷ In the above statement it is essential that we are dealing with sinusoidal partials of different frequencies. These partials are thereby uncorrelated and their power contributions simply add. This would have been different with (positively or negatively) correlated partials.

In case of (continuous) Fourier transform (8.90b) is replaced by (8.91b), which we also rewrite once again here for convenience:

$$\|x\|^2 = \int_{-\infty}^{\infty} \left|x(t)\right|^2 \mathrm{d}t = \int_{-\infty}^{\infty} \left|X(f)\right|^2 \mathrm{d}f = \|X\|^2 \tag{8.210}$$

Here the signals involved are not periodic and we don't have the average power interpretation. Instead we can notice that the left-hand side of (8.210) contains an integral of the instantaneous power over the entire time range from $-\infty$ to ∞ and thereby represents the entire energy of the signal. The right-hand side of (8.210) therefore contains individual contributions to that energy occurring at different frequencies f. However, since continuous Fourier transform has "density" semantics, the actual contributions to the sum are given by $|X(f)|^2\,\mathrm{d}f$, therefore $|X(f)|^2$ is the (spectral) energy density function, commonly referred to as *energy spectral density* or *ESD*. Equation (8.210) thereby reads:

(energy-based interpretation of Parseval's theorem for nonperiodic signals)

"the total energy of a signal is the same in time and frequency domains".

In case of discrete-time types of Fourier transform, similar interpretations are used. $|x[n]|^2$ is considered to be the instantaneous power and the sum of $|x[n]|^2$ over an interval of n is considered to be the energy on that interval.

▷ A square of the absolute magnitude of a quantum-mechanical wavefunction, instead of energy density (including power, which is "energy density with respect to time") has the probability density semantics. Respectively, we are having probability densities in different domains, while the integrals over $(-\infty, \infty)$ in (8.210) yield the total probability (which must be equal to 1 regardless of the domain, this is the usual normalization condition in quantum mechanics).

8.18.2 Energy center

The theory of relativity teaches us that energy is the same as mass ($E = mc^2$). We could therefore be interested in the "energy center" of a function, in the same way how one could be interested in the mass center of a body. The latter, as we

know, is given by the weighted average of positions of individual contributions $\rho(\vec{x})\,\mathrm{d}V$ to the total mass, the weights being given by those mass contributions:

$$\vec{x}_\mathrm{c} = \frac{\iiint \vec{x}\cdot\rho(\vec{x})\,\mathrm{d}V}{\iiint \rho(\vec{x})\,\mathrm{d}V}$$

where $\rho(\vec{x})$ is the mass density at position $\vec{x}$. It is convenient to normalize the mass density, so that the total mass of the body is unity:

$$\bar{\rho}(\vec{x}) = \frac{\rho(\vec{x})}{\iiint \rho(\vec{x})\,\mathrm{d}V}$$

in which case we obtain a simpler formula for the mass center:

$$\vec{x}_\mathrm{c} = \iiint \vec{x}\cdot\bar{\rho}(\vec{x})\,\mathrm{d}V$$

In one dimension the same formula turns into

$$x_\mathrm{c} = \int_{-\infty}^{\infty} x\cdot\bar{\rho}(x)\,\mathrm{d}x$$

Given a function $x(t)$ and treating the expression $|x(t)|^2$ as power (which is "one-dimensional" energy density with respect to time), we formally let $\rho(t) = |x(t)|^2$ be our energy density function. Without loss of generality we can assume that $\rho(t)$ is normalized:

$$\int_{-\infty}^{\infty} \rho(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} |x(t)|^2\,\mathrm{d}t = 1$$

or, in other words, that the total energy of $x(t)$ is unity (if it's not, we can always divide $x(t)$ by the square root of the total energy). Then the "energy center" of the signal $x(t)$ (with respect to time) is

(time-domain energy center of a signal)

$$M_x = \int_{-\infty}^{\infty} t\cdot|x(t)|^2\,\mathrm{d}t \qquad \text{(for normalized } |x(t)|^2\text{)} \tag{8.211}$$

It can be instructive to obtain the frequency-domain expression for M_x.

Goal 8.27. Obtain the frequency-domain formula for the time-domain energy center M_x. *Hint:* use Parseval's theorem and transform its frequency-domain part.

We cannot use Parseval's theorem length preservation formula (8.92b) directly, since there is an extra factor t under the integral. But we can rewrite the expression $t\cdot|x|^2$ as $t\cdot|x|^2 = t\cdot x^*x = x^*\cdot tx = x^*y$ where $y(t) = tx(t)$:

$$M_x = \int_{-\infty}^{\infty} x^*\cdot tx\,\mathrm{d}t = \int_{-\infty}^{\infty} x^*y\,\mathrm{d}t = \langle x, y\rangle$$

to which we can apply the full version of Parseval's theorem (8.92a):

$$M_x = \int_{-\infty}^{\infty} x^* y \,\mathrm{d}t = \int_{-\infty}^{\infty} X^*(f) Y(f) \,\mathrm{d}f$$

or, in terms of ω

$$M_x = \int_{-\infty}^{\infty} X^*(\omega) Y(\omega) \frac{\mathrm{d}\omega}{2\pi}$$

Applying the differentiation property (8.187a) to the function $y(t) = tx(t)$ yields $Y(\omega) = iX'(\omega)$ (where the prime denotes differentiation) and we obtain

$$M_x = i \int_{-\infty}^{\infty} X^*(\omega) X'(\omega) \frac{\mathrm{d}\omega}{2\pi} \tag{8.212}$$

Now (8.212) gives the same value as (8.211). But the integral expression in (8.211) clearly yields a purely real result (even for a complex $x(t)$). Therefore so must do the integral expression in (8.212). This is only possible if the integral of X^*X' is a purely imaginary value. Let's take a closer look at the real and imaginary parts of the product X^*X'.

Suppose $X = X^* = 1$. Then the real component of X' describes the rate of change of the absolute magnitude of X and the imaginary component of X' describes the rate of change of the complex argument of X:

$$\mathrm{Re}(X') = |X|' \qquad \mathrm{Im}(X') = (\arg X)' \qquad \text{for } X = 1$$

Since $X^*X' = 1 \cdot X' = X'$, the same is true for the product X^*X':

$$\mathrm{Re}(X^*X') = |X|' \qquad \mathrm{Im}(X^*X') = (\arg X)' \qquad \text{for } X = 1$$

Now if we consider a more general case of $|X| = 1$, we can realize that multiplication by X^* performs a rotation of the complex plane, aligning X back with the positive real semiaxis, and we effectively are having the same result as for $X = 1$:

$$\mathrm{Re}(X^*X') = |X|' \qquad \mathrm{Im}(X^*X') = (\arg X)' \qquad \text{for } |X| = 1$$

Finally, letting $|X|$ take completely arbitrary values, we need to multiply the previous results by $|X|$ and $|X|^2$ respectively:

$$\mathrm{Re}(X^*X') = |X| \cdot |X|' \qquad \mathrm{Im}(X^*X') = |X|^2 \cdot (\arg X)' \qquad \forall X \in \mathbb{C}$$

▷ Note that $\arg X$ here, apparently, doesn't denote the complex argument in the range $(-\pi, \pi]$. Instead $\arg X$ can vary in the entire range $(-\infty, \infty)$. The function $\varphi(\omega) = \arg X(\omega)$ is not supposed to contain any "artificial discontinuities" at the points where $\arg X = \pm\pi$, $\pm 3\pi$, $\pm 5\pi$ etc., caused by wrapping to the range $(-\pi, \pi]$. This ensures that the derivative of $\varphi(\omega)$ exists also at such points. Sometimes such unconstrained $\arg X$ is referred to as *continuous phase* or *unwrapped phase*.

The same result can be also obtained formally. A common way to separate the absolute magnitude from the complex argument is to use the complex logarithm: $\log X = |X| + i \arg X$. Let's therefore attempt to rewrite X^*X' using $\log X$, specifically rewriting the X' factor in terms of $\log X$, since we're interested in the derivative of the absolute magnitude and complex argument:

$$X^*X' = X^*(e^{\log X})' = X^* \cdot e^{\log X} \cdot (\log X)' = X^*X \cdot (\ln |X| + i \arg X)'$$

$$= |X|^2 \cdot \left(\frac{|X|'}{|X|} + i \frac{\mathrm{d}}{\mathrm{d}\omega} \arg X \right) = |X| \cdot |X|' + |X|^2 \cdot i \frac{\mathrm{d}}{\mathrm{d}\omega} \arg X \tag{8.213}$$

▷ Similarly to $\arg X$, the function $\log X$ denotes here the *continuous* or *unwrapped* complex logarithm. The imaginary part of this logarithm is not constrained to $(-\pi, \pi]$.

Thus, by (8.213), we obtained the same result

$$\begin{aligned} \operatorname{Re} X^* X' &= |X| \cdot |X|' \\ \operatorname{Im} X^* X' &= |X|^2 \cdot \frac{\mathrm{d}}{\mathrm{d}\omega} \arg X \end{aligned}$$

Integrating the real part gives

$$\begin{aligned} &\int_{-\infty}^{\infty} \operatorname{Re}(X^* X') \, \mathrm{d}\omega = \int_{-\infty}^{\infty} |X| \cdot |X|' \cdot \mathrm{d}\omega = \int_{\omega=-\infty}^{\infty} |X| \cdot \mathrm{d}(|X|) \\ &= \int_{\omega=-\infty}^{\infty} \mathrm{d}\left(\frac{|X|^2}{2} \right) = \frac{|X|^2}{2} \bigg|_{\omega=-\infty}^{\infty} = \frac{|X(+\infty)| - |X(-\infty)|}{2} = \frac{0-0}{2} = 0 \end{aligned}$$

where $X(\pm\infty) = 0$ generally in order for the inverse Fourier transform integral to converge. Even if we need to deal with more exotic cases where $X(\pm\infty) \neq 0$, we probably are going at least to have $|X(+\infty)| = |X(-\infty)|$ anyway. So the integral of $\operatorname{Re}(X^* X')$ is zero indeed and we are left just with the integral of the purely imaginary component:

$$\begin{aligned} M_x &= i \int_{-\infty}^{\infty} X^* X' \frac{\mathrm{d}\omega}{2\pi} = i \int_{-\infty}^{\infty} i \operatorname{Im}(X^* X') \frac{\mathrm{d}\omega}{2\pi} = - \int_{-\infty}^{\infty} \operatorname{Im}(X^* X') \frac{\mathrm{d}\omega}{2\pi} \\ &= - \int_{-\infty}^{\infty} |X|^2 \cdot \left(\frac{\mathrm{d}}{\mathrm{d}\omega} \arg X \right) \frac{\mathrm{d}\omega}{2\pi} \end{aligned}$$

and thus (8.212) can be rewritten as

(frequency-domain expression for the time-domain energy center of a signal)

$$M_x = \int_{-\infty}^{\infty} |X|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg X \right) \frac{\mathrm{d}\omega}{2\pi} \tag{8.214}$$

That is M_x is given by the integral of $-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg X$ weighted by $|X|^2$. Notice that if $|x(t)|^2$ is normalized (the integral of $|x(t)|^2$ is unity, then, by Parseval's theorem, so is $|X(\omega)|^2$. That is $|X|^2$ is a natural normalized weight value. It looks like the (negated) phase derivative is a key factor defining the value of M_x, the relative contributions of this factor at different frequencies being weighted by "how much of the signal $x(t)$" is present at the respective frequency ω.

▷ Readers, who studied signal processing, should recognize $-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg X$ as the "group delay". Except that, in our current discussion, $X(\omega)$ is not a filter, thus we cannot really talk of this value as the group delay.

▷ In quantum mechanics instead of power (energy density with respect to time) we have probability density with respect to position. Thus (8.211) turns into an expectation value of a particle's position. At the same time the dual domain has momentum semantics, so (8.214) integrates the phase derivative with respect to the momentum.

In wave mechanics, besides the time domain we may be using the space position domain with the wavenumber domain as its dual. Respectively (8.211) gives the energy center in space, while (8.214) expresses this energy center via the phase derivative with respect to the wavenumber.

Goal 8.27 completed. ■

8.18.3 Energy concentration

In Subsection 8.18.2 we have introduced the expression for the energy center (8.211). In a similar fashion, treating $|x|^2$ as a normalized weight function, we can compute the "standard deviation" of the energy:

$$\sigma_x = \sqrt{\int_{-\infty}^{\infty} (t - M_x)^2 \cdot |x(t)|^2 \cdot \mathrm{d}t} \tag{8.215a}$$

which can be used as a measure of how compactly the energy is concentrated around $t = M_x$.

We can rewrite (8.215a) as

$$\begin{aligned}\sigma_x^2 &= \int_{-\infty}^{\infty} (t - M_x)^2 \cdot |x(t)|^2 \cdot \mathrm{d}t = \int_{-\infty}^{\infty} (t^2 - 2M_x t + M_x x^2) \cdot |x(t)|^2 \cdot \mathrm{d}t \\ &= \int_{-\infty}^{\infty} t^2 \cdot |x(t)|^2 \cdot \mathrm{d}t - 2M_x \cdot M_x + M_x x^2 = \int_{-\infty}^{\infty} t^2 \cdot |x(t)|^2 \cdot \mathrm{d}t - M_x^2 \\ &= \int_{-\infty}^{\infty} |tx(t)|^2 \cdot \mathrm{d}t - M_x^2 = \|tx(t)\| - M_x^2 \end{aligned} \tag{8.215b}$$

(where $\|tx(t)\|$ is a somewhat loose notation for $\|y\|$ where $y(t) = tx(t)$).

By the differentiation property (8.189a) of the Fourier transform, the Fourier transform of $tx(t)$ is $-X'(f)/2\pi i$, since $X'(f)$ is Fourier transform of $-2\pi i t x(t)$. By Parseval's theorem, (8.215b) can be rewritten as

$$\sigma_x^2 = \int_{-\infty}^{\infty} \left| \frac{X'(f)}{2\pi} \right|^2 \cdot \mathrm{d}f - M_x^2 = \left\| \frac{X'(f)}{2\pi} \right\|^2 - M_x^2 \tag{8.215c}$$

In case $M_x = 0$, (8.215b) and (8.215c) turn into

(time-domain "standard deviation" of energy of a centered signal)

$$\sigma_x = \|tx(t)\| = \left\| \frac{X'(f)}{2\pi} \right\| \tag{8.216}$$

That is σ_x^2 is equal to the total energy of the signal $tx(t)$, or, equally, to the total energy of the spectrum $X'(f)/2\pi$.

▷ If $M_x \neq 0$, (8.216) will give the "width of the function $x(t)$" measured with respect to $t = 0$ instead of measured with respect to the energy center.

8.18.4 Energy concentration bound

In Section 8.17 we have observed that the concentration of a sinc function around its energy center $t = 0$ in one domain is inversely proportional to the concentration of its Fourier transform (given by the rectangular function) around its center $f = 0$ in the other domain, since a horizontal stretch by λ in one domain corresponds to a horizontal shrinking by the same factor in the other domain. Obviously, the "standard deviation" (8.215) grows proportionally to such stretching,[9] therefore there is a relationship

$$\sigma_x \sigma_X = \text{const} \tag{8.217}$$

between the respective "standard deviations" in the two domains.

▷ Apparently, until now we have introduced the "standard deviation" only in the t domain, but nothing is stopping us from formally introducing the "energy center" and the "standard deviation" in the f domain as well:

$$M_X = \int_{-\infty}^{\infty} f \cdot |X(f)|^2 \, \mathrm{d}f \tag{8.218a}$$

$$\begin{aligned} \sigma_X^2 &= \int_{-\infty}^{\infty} (f - M_X)^2 \cdot |X(f)|^2 \cdot \mathrm{d}f = \|fX(f)\| - M_X^2 \\ &\overset{(8.188a)}{=} \int_{-\infty}^{\infty} \left| \frac{x'(t)}{2\pi} \right|^2 \cdot \mathrm{d}t - M_X^2 = \left\| \frac{x'(t)}{2\pi} \right\|^2 - M_X^2 \end{aligned} \tag{8.218b}$$

where we assume normalized $|X(f)|^2$, which, by Parseval's theorem, will be ensured if $|x(t)|^2$ is normalized.

We wish to know how generic is the relationship (8.217). Is there a universal constant which is one and the same no matter what the function $x(t)$, or different kinds of functions correspond to different values of that constant?

Goal 8.28. Try to transform (8.217) into a universal constraint holding for arbitrary $x(t)$. *Hint:* the result will be an inequality rather than equality. To obtain it use Parseval's theorem and Cauchy–Schwarz inequality.

As the first step we shall notice that horizontal shifts in either domain do not affect the values σ_x and σ_X. E.g. a shift in the t domain obviously doesn't affect σ_x. By the time-shift property of Fourier transform, a t-domain shift by an amount d corresponds to multiplication by $e^{-2\pi i f d}$ in the f domain. Since $|e^{-2\pi i f d}| = 1$, this multiplication doesn't affect the value of $|X(f)|$ and respectively doesn't affect the values of M_X and σ_X. By a similar argument we find that a shift in the f domain doesn't affect the values of σ_X, M_x and σ_x. Therefore, without loss of generality, we can assume $M_x = 0$ and $M_X = 0$, as we can always shift the functions to ensure $M_x = M_X = 0$ without affecting σ_x or σ_X.

[9] We need to remember to make sure that $|x(t)|^2$ is renormalized after the stretching if we want to compute M_x and σ_x.

Therefore, let

$$\sigma_x = \|tx(t)\| = \left\| \frac{X'(f)}{2\pi} \right\| \tag{8.219a}$$

$$\sigma_X = \|fX(f)\| = \left\| \frac{x'(t)}{2\pi} \right\| \tag{8.219b}$$

We are interested in studying the behavior of the product $\sigma_x \sigma_X$.

The expressions (8.219) do not look like their product would be a constant value, even under the restriction of a unit energy of $x(t)$. Indeed, suppose we have some signal $x(t)$ with respective values of σ_x and σ_X given by (8.219). Then we could probably invent some modification of $x(t)$ which doesn't change the average values of $|x(t)|^2$ and $|tx(t)|^2$ too much, but noticeably changes the average of $|x'(t)|^2$. E.g. suppose $x(t)$ is slowly varying and we multiply it by some quickly oscillating sinusoid:

$$\tilde{x}(t) = x(t) \cos at$$

where a is sufficiently large in relation to the rate of change of $x(t)$). Then $|\tilde{x}|^2 = |x|^2 \cdot \cos^2 at$, and since the average value of $\cos^2 at$ is $\frac{1}{2}$, the total energy of $\tilde{x}(t)$ will be approximately half of the total energy of $x(t)$ (Fig. 8.17). We don't want the total energy to noticeably change, so let's multiply by $\sqrt{2} \cdot \cos at$ instead:

$$\tilde{x}(t) = x(t) \cdot \sqrt{2} \cos at$$

Now the total energy of $\tilde{x}(t)$ is approximately equal to the total energy of $x(t)$, and the total energy of $t\tilde{x}(t)$ is is approximately equal to the total energy of $tx(t)$. That is $\sigma_{\tilde{x}} \approx \sigma_x$. On the other hand, the value of $x'(t)$ is noticeably increased in magnitude:

$$\tilde{x}'(t) = (x \cdot \sqrt{2} \cos at)' = x' \cdot \sqrt{2} \cos at + \sqrt{2} \cdot ax \sin at$$

where we can see that the total energy of $x' \cdot \sqrt{2} \cos at$ is approximately equal to the total energy of $x'(t)$, but the total energy of the term $\sqrt{2} \cdot ax \sin at$ can be made arbitrarily large by choosing a sufficiently large a. So even if there are any cancellations between the two terms (if the terms are having opposite signs), at sufficiently large a these cancellations won't even matter. Thus at sufficiently large a we are going to have $\sigma_{\tilde{X}} \gg \sigma_X$.

▷ From the f-domain perspective, multiplication by $\cos at$ is going to significantly "widen" the function $X(f)$. Indeed, $\cos at$ can be expanded into a sum of two complex sinusoids:

$$\cos at = \frac{1}{2} e^{at} + \frac{1}{2} e^{-at}$$

Now recall the frequency shift property (8.113). By this property, multiplication with a complex sinusoid in time domain corresponds to a horizontal shift (translation) in frequency domain. So multiplication by each of the terms e^{2at} and e^{-2at} is going to cause two frequency shifts in opposite directions, the amounts of shifts being proportional to a. If a is large compared to the original width of $X(f)$, the function $X(f)$ will be thereby "widened".

Since $\sigma_x \approx \sigma_{\tilde{x}}$ and $\sigma_{\tilde{X}} \gg \sigma_X$ in our example, it follows that $\sigma_{\tilde{x}} \sigma_{\tilde{X}} \gg \sigma_x \sigma_X$. Respectively, $\sigma_x \sigma_X$ cannot be invariant under arbitrary choice of $x(t)$. Thus,

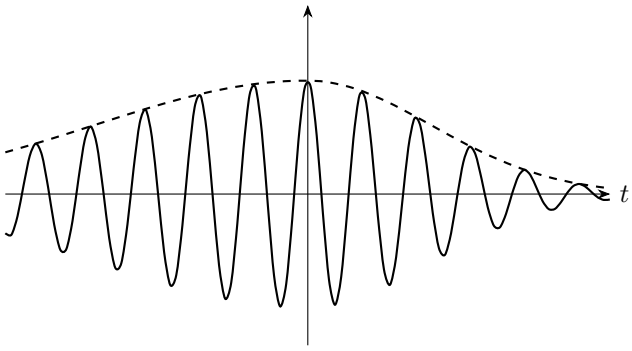


Figure 8.17: Multiplication by $\cos at$ of a sufficiently high frequency a approximately halves the energy of the signal. The dashed line represents the signal prior to multiplication.

the relationship (8.217) cannot hold for arbitrary changes of $x(t)$, but only holds "per family of $x(t)$" (where such family e.g. could consist of functions of identical shapes, varying only in horizontal and vertical dimensions).

So it looks like our original conjecture about $\sigma_x\sigma_X$ always evaluating to one and the same constant is wrong. But we still might be onto something with our expressions (8.219). In our previous experiment we managed to increase the value of $\sigma_x\sigma_X$ by a pretty much arbitrarily large amount. However, it's not that obvious how this value could be decreased for an arbitrary given signal. This might lead to the idea that, despite not being constant, the expression $\sigma_x\sigma_X$ might have some kind of lower boundary.

From our discussion of Fourier transform unitarity and Parseval's theorem we should recall that signal energies can be seen as squared "vector lengths", which is also explicitly suggested by the notation in (8.219). Thereby the product $\sigma_x\sigma_X$ can be interpreted as a product of vector lengths:

$$\sigma_x\sigma_X = \|tx(t)\| \cdot \|fX(f)\|$$

or, writing both vectors in terms of "the same coordinate system":

$$\sigma_x\sigma_X = \|tx(t)\| \cdot \left\| \frac{x'(t)}{2\pi} \right\|$$

This product is bounded from below by the "dot product" of the respective vectors, since the dot product cannot exceed the product of the lengths of the vectors ("the absolute magnitude of the cosine of the angle between the vectors cannot exceed 1"). It is the statement of the *Cauchy–Schwarz inequality*:

$$\sigma_x\sigma_X = \|tx(t)\| \cdot \left\| \frac{x'(t)}{2\pi} \right\| \geq \left| \left\langle tx(t), \frac{x'(t)}{2\pi} \right\rangle \right| \tag{8.220}$$

where $\langle \cdot, \cdot \rangle$ is defined by (8.91a). Now let's take a closer look at the "dot product" in the right-hand side of (8.220) and see if we could get any estimate for it.

By (8.91a)

$$\left\langle tx(t), \frac{x'(t)}{2\pi} \right\rangle = \int_{-\infty}^{\infty} tx^*(t) \frac{x'(t)}{2\pi} \, \mathrm{d}t \tag{8.221}$$

By (8.213) (which holds for any complex function of a real argument)

$$\begin{aligned} \left\langle tx(t), \frac{x'(t)}{2\pi} \right\rangle &= \frac{1}{2\pi} \int_{-\infty}^{\infty} tx^*(t) x'(t) \, \mathrm{d}t \\ &\overset{(8.213)}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} t \cdot \left(|x| \cdot |x|' + i \cdot |x|^2 \cdot (\arg x)' \right) \mathrm{d}t \end{aligned} \tag{8.222}$$

Now if $x(t) \in \mathbb{R}$ then $(\arg x)' \equiv 0$ and we are left only with the first term in (8.222).

Strictly speaking, $\arg x'$ can become infinitely large when the sign of real $x(t)$ flips. But instead, for real x, we could "incorporate the sign into $|x|$ in (8.222)", by rewriting (8.221) as

$$\begin{aligned} \left\langle tx(t), \frac{x'(t)}{2\pi} \right\rangle &= \frac{1}{2\pi} \int_{-\infty}^{\infty} tx(t) x'(t) \, \mathrm{d}t = \frac{1}{2\pi} \int_{t=-\infty}^{\infty} tx(t) \, \mathrm{d}x(t) = \frac{1}{2\pi} \int_{t=-\infty}^{\infty} t \cdot \mathrm{d}\frac{x^2(t)}{2} \\ &= \frac{1}{4\pi} \int_{t=-\infty}^{\infty} t \cdot \mathrm{d}(x^2(t)) = \frac{1}{4\pi} \left(tx^2(t) \Big|_{t=-\infty}^{\infty} - \int_{-\infty}^{\infty} x^2(t) \, \mathrm{d}t \right) \\ &= \frac{1}{4\pi} \left(tx^2(t) \Big|_{t=-\infty}^{\infty} - 1 \right) \end{aligned} \tag{8.223a}$$

(remember that we assume that $x^2(t)$ is normalized, so the total energy of $x^2(t)$ is unity). Now usually $x(t)$ will decay sufficiently quickly to zero as $t \to \pm\infty$ (remember, at $1/t$ decay speed we only have conditional convergence at best, so its more common for $x(t)$ to decay faster than that). Assuming $x(t) = O(1/t)$ we have $tx^2(t) \to 0$ as $t \to \pm\infty$ and respectively the first term in the sum in (8.223a) turns to zero, and we obtain

$$\left\langle tx(t), \frac{x'(t)}{2\pi} \right\rangle = -\frac{1}{4\pi} \qquad \text{for } x(t) \in \mathbb{R} \tag{8.223b}$$

Now if $x(t)$ is not purely real, the second term in (8.222) becomes nonzero, while integrating the first term yields the same result as (8.223) and we have

$$\begin{aligned} \left| \left\langle tx(t), \frac{x'(t)}{2\pi} \right\rangle \right| &= \frac{1}{2\pi} \cdot \left| \int_{-\infty}^{\infty} t \cdot \left(|x| \cdot |x|' + i \cdot |x|^2 \cdot (\arg x)' \right) \mathrm{d}t \right| \\ &\geq \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} t \cdot |x| \cdot |x|' \cdot \mathrm{d}t \right| + \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} it \cdot |x|^2 \cdot (\arg x)' \cdot \mathrm{d}t \right| \\ &\geq \left| -\frac{1}{4\pi} \right| + \left| \frac{i}{2\pi} \int_{-\infty}^{\infty} t \cdot |x|^2 \cdot (\arg x)' \cdot \mathrm{d}t \right| \geq \frac{1}{4\pi} \end{aligned} \tag{8.224}$$

where the first integral's equality to $-1/4\pi$ is by the same argument as in (8.223).

▷ The difficulty with $(\arg x)'$ in (8.224) becoming infinitely large whenever $x(t)$ is going through the origin can be resolved by using Dirac delta to represent the derivative's value at these points. Upon integration $\delta(t)$ will produce a finite step, as we found in (8.197), thus producing a finite contribution to the value of the integral of $|x|^2 \cdot (\arg x)'$. But actually, the exact value of the contribution does not matter, what matters is that the respective imaginary term can only increase the total magnitude of the sum.

Infinitely large values of $|x|'$ can be dealt with by converting the Riemann integral of $t \cdot |x| \cdot |x|' \cdot \mathrm{d}t$ to the Riemann–Stieltjes integral of $t \cdot \frac{1}{2}\mathrm{d}(|x|^2)$.

Combining (8.220) with (8.224) yields

$$\sigma_x \sigma_X \geq \left| \left\langle tx(t), \frac{x'(t)}{2\pi} \right\rangle \right| \geq \frac{1}{4\pi} \qquad (\text{for } X = X(f)) \tag{8.225a}$$

Since $X(\omega)$ is 2π times wider than $X(f)$, for $X(\omega)$ we obtain a 2π times larger σ_X and a 2π times larger lower bound:

(Heisenberg uncertainty relationship as an abstract property of Fourier transform)

$$\sigma_x \sigma_X \geq \frac{1}{2} \qquad (\text{for } X = X(\omega)) \tag{8.225b}$$

The inequality (8.225) puts the lower bound on the product $\sigma_x \sigma_X$, thereby energy cannot become indefinitely concentrated in both Fourier transform domains simultaneously. Once the limit (8.225) is reached, trying to make the energy more concentrated in one domain will necessarily make it less concentrated in the other domain. We already have seen in (8.208) that if we try to shrink a function horizontally, this will stretch the function by exactly the same factor in the other domain. However, even if we try to get smarter and try to use some other means of horizontally shrinking the function, we still won't get past the hard limit (8.225).

The inequality (8.225) bears different names in different application contexts. Most famously, in quantum mechanics it is known as the *Heisenberg uncertainty relationship* (also as *Kennard inequality*). In signal processing it is known as *Gabor limit*.

Goal 8.28 completed. ■

8.18.5 Maximally concentrated energy

Goal 8.29. Find out which shape(s) of $x(t)$ have the energy simultaneously maximally concentrated in both domains. That is find out, which $x(t)$ turn (8.225) into an equality. *Hint:* (8.225) is a form of Cauchy–Schwarz inequality.

The Heisenberg uncertainty relationship (8.225) is essentially the Cauchy–Schwarz inequality (8.220), therefore it turns into an equality when both vectors $tx(t)$ and $x'(t)/2\pi$ become collinear:

$$a \cdot tx(t) = \frac{x'(t)}{2\pi} \qquad (a \neq 0) \tag{8.226a}$$

Although, more precisely, collinearity is not sufficient, since actually the relationship (8.225) is a transitive combination of the Cauchy–Schwarz inequality (8.220) and the inequality (8.224). While collinearity is sufficient and necessary to turn (8.220) into an equality, this doesn't guarantee that (8.224) is an equality.

We can recall that (8.224) turns into an equality (8.223) if $x(t)$ is purely real, so we should require $x(t) \in \mathbb{R}$ and respectively $a \in \mathbb{R}$ in (8.226). Although, more precisely speaking, the extra imaginary term in (8.222) and (8.224) can still turn to zero for certain essentially complex $x(t)$:

$$\int_{-\infty}^{\infty} t \cdot |x|^2 \cdot (\arg x)' \cdot \mathrm{d}t = 0 \tag{8.226b}$$

Thus if both (8.226) are satisfied, the Heisenberg uncertainty relationship (8.225) becomes an equality and we thereby are having the minimum possible value of the product $\sigma_x \sigma_X$.

Solving the differential equation (8.226a) we obtain

$$\frac{x'(t)}{x(t)} = 2\pi a t$$

$$(\ln x(t))' = 2\pi a t$$

$$\ln x(t) = \pi a t^2 + C$$

$$x(t) = A e^{\pi a t^2} \tag{8.227}$$

where $A = e^C$. At real $a < 0$ the function (8.227) essentially represents the well known Gaussian distribution function. At $\operatorname{Re} a \geq 0$ Fourier transform diverges (as well as the energy of $x(t)$ is infinite). Thus, Gaussian distribution function minimizes the product $\sigma_x \sigma_X$. In that sense it's a function where the energy is maximally concentrated simultaneously in both Fourier transform domains.

We would still like to know what happens for complex values of a such that $\operatorname{Re} a < 0$. Considering $x(t) = e^{\pi a t^2}$ for such complex a in the context of the condition (8.226b) we notice that $x(t)$ is an even function of t and therefore $\arg x$ and $|x|$ are even functions of t. Respectively $(\arg x)'$ is an odd function and so is the product $|x|^2 \cdot (\arg x)'$, while the total product under the integral sign $t \cdot |x|^2 \cdot (\arg x)'$ is an even function again and therefore the condition (8.226b) can be rewritten as

$$\int_{0}^{\infty} t \cdot |x|^2 \cdot (\arg x)' \cdot \mathrm{d}t = 0 \tag{8.228}$$

Further noticing that

$$\arg x = \operatorname{Im}(\pi a t^2) = \pi t^2 \operatorname{Im} a$$

and

$$(\arg x)' = 2\pi \operatorname{Im} a \cdot t$$

we conclude that the sign of $\arg x$ doesn't change for $t > 0$ and coincides with the sign of $\operatorname{Im} a$. The signs of t and $|x|^2$ do not change either, and these expressions are strictly positive for $t > 0$. Thus the integral in (8.228) is nonzero and therefore (8.228) doesn't hold and (8.226b) doesn't hold either. Thus $\operatorname{Re} a < 0$ doesn't turn Heisenberg uncertainty relationship into an equality, and we do

need a purely real $x(t)$ and $a < 0$ in (8.227). That is $x(t)$ must be the Gaussian distribution function.

Goal 8.29 completed. ■

If $x(t) = Ae^{\pi a t^2}$, then what is its Fourier transform $X(\omega)$? By the duality of Fourier transform we can guess that $X(\omega)$ must also satisfy both of (8.226) and therefore also must be the Gaussian distribution function.

▷ Technically the dot product $\langle tx, x' \rangle$ can be computed in the f domain instead of t domain, yielding (up to some constant factor, which we don't bother with) $\langle X', \omega X \rangle$. Both expressions represent one and the same dot product having one and the same value (by the unitarity of Fourier transform). The lengths of the two vectors involved in both products also shouldn't change. Thus the Cauchy–Schwarz inequality (8.220) can be rewritten in the f-domain terms via the dot product of the vectors X' and ωX. Therefore (8.226) can be rewritten in the f-domain terms as well and so can be (8.227) yielding $X(\omega) = Be^{\pi b \omega^2}$ (for some B and b)

It would be still instructive to obtain $X(\omega)$ directly as Fourier transform of $x(t) = Ae^{-at^2}$ (where for convenience we switched the notation from $e^{\pi a t^2}$ to e^{-at^2}). For simplicity we will let $x(t) = e^{-t^2}$. If necessary, the results for $A \neq 1$ and $a \neq 1$ can be obtained applying linearity of Fourier transform (for $A \neq 1$) and the scaling property (8.208b) (for $a \neq 1$). By (8.71a)

$$\begin{aligned} X(\omega) &= \int_{-\infty}^{\infty} e^{-t^2} e^{-i\omega t}\,\mathrm{d}t = \int_{-\infty}^{\infty} e^{-t^2 - i\omega t}\,\mathrm{d}t \\ &= e^{-\omega^2/4} \int_{-\infty}^{\infty} e^{-(t^2 + 2t(i\omega/2) - \omega^2/4)}\,\mathrm{d}t = e^{-\omega^2/4} \int_{-\infty}^{\infty} e^{-(t+i\omega/2)^2}\,\mathrm{d}t \\ &= e^{-\omega^2/4} \int_{-\infty+i\omega/2}^{+\infty+i\omega/2} e^{-z^2}\,\mathrm{d}z \end{aligned}$$

where the integration in the complex plane is done along a straight line from $z = -\infty + i\omega/2$ to $z = +\infty + i\omega/2$ (where $+\infty$ and $-\infty$ denote the infinities of the two ends of the real axis). Now since e^{-z^2} is analytic, we can change the integration path into one consisting of three linear segments:

$$\int_{-\infty+i\omega/2}^{+\infty+i\omega/2} e^{-z^2}\,\mathrm{d}z = \int_{-\infty+i\omega/2}^{-\infty} e^{-z^2}\,\mathrm{d}z + \int_{-\infty}^{+\infty} e^{-z^2}\,\mathrm{d}z + \int_{+\infty}^{+\infty+i\omega/2} e^{-z^2}\,\mathrm{d}z$$

(where infinities are understood in the limit sense). The first and the third integrals integrate infinitely small e^{-z^2} over a finite range each and therefore are equal to zero, therefore

$$\int_{-\infty+i\omega/2}^{+\infty+i\omega/2} e^{-z^2}\,\mathrm{d}z = \int_{-\infty}^{+\infty} e^{-z^2}\,\mathrm{d}z = \sqrt{\pi}$$

where in the right-hand side we're having the famous Gaussian integral over the real axis, which is known to be equal to $\sqrt{\pi}$. We therefore obtain

$$X(\omega) = e^{-\omega^2/4} \int_{-\infty+i\omega/2}^{+\infty+i\omega/2} e^{-z^2}\,\mathrm{d}z = e^{-\omega^2/4} \int_{-\infty}^{\infty} e^{-t^2}\,\mathrm{d}t = e^{-\omega^2/4} \cdot \sqrt{\pi}$$

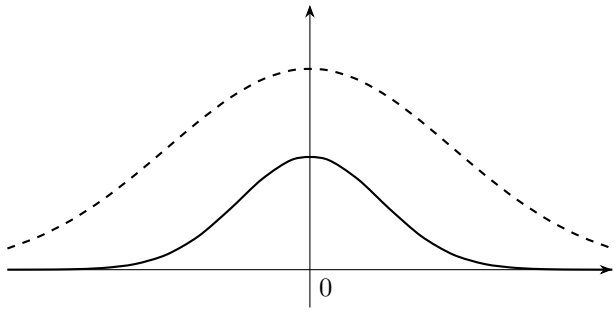


Figure 8.18: Fourier transform of a Gaussian (solid) is another Gaussian (dashed). Note that the wider Gaussian is also higher in this picture because we assume the density-like version (8.71) of the Fourier transform. In the symmetric version (8.73) a wider Gaussian would be lower to keep the area under the squared function constant, as required by Parseval's theorem.

Fig. 8.18 illustrates.

So $X(\omega)$ is almost identical to $x(t)$. Can we make both fully identical somehow? If we stretch $x(t)$ by $\sqrt{2}$, obtaining $x(t) = e^{-t^2/2}$, then by the scaling property (8.208b) $X(\omega)$ should shrink by the same factor, yielding $Be^{-\omega^2/2}$ for some B. Therefore we let

$$y(t) = x(t/\sqrt{2}) = e^{-t^2/2}$$

By (8.208b) we have

$$Y(\omega) = \sqrt{2} \cdot X(\omega \cdot \sqrt{2}) = \sqrt{2\pi} \cdot e^{-\omega^2/2} = \sqrt{2\pi} \cdot y(\omega)$$

Thus $e^{-t^2/2}$ is an *eigenfunction* of Fourier transform corresponding to the eigenvalue $\sqrt{2\pi}$. Notice that in the fully symmetric version of Fourier transform (8.73) the resulting $Y(t)$ is additionally divided by $\sqrt{2\pi}$ and the eigenvalue corresponding to the eigenfunction $e^{-t^2/2}$ turns to 1:

(Gaussian distribution is an eigenfunction of Fourier transform)

$$x(t) = e^{-t^2/2} \iff X(\omega) = e^{-\omega^2/2} \qquad \text{(for symmetric transform (8.73))}$$

Thus the eigenfunction $e^{-t^2/2}$ of the Fourier transform is obtained as the solution of the collinearity equation (8.226a) at $a = -1/2\pi$, where (8.226a) simply takes the form

$$tx(t) + x'(t) = 0$$

▷ In quantum mechanics the Gaussian distribution function minimizes the product of the position and momentum uncertainties Δx and Δp, which, in terms of Fourier transform, correspond to σ_x and σ_X. It also describes the wavefunction corresponding to the minimum energy level of the quantum harmonic oscillator. So the minimum energy corresponds to the minimum value of the product of uncertainties $\Delta x \Delta p$. The Hamiltonian for this oscillator is symmetric with respect to swapping the position and momentum operators, which suggests that solutions to the corresponding Schrödinger equation (given by the eigenfunctions of the Hamiltonian) should be identical in both position and momentum domains.

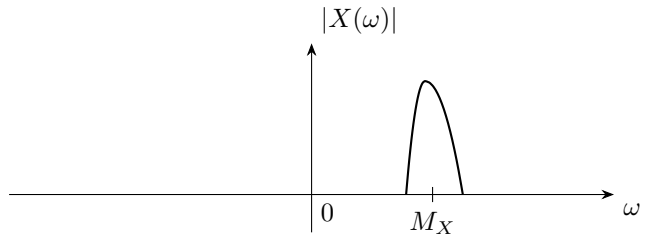


Figure 8.19: Narrowband spectrum.

8.18.6 Energy center shift

Narrowband spectrum

Let $x(t)$ and $y(t)$ be two functions with their respective Fourier transforms $X(\omega)$ and $Y(\omega)$ and let $U(\omega) = X(\omega)Y(\omega)$, that is we are having the same "filtering setup" as the one used in the discussion of convolution in Section 8.11. By (8.109) $u(t) = x * y$.

We can notice that, by the time-shift property (8.112) of Fourier transform, we can achieve a shift of $x(t)$ by the amount x_0 by letting $Y(\omega) = e^{-i\omega t_0}$:

$$Y(\omega) = e^{-i\omega t_0} \iff u(t) = x(t - t_0)$$

Notice that the phase part of $Y(\omega)$ is thereby a linear function $\varphi_y(\omega) = -\omega t_0$.

Now suppose $X(\omega)$ is concentrated in a small interval $[\omega_1, \omega_2]$, being practically zero outside of that interval (Fig. 8.19). This suggests that $M_X \in [\omega_1, \omega_2]$ and, since we said that the interval is small, σ_X is small. In this case we don't care about the values of $Y(\omega)$ outside of $[\omega_1, \omega_2]$, since $X(\omega)$ is practically zero there and thus $Y(\omega)$ doesn't noticeably affect the value of $U(\omega)$. So we can have a pretty much arbitrary function $Y(\omega)$ as long as $Y(\omega) \approx e^{-i\omega t_0}$ and respectively $\varphi_y(\omega) = -\omega t_0$ for $\omega \in [\omega_1, \omega_2]$.

Now suppose we also lift that restriction, that is we do not require $Y(\omega) \approx e^{-i\omega t_0}$ for $\omega \in [\omega_1, \omega_2]$. That is $Y(\omega)$ is almost completely arbitrary. Let's expand the phase function $\varphi_y(\omega) = \arg Y(\omega)$ into Taylor series up to the 1st-order term around the point ω_0, which we choose at the energy center of $X(\omega)$, that is $\omega_0 = M_X$:

$$\varphi_y(\omega) \approx \varphi_y(\omega_0) + \varphi_y'(\omega_0) \cdot (\omega - \omega_0) = \big(\varphi_y(\omega_0) - \varphi_y'(\omega_0)\omega_0\big) + \varphi_y'(\omega_0)\omega \quad (8.229)$$

Fig. 8.20 illustrates. The main difference from our previous case is contained in the constant term $\varphi_y(\omega_0) - \varphi_y'(\omega_0)\omega_0$. Indeed, if this term hadn't been there, we would have had $\varphi_y(\omega) = \varphi_y'(\omega_0)\omega = -\omega\tau_0$, where $\tau_0 = -\varphi_y'(\omega_0)$, and respectively $Y(\omega) \approx e^{-i\omega\tau_0}$ for $\omega \approx \omega_0$. Let's therefore study the effect of this constant term.

Let therefore

$$\varphi_y(\omega) \approx \varphi_0 - \tau_0\omega \qquad \text{for } \omega \approx \omega_0$$

where

$$\varphi_0 = \varphi_y(\omega_0) - \varphi_y'(\omega_0)\omega_0$$

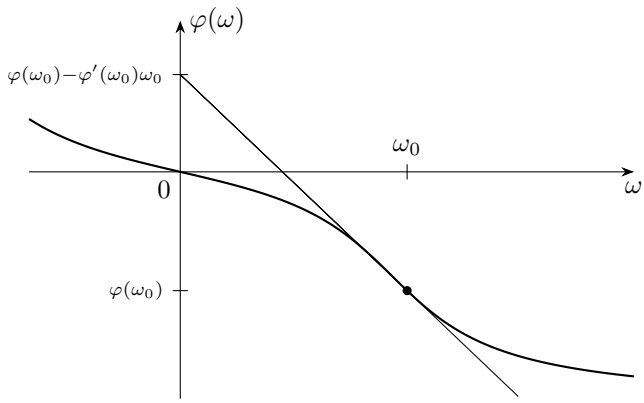


Figure 8.20: Approximate linearization of the spectral phase.

$$\tau_0 = -\varphi_y'(\omega_0)$$

Then, in the neighborhood of ω_0,

$$Y(\omega) \approx |Y(\omega_0)| \cdot e^{-i\varphi_y(\omega)} = |Y(\omega_0)| \cdot e^{i(\varphi_0 - \tau_0\omega)} = |Y(\omega_0)| \cdot e^{i\varphi_0} \cdot e^{-\tau_0\omega}$$

The effect of $Y(\omega)$ on $X(\omega)$ consists therefore of 3 parts:

1. Scaling the absolute magnitude by $|Y(\omega_0)|$. This effect is identical in both t- and f-domains due to the linearity of Fourier transform.

2. Shifting the phase component of $X(\omega)$ by the constant term φ_0.

3. Additionally shifting the phase component of $X(\omega)$ by the linear term $-\tau_0\omega$. We already know that in the t domain it shifts $x(t)$ by τ_0.

We are therefore still not certain what is the t-domain effect of the phase shift by the constant term, caused by the factor $e^{i\varphi_0}$ of $Y(\omega)$. We can however quickly realize that multiplication of the spectrum $X(\omega)$ by a complex constant $e^{i\varphi_0}$ just multiplies $x(t)$ by the same constant $e^{i\varphi_0}$. There is no difference from multiplication by $|Y(\omega_0)|$, it's still just the linearity of Fourier transform. If $x(t)$ was just a complex sinusoid of the form $e^{i\omega_x t}$ of some frequency ω_x, multiplication by $e^{i\varphi_0}$ would have just changed the initial phase of that sinusoid. But we are having a somewhat more general $x(t)$, which doesn't look like a constant-frequency sinusoid. Or does it?

Notice that, since $X(\omega)$ is concentrated around $\omega_0 = M_x$, it behaves somewhat similarly to $X(\omega) = \delta(\omega - \omega_0)$, the latter in the t domain looking like a complex sinusoid $e^{i\omega_0 t}$. We might therefore suspect that our $x(t)$ retains some features of that sinusoid.

To make the comparison to $\delta(\omega - \omega_0)$ more explicit, let's represent $X(\omega)$ as $X(\omega) = \bar{X}(\omega - \omega_0)$ where thereby $M_{\bar{X}} = 0$ and $\sigma_{\bar{X}} = \sigma_X$. The value of σ_X being small corresponds to the fact that we are attempting to see $\bar{X}(\omega)$ as an approximation of $\delta(\omega)$. By the frequency-shift property (8.113) of Fourier transform,

$$x(t) = \bar{x}(t)e^{i\omega_0 t}$$

which suggests that $x(t)$ is has a sinusoidal component (factor) of frequency ω_0. Theoretically this sinusoid could be masked by other features of $\bar{x}(t)$, but how likely is it that $\bar{x}(t)$ is strongly varying? We can notice that, by (8.218b), strong variations in $\bar{x}(t)$ imply large $\sigma_{\bar{X}}$. But we said that $\sigma_{\bar{X}}$ is small. Therefore $\bar{x}(t)$ is not expected to have strong variations, but should be rather varying slowly, and respectively $x(t) = \bar{x}(t)e^{i\omega_0 t}$ implies a prominent sinusoidal component of frequency ω_0 in $x(t)$.

Now if we multiply the spectrum $X(\omega)$ of $x(t) = \bar{x}(t)e^{i\omega_0 t}$ by the constant $e^{i\varphi_0}$, this also multiplies $x(t)$ by the same constant and $x(t)$ turns into

$$x_\varphi(t) = \bar{x}(t)e^{i\omega_0 t} \cdot e^{i\varphi_0} = \bar{x}(t)e^{i(\omega_0 t + \varphi_0)}$$

That is shifting the phases of the spectrum $X(\omega)$ by φ_0 in the t domain looks like phase-shifting the sinusoidal component of $x(t) = \bar{x}(t)e^{i\omega_0 t}$.

The t-domain effect of multiplication by $Y(\omega) \approx |Y(\omega_0)| \cdot e^{i\varphi_0} \cdot e^{-\tau_0\omega}$ (if the spectrum of $X(\omega)$ is concentrated around $\omega_0 = M_X$) is thereby

1. Scaling the absolute magnitude by $|Y(\omega_0)|$.
2. Shifting the phase of the sinusoidal component $e^{i\omega_0 t}$ of $x(t)$ by φ_0. This doesn't change M_x or σ_x, since the absolute magnitude of $x(t)$ is unaffected by such change.
3. Shifting the entire function $x(t)$ by τ_0. Thereby M_x is also shifted by τ_0, while σ_x is unchanged.

▷ In signal processing the value $\tau_y(\omega) = -\varphi_y'(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\omega}\arg Y(\omega)$ is referred to as *group delay*.

Hermitian narrowband spectrum

So far we have been assuming a generic case of a complex $x(t)$. Our results do not immediately apply to a real $x(t)$. Indeed, the spectrum of a real $x(t)$ must be Hermitian, therefore it cannot be concentrated around a single point $\omega_0 \neq 0$. However, it can be concentrated around two points ω_0 and $-\omega_0$. We can therefore assume that $X(\omega)$ is representable as

$$X(\omega) = \frac{\bar{X}(\omega - \omega_0) + \bar{X}^*(-(\omega - \omega_0))}{2}$$

Fig. 8.21 illustrates.

Respectively, by (8.54b), there will be a real sinusoidal component in $x(t)$:

$$\begin{aligned} x(t) &= \frac{\bar{x}(t)e^{i\omega_0 t} + \bar{x}^*(t)e^{-i\omega_0 t}}{2} = \mathrm{Re}\big(\bar{x}(t)e^{i\omega_0 t}\big) \\ &= |\bar{x}(t)| \cdot \cos(\omega_0 t + \arg \bar{x}(t)) \end{aligned}$$

where the amplitude $|\bar{x}(t)|$ and the phase $\arg \bar{x}(t)$ are changing slowly compared to the sinusoid's frequency. We are therefore having a situation similar to the one shown earlier in Fig. 8.17 on page 492, except that now there is additional variation in the sinusoid's phase. The slowly changing amplitude $|\bar{x}(t)|$ works as a kind of an "envelope" applied to the "carrier" sinusoidal signal, and we

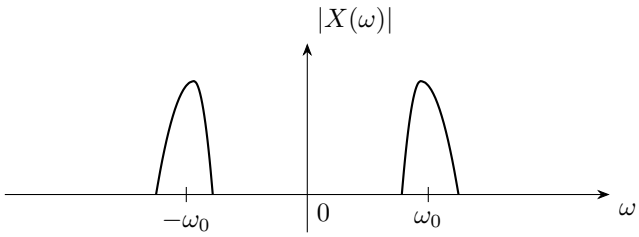


Figure 8.21: Hermitian narrowband spectrum of a real $x(t)$.

can understand $\arg \bar{x}(t)$ as a "phase envelope". This phase envelope is also essentially responsible for the sinusoidal component not really having a defined frequency, since the actual true frequency of this sinusoid, being equal to the derivative of the full phase, is equal to $\omega_0 + \frac{\mathrm{d}}{\mathrm{d}t} \arg \bar{x}(t)$).

Now, if $y(t)$ is also real and $Y(\omega)$ is Hermitian, we have $\varphi(\omega_0) = -\varphi(-\omega_0)$ and $\varphi'_y(\omega_0) = \varphi'_y(-\omega_0)$. Thus the t-domain counterparts of both spectral components $\bar{X}(\omega - \omega_0)$ and $\bar{X}(-(\omega - \omega_0))$ are shifted by exactly the same amount and in the same direction. The value $\tau_y(\omega_0) = -\varphi'_y(\omega_0)$ can be therefore seen as defining the t-domain shift of the envelopes, while the value $\varphi_y(\omega_0)$ (in combination with $\varphi'_y(\omega_0)$) determines the phase shift of the sinusoidal carrier:

- envelopes are approximately shifted by $\tau_y(\omega_0) = -\varphi'_y(\omega_0)$
- the phase of the sinusoidal carrier of $x(t)$ is shifted by $\varphi_y(\omega_0) - \varphi'_y(\omega_0)\omega_0$.

General case

It can be instructive to find out how do our results align with the equation (8.214), which is the one that can be applied in the general case.

Goal 8.30. Use (8.214) to reobtain the results obtained earlier from the approximate linearization of $\varphi_y(\omega)$ at $\omega = \omega_0$. *Hint:* apply (8.214) to $U(\omega) = X(\omega)Y(\omega)$.

By (8.214) the energy center of $u(t) = x * y$ is

$$\begin{aligned} M_u &= \int_{-\infty}^{\infty} |U|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg U\right) \frac{\mathrm{d}\omega}{2\pi} \\ &= \int_{-\infty}^{\infty} |X|^2 \cdot |Y|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega}(\arg X + \arg Y)\right) \frac{\mathrm{d}\omega}{2\pi} \end{aligned} \tag{8.230}$$

since $|U| = |X| \cdot |Y|$, $\arg U = \arg X + \arg Y$. We can immediately notice the addition of the extra term equal to $\tau_y(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\omega} Y(\omega)$ under the integral sign, suggesting that M_x might become shifted by $\tau_y(\omega_0)$, as in our previous results. But that requires some additional constraints on $X(\omega)$ and/or $Y(\omega)$. We also need to mind the proper (re-)normalization of the product $|X|^2 \cdot |Y|^2$.

We could immediately impose similar constraints on $X(\omega)$ as in the previous discussion, but it would be quite instructive to first consider the case where we impose a constraint $|Y| \equiv 1 \ \forall \omega$ on $Y(\omega)$, while $X(\omega)$ is left without constraints. In this case no additional renormalization of the product $|X|^2 \cdot |Y|^2$ is necessary, since $|U| = |X|$.

▷ Usually we assume $Y \to 0$ as $\omega \to \infty$ to ensure the convergence of the inverse Fourier transform. However we also have been dealing with Dirac delta function $\delta(t)$ which has Fourier transform equal to 1. Similarly to $\delta(t)$, there can be further functions $y(t)$ such that $|Y(\omega)| \equiv 1$ (in signal processing, filters having such $Y(\omega)$ are referred to as *allpass filters*).

The simplest example is probably $y(t) = \delta(t - t_0)$ where t_0 is some constant value. By the time-shift property $Y(\omega) = e^{-i\omega t_0}$, therefore $|Y(\omega)| \equiv 1$ (from the signal processing perspective, such filter is simply a delay). Notice that even though this function, like $\delta(t)$, has infinite energy (which is easiest to verify by computing the energy in the f domain), it still has an energy center at $t = t_0$ (where we can imagine normalizing the function from the infinite energy to unit energy). In the t domain the answer $t = t_0$ for the energy center is obvious, but also in the f domain by (8.214) we have

$$M_y = \int_{-\infty}^{\infty} |Y|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg Y\right) \frac{\mathrm{d}\omega}{2\pi} = \int_{-\infty}^{\infty} |Y|^2 \cdot (-(-t_0)) \frac{\mathrm{d}\omega}{2\pi} = \int_{-\infty}^{\infty} |Y|^2 \cdot t_0 \frac{\mathrm{d}\omega}{2\pi}$$

It's not difficult to see that if Y is normalized so that is has unit energy, we obtain $M_y = t_0$.

Goal 8.30a. Compute the amount of energy center shift of (unconstrained) $x(t)$ under the restriction $|Y| \equiv 1$.

Given $|Y| \equiv 1$, (8.230) turns into

$$\begin{aligned} M_u &= \int_{-\infty}^{\infty} |X|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega}(\arg X + \arg Y)\right) \frac{\mathrm{d}\omega}{2\pi} \\ &= \int_{-\infty}^{\infty} |X|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg X\right) \frac{\mathrm{d}\omega}{2\pi} + \int_{-\infty}^{\infty} |X|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg Y\right) \frac{\mathrm{d}\omega}{2\pi} \\ &= M_x + \int_{-\infty}^{\infty} |X|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg Y\right) \frac{\mathrm{d}\omega}{2\pi} \end{aligned}$$

That is the energy center shift is defined by

(frequency-domain expression for the time-domain energy shift)

$$\begin{aligned} \Delta M_x &= M_u - M_x \\ &= \int_{-\infty}^{\infty} |X|^2 \cdot \left(-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg Y\right) \frac{\mathrm{d}\omega}{2\pi} = \int_{-\infty}^{\infty} |X|^2 \cdot \tau_y \frac{\mathrm{d}\omega}{2\pi} \end{aligned} \tag{8.231}$$

We see that similarly to how M_x was defined by $-\frac{\mathrm{d}}{\mathrm{d}\omega} \arg X$ in (8.214), ΔM_x is defined by $\tau_y(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\omega} \arg Y$, while the weights are still given by $|X|^2$.

▷ Notice that, differently from our previous results, the value of ΔM_x is given exactly by (8.231) and not just approximately. This holds, however, under the assumption that $Y(\omega)$ defines an allpass filter: $|Y| \equiv 1$.

Goal 8.30a completed. ■

Still retaining the condition $|Y| \equiv 1$, let's additionally impose the previous constraints on $X(\omega)$. That is, let $X(\omega)$ be concentrated around $\omega_0 = M_X$ with a small σ_X. In this case only the values of $Y(\omega)$ in the vicinity of $\omega = \omega_0$ will

affect ΔM_x given by (8.231), while everywhere else $Y(\omega)$ can have arbitrary values. Assuming $\varphi_y'(\omega) \approx \varphi_y'(\omega_0) = -\tau_y(\omega_0) = -\tau_0$ in the vicinity of $\omega = \omega_0$ (where $\varphi_y(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\omega}Y$) we obtain from (8.231)

$$\Delta M_x \approx \int_{-\infty}^{\infty} |X|^2 \cdot \tau_0 \, \frac{\mathrm{d}\omega}{2\pi} = \tau_0$$

matching our previous discussion.

The next step is to allow $Y(\omega)$ to become arbitrary also in the vicinity of $\omega = \omega_0$. We shall however assume that $Y(\omega)$ doesn't change drastically in that range: $Y(\omega) \approx Y(\omega_0)$ for $\omega \approx \omega_0$. We can notice that in this case we need to renormalize $U(\omega)$, which can be approximately achieved by renormalizing $Y(\omega)$ so that $|Y(\omega_0)| \approx 1$, and thus we obtain the same result $\Delta M_x \approx \tau_0$.

Thereby our results obtained earlier by the linearization of the phase part of $Y(\omega)$ match the ones obtained from (8.214).

Goal 8.30 completed. ■

8.19 Finite computability

Of the 4 types of Fourier transform, 3 contain infinite summation and/or integration. Neither of these operations can be performed numerically in finite time. Thus these 3 types can be used only for theoretical analysis, but are unsuitable for practical numeric computations. Actually it's even not possible to just store infinitely long discrete-domain functions or continuous functions (regardless of the domain length) in finite memory. Therefore the only Fourier transform type suitable for numeric computations is DFT. In this section we shall discuss the application of DFT for spectral analysis of functions which are neither discrete nor periodic (or at least not both).

▷ One more aspect requiring infinite computation time and memory is processing of real numbers, which require an infinite amount of memory to be stored with infinite precision. In practice, limited precision storage must be used. If this precision is low, the *quantization effects* can become pretty prominent and also show up in the spectral domain. We won't discuss these effects here, as the standard precisions offered by modern-day computers are usually enough, and it would be a pretty specialized topic of *digital signal processing* anyway.

8.19.1 Sampling

Suppose we are given a real-world function $x(t)$. This function can be a signal, if t has time semantics, or it can be something else. We assume that the shape of the function is pretty much arbitrary, therefore it cannot be represented by an analytical expression such as $x(t) = \sin t$. Instead we need to store a finite number of the function's samples. This means that we want take a continuous-domain function $x(t)$ (potentially defined on the entire real axis, or at least on a pretty long interval of t) and convert it to a discrete-domain function $\hat{x}[n]$, defined on a sufficiently short interval of n to make the required storage amount not just finite, but small enough to be practical. For now let's concentrate on the conversion of $x(t)$ defined on $t \in \mathbb{R}$ to $\hat{x}[n]$ defined on $n \in \mathbb{Z}$, that is

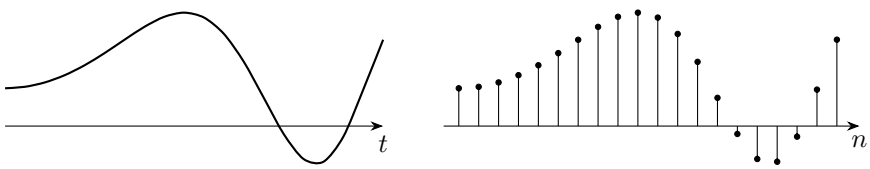


Figure 8.22: Sampling of a continuous-argument function.

on the *sampling* of $x(t)$. Truncation of the domain of x will be discussed in Subsection 8.19.2.

An obvious way to obtain $\hat{x}[n]$ from $x(t)$ is to let

$$\hat{x}[n] = x(nT) \tag{8.232a}$$

where T is the *sampling period*. The conversion described by (8.232a) is referred to as *regular sampling*, since the samples are taken at regularly spaced points. It is convenient and customary to assume $T = 1$, in which case other sampling periods can be handled by a change of the units of time. Thus, instead of (8.232a), let

$$\hat{x}[n] = x(n) \tag{8.232b}$$

Fig. 8.22 illustrates.

As mentioned, we still assume that n goes from $-\infty$ to ∞, thus $\hat{x}[n]$ is a bilateral infinite sequence. The spectrum of such sequence needs to be obtained by discrete-time Fourier transform. However, we should recall that DTFT spectra are 1-periodic, while the spectrum of the original $x(t)$ is obtained by integral Fourier transform and therefore is infinitely long. How do the two spectra relate to each other, in other words, what is the effect of the sampling (8.232) on the spectrum of $x(t)$?

To answer that question let's start with the direct DTFT transform equation (8.58a) and try to rewrite it in the form of the direct integral Fourier transform equation (8.70a). To achieve that we'd need to rewrite the sum in (8.58a) as an integral in (8.70a). We already discussed a way to rewrite sums as integrals in Subsection 8.13.10. Applying the same technique we obtain

$$\begin{aligned}\hat{X}(f) &= \sum_{n=-\infty}^{\infty} \hat{x}[n]e^{-2\pi i f n} = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} x(t)e^{-2\pi i f t}\delta(t-n)\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} \left(\sum_{n=-\infty}^{\infty} x(t)\delta(t-n)\right) e^{-2\pi i f t}\,\mathrm{d}t = \int_{-\infty}^{\infty} \hat{x}(t)e^{-2\pi i f t}\,\mathrm{d}t \end{aligned} \tag{8.233a}$$

where

$$\hat{x}(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t-n) = x(t)\sum_{n=-\infty}^{\infty} \delta(t-n) = x(t)\delta_1(t) \tag{8.233b}$$

That is the DTFT spectrum $\hat{X}(f)$ of $\hat{x}[n]$ is also an integral Fourier transform of the function $\hat{x}(t)$ which is the product of $x(t)$ and the 1-periodic Dirac comb $\delta_1(t)$. But the spectrum of $x(t)\delta_1(t)$, by the convolution property, is the convolution of spectra of $x(t)$ and $\delta_1(t)$. The integral Fourier transform spectrum of $x(t)$ is $X(f)$, what is the integral Fourier transform spectrum of $\delta_1(t)$?

Applying direct integral Fourier transform (8.70a) yields

$$\begin{aligned}\Delta_1(f) &= \int_{-\infty}^{\infty} \delta_1(t) e^{-2\pi i f t}\,\mathrm{d}t \overset{(8.158a)}{=} \int_{-\infty}^{\infty} \left(\sum_{n=-\infty}^{\infty} \delta(t-n)\right) e^{-2\pi i f t}\,\mathrm{d}t \\ &= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(t-n) e^{-2\pi i f t}\,\mathrm{d}t = \sum_{n=-\infty}^{\infty} e^{-2\pi i f n} \\ &\overset{(8.132)}{=} \lim_{N\to\infty} D_N(-2\pi f) = \lim_{N\to\infty} D_N(2\pi f) \overset{(8.138)}{=} \delta_1(f)\end{aligned}$$

▷ In the application of (8.132) we have matched $\sum_k e^{2\pi i k F t}$ in (8.132) against $\sum_n e^{-2\pi i f n}$ in the above. In doing so we replaced k by n and Ft by $-f$.

Thus

(Fourier transform of $\delta_1(t)$ is $\delta_1(t)$)

$$\Delta_1(f) = \int_{-\infty}^{\infty} \delta_1(t) e^{-2\pi i f t}\,\mathrm{d}t = \delta_1(f) \tag{8.234}$$

Therefore 1-periodic Dirac comb is another eigenfunction of Fourier transform.

Thus, by the convolution property, the spectrum of $\hat{x}(t) = x(t)\delta_1(t)$ is

$$\begin{aligned}\hat{X}(f) = X(f) * \delta_1(f) &= \int_{-\infty}^{\infty} X(f-\nu)\delta_1(\nu)\,\mathrm{d}\nu \\ &= \int_{-\infty}^{\infty} X(f-\nu)\left(\sum_{n=-\infty}^{\infty} \delta(\nu-n)\right)\mathrm{d}\nu = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} X(f-\nu)\delta(\nu-n)\,\mathrm{d}\nu \\ &= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} X(f-n-\nu)\delta(\nu)\,\mathrm{d}\nu = \sum_{n=-\infty}^{\infty} (X(f-n) * \delta(f))\end{aligned}$$

that is simply

(spectrum of a sampled signal)

$$\hat{X}(f) = X(f) * \delta_1(f) = \sum_{n=-\infty}^{\infty} X(f-n) \tag{8.235}$$

Thus, the spectrum of $\hat{x}[n]$ and $\hat{x}(t)$ is obtained by adding up shifted copies of the spectrum of $X(t)$. It is not difficult to see that $\hat{X}(f)$ given by (8.235) is 1-periodic. This is exactly what is expected of a DTFT spectrum, in a way supporting our statement that the spectra of $\hat{x}[n]$ and $\hat{x}(t)$ are identical.

If we restrict f in $\hat{X}(f)$ to the standard interval $[-\frac{1}{2}, \frac{1}{2}]$, we can see that the values of $\hat{X}(f)$ in that interval are obtained by adding up $X(f)$ on all intervals $[n-\frac{1}{2}, n+\frac{1}{2}]$ $\forall n \in \mathbb{Z}$. Outside of $[-\frac{1}{2}, \frac{1}{2}]$ the spectrum $\hat{X}(f)$ simply repeats periodically, being a 1-periodic function.

▷ Recall that in the very beginning by (8.6b) we have established that the sinusoidal functions which we use as the basis for the Fourier expansion of a function are periodic not only in the original (time) domain but also in the transform (frequency) domain. We could rewrite (8.6b) for the DTFT case as $h_{f+1}[n] = h_f[n]$

where $h_f[n] = e^{2\pi i f n}$ are the basis functions, where the subscript f in h_f is real rather than integer. Notice that there is no DTFT counterpart of (8.6a) since $h_f[n]$ are not periodic with respect to n.

The equality (8.235) in a way reflects the fact that we cannot distinguish between h_{f+k} and h_f $\forall k \in \mathbb{Z}$, therefore it simply "lumps" all those equivalent basis functions together. This aliasing of the basis sinusoidal functions is exactly what is referred to as *aliasing* in the context of sampling.

The overlapping of spectra in (8.235) is commonly amended by filtering the spectrum of $x(t)$ to erase all spectral components outside of the $(-\frac{1}{2}, \frac{1}{2})$ range. This process is known as *bandlimiting*. By bandlimiting $x(t)$ we potentially lose a good part of the spectrum, but at least avoid the "lumping" of aliased sinusoidal components.

From the spectral point of view, bandlimiting to $(-\frac{1}{2}, \frac{1}{2})$ (which is the most common form of bandlimiting) can be performed by multiplying the spectrum $X(f)$ by $\operatorname{rect}(f)$. By the convolution property, in the t domain this corresponds to computing convolution with the sinc function (the latter being given by (8.206) at $F = 1$):

$$\bar{x}(t) = x(t) * \operatorname{sinc} \pi t$$

▷ Apparently bandlimiting doesn't have to be necessarily done to the range $(-\frac{1}{2}, \frac{1}{2})$. We can pick up any other range of a unit length (or even a combination of non-aliasing ranges with a total length equal to unity). However, unless this range is symmetric with respect to the origin, the function after bandlimiting will be essentially complex. So, if we want to keep the function real, we'd need a range which is simultaneously symmetric and non-aliasing. So e.g. the range $(-\frac{3}{4}, -\frac{1}{4}) \cup (\frac{1}{4}, \frac{3}{4})$ won't do, as the positive and negative halves of the range will alias to each other, while the range $(-1, -\frac{1}{2}) \cup (\frac{1}{2}, 1)$ is okay.

Restoration from sampled signal

In case $\hat{x}[n]$ is obtained by sampling a bandlimited function $x(t)$ (that is $x(t)$ denotes the already bandlimited function, so that $x(t) = \bar{x}(t)$), the formula (8.235) can suggest a simple way to convert $\hat{x}[n]$ back to $x(t)$. Since the spectrum $X(f)$ is restricted to the range $(-\frac{1}{2}, \frac{1}{2})$ anyway and is zero everywhere else, this means that $X(f) = \hat{X}(f)$ $\forall f \in (-\frac{1}{2}, \frac{1}{2})$ and we can formally recover $X(f)$ from $\hat{X}(f)$ by multiplying $\hat{X}(f)$ with the rectangular function

$$X(f) = \hat{X}(f) \operatorname{rect}(f) \qquad \text{if } X(f) = 0 \ \forall f \notin (-\tfrac{1}{2}, \tfrac{1}{2}) \tag{8.236}$$

that is, effectively we want to bandlimit the spectrum once again. By the convolution property we can do this by convolving $\hat{x}(t)$ with the sinc function:

$$\begin{aligned} x(t) = \hat{x}(t) * \operatorname{sinc} \pi t &= \left(\sum_{n=-\infty}^{\infty} x(t)\delta(t-n) \right) * \operatorname{sinc} \pi t \\ &= \sum_{n=-\infty}^{\infty} \big(x(t)\delta(t-n) * \operatorname{sinc} \pi t \big) \\ &= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} x(\tau)\delta(\tau - n) \cdot \operatorname{sinc} \pi(t-\tau) \, \mathrm{d}\tau \end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \big(x(\tau) \cdot \operatorname{sinc} \pi(t-\tau)\big) \delta(\tau - n) \, \mathrm{d}\tau \\
&= \sum_{n=-\infty}^{\infty} \big(x(n) \cdot \operatorname{sinc} \pi(t-n)\big) \\
&= \sum_{n=-\infty}^{\infty} \big(\hat{x}[n] \cdot \operatorname{sinc} \pi(t-n)\big) = \hat{x}[n] * \operatorname{sinc} \pi t
\end{aligned}
$$

where we effectively have introduced the idea of a convolution of a discrete-domain function $\hat{x}[n]$ with a continuous-domain function $\operatorname{sinc} \pi t$. Thus

(restoration of continuous-time signal from its sampled version, provided the original signal is bandlimited)

$$
x(t) = \hat{x}(t) * \operatorname{sinc} \pi t = \hat{x}[n] * \operatorname{sinc} \pi t \tag{8.237}
$$

▷ Notably, piecewise-polynomial interpolations of $\hat{x}[n]$ often well approximate the right-hand side of (8.237), the higher the interpolation order the better the approximation. Such interpolations can be written as a mixed continuous/discrete convolution with some function $g(t)$:

$$
x(t) \approx \hat{x}[n] * g(t)
$$

where $g(t)$ depends on the interpolation scheme (and is usually compactly supported). At higher interpolation orders $g(t) \approx \operatorname{sinc} \pi t$ for interpolation schemes such as Lagrange interpolation.

For that reason the restoration of the original function $x(t)$ from its sampled version $\hat{x}[n]$ by (8.237) is often referred to as *sinc interpolation*.

Sampling frequency

We have been assuming unit sampling period $T = 1$ during our discussion. Switching to the general case of $T \neq 1$ means that DTFT is no longer 1-periodic but $1/T$-periodic. Respectively the $(-\frac{1}{2}, \frac{1}{2})$ f-domain interval turns into $(-\frac{1}{2T}, \frac{1}{2T})$. Introducing the *sampling frequency* $f_s = 1/T$, the interval $(-\frac{1}{2T}, \frac{1}{2T})$ can be rewritten as $(-\frac{f_s}{2}, \frac{f_s}{2})$. The frequency $f_s/2$, determining the upper bound on the spectrum range, is known as *Nyquist frequency*.

In this case $\operatorname{rect}(f)$ turns into $\operatorname{rect}(f/f_s)$ and respectively $\operatorname{sinc} \pi t$ turns into $f_s \operatorname{sinc} \pi f_s t$.

▷ Practically the bandlimiting of $x(t)$ needs to be done in the real world, where we don't have a physical process which performs convolution. Instead some other physical processes are employed which perform some "soft bandlimiting", where the spectrum, instead of (effectively) being multiplied with $\operatorname{rect}(f)$, is multiplied with some other smooth function, slowly rolling off towards $f \to \infty$. The result can be considered "approximately bandlimited" to some relatively high frequency bound.

This bound requires a pretty high sampling frequency, so another bandlimiting step in then performed digitally in the discrete domain, after the initial sampling of the "soft-bandlimited" signal. In the discrete digital domain we however still cannot compute a convolution with sinc in finite time, since sinc has infinitely large support.

Instead, convolution with truncated or windowed sinc will be done (spectral domain effects of truncation or windowing will be discussed in Subsection 8.19.2), which makes this second bandlimiting step also approximate.

The same in reverse applies to the restoration of the original (bandlimited) $x(t)$ from $\hat{x}[n]$.

8.19.2 Domain truncation

In Subsection 8.19.1 we converted a continuous-domain function $x(t)$ into a discrete-domain function $x[n]$. Effectively we had to truncate the spectrum of $x(t)$ (by bandlimiting) in order to be able to turn $x(t)$ into a function of a discrete argument without loss of information (or we could say that we thereby choose exactly what information is going to be lost). However the domain of $x[n]$ still might be infinite. Actually, it *will* be infinite, since, by the discussion in Subsection 8.16.7, a bandlimited function of a continuous argument (which thereby has compactly supported spectrum) must be infinitely long, and so must be its sampled version.[10] We therefore now need to truncate the domain of $x[n]$ to arrive at a function consisting of a finite number of discrete samples.

It would be more instructive and intuitive to study the effect of domain truncation for continuous-argument functions $x(t)$. The extension of the results to discrete-argument functions $x[n]$ can be then done by the application of the material of Subsection 8.19.1. We shall discuss this extension later in Subsection 8.19.7.

▷ The domain of $x(t)$ or $x[n]$ will usually have time semantics. The process of truncating the domain of x in this case is referred to as *time-limiting*, similarly to how truncation of the spectral domain was referred to as *band-limiting*.

So let's take a continuous-argument $x(t)$ defined on the entire real axis and truncate its domain to $[-T/2, T/2]$, obtaining a truncated-domain function $\bar{x}(t)$ defined just on $[-T/2, T/2]$ (notice that T here is not the sampling period but simply the length of the truncation result).

▷ Without loss of generality we can assume the truncated domain range to be symmetric for simplicity. The results can be later generalized to non-symmetric ranges by using the time-shift property of Fourier transform, if necessary.

Attempting to apply the direct Fourier transform formula (8.70a) to such $\bar{x}(t)$, we won't be able to integrate over the entire real axis (as the function is formally undefined outside of $[-T/2, T/2]$) and will have to do with integrating just over $[-T/2, T/2]$ instead:

$$\bar{X}(f) = \int_{-T/2}^{T/2} \bar{x}(t) e^{-2\pi i f t} \, \mathrm{d}t \tag{8.238}$$

▷ Notice that (8.238) looks essentially identical to the density version of Fourier series coefficients formula (8.68a), if we replace f in (8.238) by kF, where $F = 1/T$ is the base frequency of a T-periodic signal.

[10] Unless the function's zeros happen to exactly coincide with the sampling points, which is rather improbable.

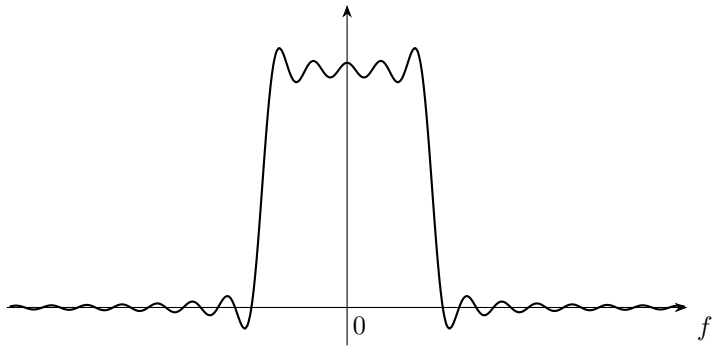


Figure 8.23: Spectrum of a time-limited sinc.

Limiting the integration range to $[-T/2, T/2]$ can be alternatively seen as forcing $x(t) = 0\ \forall t \notin [-T/2, T/2]$, which can be achieved by letting

$$\bar{x}(t) = \operatorname{rect}\left(\frac{t}{T}\right) x(t) \tag{8.239}$$

in which case we can compute the direct Fourier transform of $\bar{x}(t)$ in the usual way by integrating over the entire real axis in (8.70a):

$$\begin{aligned}\bar{X}(f) &= \int_{-\infty}^{\infty} \bar{x}(t) e^{-2\pi i f t}\,\mathrm{d}t = \int_{-\infty}^{\infty} \operatorname{rect}\left(\frac{t}{T}\right) x(t) e^{-2\pi i f t}\,\mathrm{d}t \\ &= \int_{-T/2}^{T/2} x(t) e^{-2\pi i f t}\,\mathrm{d}t\end{aligned}$$

▷ In case t has time semantics, this kind of domain truncation or zeroing of the function outside of some interval is referred to as *timelimiting* of $x(t)$.

Multiplication of $x(t)$ by $\operatorname{rect}(t/T)$ produces two jump discontinuities at $t = \pm T/2$ in $\bar{x}(t)$. By the discussion in Subsection 8.16.9, these discontinuities will make two contributions to the spectrum of $\bar{x}(t)$: they will make the spectrum roll off as $1/f$ (or slower), and the spectrum (as a function of f) will contain sinusoidal ripples of frequency T. If the spectrum was originally a slowly changing and quickly decaying function, this change could be quite drastic. E.g. by truncating the domain of a sinc function to $[-T/2, T/2]$, its spectrum, previously being equal to a rectangular function, attains sinusoidal ripples of frequency $T/2$, decaying as $1/f$ at $f \to \infty$ (Fig. 8.23).

To obtain the analytic expression for the function in Fig. 8.23, we could apply the convolution property to

$$\bar{x}(t) = \operatorname{rect}(t/T) \cdot \operatorname{sinc} \pi t$$

By (8.206), Fourier transform of $\operatorname{sinc} \pi t$ is $\operatorname{rect}(f)$ and, by its dual, Fourier transform of $\operatorname{rect}(t/T)$ is

$$x(t) = \operatorname{rect}(t/T) \iff X(f) = T \operatorname{sinc}(-\pi f T) = T \operatorname{sinc} \pi f T \tag{8.240}$$

Therefore

$$\bar{X}(t) = T \operatorname{sinc} \pi f T * \operatorname{rect}(f) = T \int_{-\infty}^{\infty} \big(\operatorname{sinc} \pi (f - \nu) T\big) \operatorname{rect} \nu\,\mathrm{d}\nu$$

$$
\begin{aligned}
&= T\int_{[-\frac{1}{2},\frac{1}{2}]} \big(\operatorname{sinc}\pi(f-\nu)T\big)\,\mathrm{d}\nu = T\int_{[f-\frac{1}{2},f+\frac{1}{2}]} \operatorname{sinc}\pi\nu T\,\mathrm{d}\nu \\
&= T\int_{f-\frac{1}{2}}^{f+\frac{1}{2}} \operatorname{sinc}\pi\nu T\,\mathrm{d}\nu = \frac{T}{\pi T}\cdot\big(\operatorname{Si}\pi(f+\tfrac{1}{2})T - \operatorname{Si}\pi(f-\tfrac{1}{2})T\big) \\
&= \frac{1}{\pi}\cdot\big(\operatorname{Si}\pi(f+\tfrac{1}{2})T - \operatorname{Si}\pi(f-\tfrac{1}{2})T\big) \qquad (8.241)
\end{aligned}
$$

As we might remember from the earlier discussion, the integral sine function $\operatorname{Si} f$ has ripples of unit cyclic frequency (or of ordinary frequency equal to $1/2\pi$) decaying as $1/f$. This is exactly what we should have expected due to the two jump discontinuities appearing in the time-limited sinc at $t = \pm T/2$, as the frequency of the ripples in $\operatorname{Si}\pi fT$ is πT times larger than the one in $\operatorname{Si} f$, being thus equal to $\frac{1}{2\pi}\pi T = T/2$.

What we have just constructed is the simplest possible kind of *short-time Fourier transform* or shortly *STFT*. As we have seen so far, the spectrum obtained by our example STFT seems to have rather strong undesired artifacts, and it also seems that it's a general problem with the domain truncation approach. Maybe we can further improve the details of the STFT technique?

8.19.3 Windowing

To introduce some degrees of freedom into the domain truncation process we can replace $\operatorname{rect}(t/T)$ in (8.239) by some generic function $w(t)$, so that (8.239) becomes

(windowing)

$$\tilde{x}(t) = w(t)x(t) \qquad (8.242a)$$

The multiplication of $x(t)$ by $w(t)$ in (8.242a) is referred to as *windowing*, respectively the function $w(t)$ is referred to as the *window function* or simply *window*. We would be interested in whether there are better (whatever that means) window functions than the *rectangular window* function $w(t) = \operatorname{rect}(t/T)$ that we have used so far.

▷ Notice that, provided $w(t) = 0\ \forall t \notin [-T/2, T/2]$, we still can use the formula (8.238) to compute the spectrum of the windowed function $\tilde{x}(t)$, instead of computing its direct Fourier transform by integration over the entire real axis. If it hadn't been the case, this would have pretty much defeated our original intention, which was exactly to be able to compute direct Fourier transform by integrating over a finite range.

The effects of windowing on the spectrum of $x(t)$ are obviously given by the convolution property:

(spectral effect of windowing)

$$\tilde{X}(f) = W(f) * X(f) \qquad (8.242b)$$

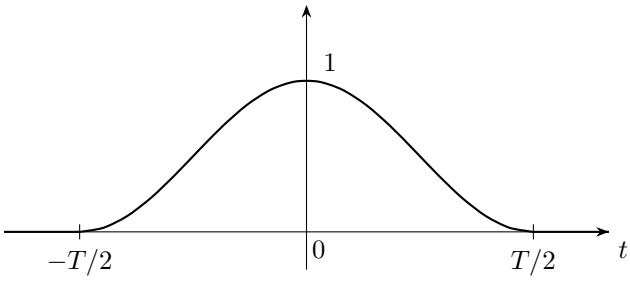


Figure 8.24: Raised-cosine window.

where $W(f)$ and $X(f)$ are Fourier transforms of $w(t)$ and $x(t)$. E.g. (8.241) was obtained by the convolution of $X(f)$ with $W(f) = T \operatorname{sinc} \pi f T$. The ripples present in the sinc function contained in $W(f)$ exposed themselves in the convolution result. Therefore we might be looking for such window functions $w(t)$ that the ripples in their spectra are not too prominent, where we might be paying attention to both the amplitude of the ripples as well as their decay speed.

We know that the primary factor affecting the decay rate of the ripples in the spectrum is the continuity properties of the original function. Therefore we might be interested in $w(t)$ which do not have jump discontinuities, and maybe also no jump discontinuities in the 1st derivative, or even higher-order derivatives. One possible window which is continuous up to the 1st derivative inclusively is the *raised-cosine window*

$$w(t) = \begin{cases} \frac{1}{2}\left(1 + \cos 2\pi \frac{t}{T}\right) & \text{if } |t| < T/2 \\ 0 & \text{otherwise} \end{cases}$$

(Fig. 8.24).

Computing the direct Fourier transform for the unit-length ($T = 1$) raised-cosine window yields

$$\begin{aligned} W(t) &= \frac{1}{2}\int_{-1/2}^{1/2} (1 + \cos 2\pi t) e^{-2\pi i f t} \mathrm{d}t \\ &= \frac{1}{2}\int_{-1/2}^{1/2} \left(1 + \frac{e^{2\pi i t}}{2} + \frac{e^{-2\pi i t}}{2}\right) e^{-2\pi i f t} \mathrm{d}t \\ &= \frac{1}{2}\int_{-1/2}^{1/2} e^{-2\pi i f t} \mathrm{d}t + \frac{1}{4}\int_{-1/2}^{1/2} e^{2\pi i t} e^{-2\pi i f t} \mathrm{d}t + \frac{1}{4}\int_{-1/2}^{1/2} e^{-2\pi i t} e^{-2\pi i f t} \mathrm{d}t \end{aligned}$$

The first of the 3 integrals is the direct Fourier transform of $\operatorname{rect}(t)$, which by (8.240) is $\operatorname{sinc} \pi f$. The second and third integrals can be computed by using the frequency shift property, which suggests that the spectra given by the 2nd and the 3rd integral are frequency-shifted versions of the spectrum given by the 1st integral, the shift amount being equal to 1. Thereby

$$W(t) = \frac{1}{2} \operatorname{sinc} \pi f + \frac{1}{4} \operatorname{sinc} \pi (f - 1) + \frac{1}{4} \operatorname{sinc} \pi (f + 1)$$

$$
\begin{aligned}
&= \frac{1}{2}\frac{\sin \pi f}{\pi f} + \frac{1}{4}\frac{\sin \pi(f-1)}{\pi(f-1)} + \frac{1}{4}\frac{\sin \pi(f+1)}{\pi(f+1)} \\
&= \frac{1}{2\pi}\sin \pi f \cdot \left(\frac{1}{f} - \frac{1}{2}\frac{1}{f-1} - \frac{1}{2}\frac{1}{f+1}\right) \\
&= \frac{1}{2\pi}\sin \pi f \cdot \frac{2(f-1)(f+1) - f(f+1) - f(f-1)}{f(f-1)(f+1)} \\
&= \frac{1}{2\pi}\sin \pi f \cdot \frac{2f^2 - 2 - f^2 - f - f^2 + f}{f(f^2-1)} \\
&= \frac{1}{2\pi}\sin \pi f \cdot \frac{-2}{f(f^2-1)}
\end{aligned}
$$

The spectrum, as expected, rolls off as $1/f^3$.

Other examples of commonly used window functions, as well as further considerations about the factors affecting the quality of a window can be found in signal processing literature.

8.19.4 Window width and height

Given a window function $w(t)$ we can vary the window width by using $w(t/T)$ instead of $w(t)$. In this case the window width (which is also referred to as *window length*) grows proportionally to T. The larger is T, the bigger part of $x(t)$ is included into the computation of the spectrum of $x(t)$.

Apparently, we never compute the true spectrum of $x(t)$, always computing the spectrum of the windowed function $\tilde{x}(t) = w(t)x(t)$ instead, which, by the convolution property, is equal to $W(f) * X(f)$. That is the true spectrum $X(f)$ is being "smeared" by the convolution with $W(f)$. To reduce that smearing we need to make $W(f)$ more narrow, which, by the domain scaling property (8.208) of Fourier transform, can be achieved by making the window $w(t)$ itself wider.

Intuitively, this also makes sense. The wider is $w(t)$, the more of $x(t)$ we include into the computation of the spectrum, therefore the more precise the computed spectrum becomes. Here we again observe a reciprocal duality between the two domains of Fourier transform, similar to the reciprocal duality observed in the domain scaling property (8.208) as well as in the Heisenberg uncertainty relationship inequality (8.225). In the case of windowing this reciprocal relationship says: the wider the window, the finer the frequency resolution of the computed spectrum (and vice versa).

> ▷ This reciprocal relationship is also one of the factors (if not the main one) creating a difficulty in finding a good window shape. Suppose we try to increase the decay rate of the ripples in $W(f)$ by making $w(t)$ "more continuous" at the ends of the window. This however will make the window shape more concentrated towards $t = 0$. E.g. the raised cosine window is stronger concentrated around $t = 0$ compared to the rectangular window. Thus, while we improve the ripple decay speed, we might somewhat degrade the frequency resolution of the obtained spectrum.

So, the longer the window, the better. However, what about the window height? Do we need to scale it? And what should the window height be to begin with?

One simple consideration for the window height is that the area under $W(f)$ should be unity, so that upon convolution with $W(f)$ (which can be seen as computing a "weighted average", the values $W(f)\,\mathrm{d}f$ being the weights) the values of $X(f)$ kind of retain their "average magnitude". This particularly makes sense if $w(t)$ is a real even function (which is typically the case), therefore so is $W(f)$. By the "constant term" property, the area under $W(f)$ is simply equal to $w(0)$, therefore we obtain the condition $w(0) = 1$.

Therefore, at the first look, as we vary the window length, we shouldn't change the window height. This conclusion is, however, deceiving. In obtaining this conclusion we have assumed that the "true" $X(f)$ corresponding to the unwindowed $x(t)$ is the "ultimate truth" and our point of reference. We should however consider that in practice our window lengths will be much smaller than the "full length" of $x(t)$. If the magnitude of $x(t)$ doesn't vary too much over the range where we consider to apply the window, for all practical purposes we can also say that $x(t)$ doesn't decay to zero as $t \to \infty$. We can consider $x(t)$ as a non-decaying infinitely long function, pretty much as if it was a periodic function. At any rate we can assume that the average power of $x(t)$ (averaged over some sufficiently large interval, but not too large, this interval should be still smaller than our typical window lengths) doesn't vary much across the t axis.

Under this assumption of an approximately constant average, the total energy of the windowed $x(t)$ will grow approximately proportionally to the window length, and, by Parseval's theorem, so will the total energy of the spectrum $\tilde{X}(f)$ obtained for the windowed function $\tilde{x}(t) = x(t)w(t)$, that is the area under $|\tilde{X}(f)|$ should grow. This generally means that either the width of $\tilde{X}(f)$ or the height of $\tilde{X}(f)$ (without being particularly strict what the "width" and "height" mean) must also grow proportionally to the same factor. Or this growth could be distributed somehow between the width and the height, so both would grow.

Now it looks highly unlikely that the width of $\tilde{X}(f)$ would grow. Because if the width grew, it would mean that more and more new high-frequency components appear in $x(t)$ as the window length grows, and that doesn't look very realistic. Indeed, by (8.218b),

$$\sigma_{\tilde{X}} = \frac{\|\tilde{x}'\|}{\|\tilde{x}\|} - M_{\tilde{x}}^2$$

(where we wrote the normalization by $\|\tilde{x}\|$ out explicitly). Now since the average power of $x(t)$ is assumed to be approximately constant, the value $\|\tilde{x}\|$ should grow approximately proportionally to the window length. Also, assuming the average power of $x'(t)$ is approximately constant, the value $\|\tilde{x}'\|$ should also grow approximately proportionally to the window length. Also, assuming $w(t)$ is real even (or at least symmetric with respect to the origin), $M_{\tilde{x}}$ is likely to be close to zero. It follows that $\sigma_{\tilde{X}}$ must be approximately constant as the window length grows. Thus, it's more likely that the energy growth of $\tilde{x}(t)$ will manifest solely in growth of the "height" of $\tilde{X}(f)$, that is the magnitudes of $\tilde{X}(f)$ would grow.

This growth of the spectral values doesn't feel very intuitive. As we increase the window length, we intuitively expect to attain a better frequency resolution in the obtained spectrum, not to obtain a very different (if only by the absolute magnitude) spectrum.

The reason for this problem is that the "true" Fourier transform $X(f)$, which we wanted to use as our reference, is "designed" to handle functions of finite energy. However in our current situation, even if $x(t)$ in reality has a finite energy, most of this energy is contained outside of our window, therefore for all practical purposes $x(t)$ behaves as if it had infinite energy (but constant average power), pretty much like periodic functions do. This suggests that we should additionally divide $\tilde{X}(f)$ by the square root of the window length, thereby keeping the energy approximately constant. This additional factor of $1/\sqrt{T}$ (where T is the window length) can be incorporated into any stage of STFT. One possibility is to incorporate it into the window function, suggesting that the window height must fall off as $1/\sqrt{T}$ as the window length grows. That is, instead of requiring the area under $W(f)$ to stay constant and equal to 1, we require the total energy of the window to stay constant. For definiteness, we might just require the window to have unit energy: $\|w\| = \|W\| = 1$. In this case the total energy of $\tilde{x}(t)$ is

$$\|\tilde{x}\|^2 = \|x(t)w(t)\|^2 = \int_{-\infty}^{\infty} |x(t)w(t)|^2 \cdot \mathrm{d}t = \int_{-\infty}^{\infty} |x(t)|^2 \cdot |w(t)|^2 \cdot \mathrm{d}t$$

that it is a weighted average of $x(t)$, where $|w(t)|^2$ is the normalized weighting function. This means that $\|\tilde{x}\|^2$ will be simply the average power of $x(t)$, the averaging being done by $|w(t)|^2$.

Thinking of it, recall that we assumed that $x(t)$, instead of having a finite energy (which is a typical Fourier transform setup), has approximately constant average power, not unlike periodic signals do. In this case it only makes sense that the spectra of $x(t)$, that we obtain, should use the average power as the reference value for their normalization.

▷ Instead of including the $1/\sqrt{T}$ factor into the window, one other option would be to include it into the transform formula (8.238), which thereby attains the form

$$\tilde{X}(f) = \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} \tilde{x}(t) e^{-2\pi i f t} \, \mathrm{d}t \tag{8.243}$$

which looks like Fourier series coefficient computation formula with symmetric normalization placement.

▷ If $x(t)$ is truly periodic, rather than simply having approximately constant average power, then, as the window length T grows, the spectrum of $\tilde{x}(t)$ will consist of more and more narrow and high peaks, occurring at the multiples of the base frequency of $x(t)$, corresponding to the fact that by (8.172) the spectrum of a periodic $x(t)$ consists of regularly spaced Dirac delta functions. Respectively, $\tilde{X}(f)$ becomes "infinitely high", despite the normalization by $1/\sqrt{T}$. In this case we actually might want to normalize by $1/T$, as it is commonly the case with Fourier series (the latter being designed explicitly for handling periodic signals).

Thus, there is no perfect normalization which handles all cases. We will return to this question once more in in the discussion of stochastic Fourier transform in Section 8.20.

8.19.5 Spectral-domain sampling

Assuming the window function $w(t)$ is limited to $t \in [-T/2, T/2]$, we can use integration over the finite range (8.238) to obtain direct Fourier transform. The

right-hand side of (8.238) looks essentially identical to the Fourier series coefficients formula in its density version (8.68a). Matching the left- and right-hand sides of (8.238) and (8.68a), we obtain, by equating their matched left-hand sides,

$$\tilde{X}(k/T) = \tilde{X}[k]$$

This suggests that we kind of do not need to compute the spectrum $\tilde{X}(f)$ of $\tilde{x}(t)$ for all values of f. Computing these values just at $f = kF = k/T$ (where $F = 1/T$ is the base frequency corresponding to the period of T) should suffice. Indeed, having obtained the values $\tilde{X}(k/T)$, we can apply the Fourier series formula (8.68b) to convert $\tilde{X}[k] = \tilde{X}(k/T)$ back to $\tilde{x}(t)$. In that sense all information about $\tilde{x}(t)$ is contained within the values $\tilde{X}(k/T)$. Except that $\tilde{x}(t)$ given by the Fourier series will be T-periodic, but we can handle that by explicitly forcing the result to zero outside of the primary period $[-T/2, T/2]$.

Okay, so we don't really need the values of $\tilde{X}(f)$ for all f and can do with just $\tilde{X}(k/T)$. But what if we still do want to compute the values $\tilde{X}(f)$ for all f? Of course, nothing is stopping us from directly computing them by (8.238), but it could be instructive to find the relationship between $\tilde{X}[k] = \tilde{X}(k/T)$ and $\tilde{X}(f)$.

In this regard we can notice the duality of our current situation to the sampling of a bandlimited function discussed in Subsection 8.19.1. Instead of a bandlimited function in in Subsection 8.19.1 we are now having a timelimited function, and instead of t-domain sampling we are now effectively doing f-domain sampling:

$$\tilde{X}[k] = \tilde{X}(k/T)$$

Respectively the discrete spectrum $\tilde{X}[k]$ corresponds to a periodically extended $\tilde{x}(t)$, as given by the Fourier series, while continuous spectrum $\tilde{X}(f)$ corresponds to a non-periodic $\tilde{x}(t)$. We can see this non-periodic $\tilde{x}(t)$ as having been obtained from the periodically extended $\tilde{x}(t)$ by erasing all its repetitions outside of the primary period $[-T/2, T/2]$. Summing up: by going from $\tilde{X}(f)$ to $\tilde{X}[k]$ we periodically extend $\tilde{x}(t)$, and by going back from $\tilde{X}[k]$ to $\tilde{X}(f)$ we erase all period repetitions of $\tilde{x}(t)$, except the primary period.

This suggests that the relationship between $\tilde{X}(f)$ and $\tilde{X}[k]$ is pretty much the same which we had between the bandlimited function $x(t)$ and its sampled version $\hat{x}[n]$ in Subsection 8.19.1, which was given by (8.237). So we should obtain a pretty much the same relationship here. Temporarily letting $T = 1$ so $\tilde{X}[k] = \tilde{X}(k)$, we obtain

$$\tilde{X}(f) = \tilde{X}[k] * \operatorname{sinc} \pi f$$

If $T \neq 1$, then by the scaling property (8.208)

$$\frac{1}{T}\tilde{X}(k/T) = \tilde{X}[k] * \operatorname{sinc} \pi f T$$

or

$$\tilde{X}(k/T) = \tilde{X}[k] * T \operatorname{sinc} \pi f T \tag{8.244}$$

Thus $\tilde{X}(k/T)$ can be simply obtained by interpolation of $\tilde{X}[k]$. This corresponds to multiplication of the T-periodic version of $\tilde{x}(t)$ by a rectangular function $\operatorname{rect}(t/T)$ (thereby making $\tilde{x}(t)$ timelimited and nonperiodic), which in the spectral domain looks like a convolution with sinc.

8.19.6 Localization of frequency

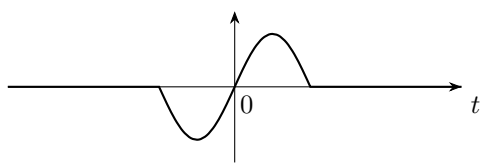
At this point it would be instructive to discuss the distinction between the two different ideas of frequency occurring in connection with Fourier transform. One is the frequency of the spectral components, that is the frequency of sinusoidal components into which Fourier transform expanded the function. The other one is instantaneous frequency. The distinction between the two is probably best illustrated by a time-limited sinusoidal signal of frequency F:

$$x(t) = \operatorname{rect}\left(\frac{t}{T}\right) e^{2\pi i F t} \tag{8.245a}$$

▷ It might have been more intuitive to think of a time-limited purely real sinusoid

$$x(t) = \operatorname{rect}\left(\frac{t}{T}\right) \sin 2\pi F t$$

which could look e.g. as



But then we'd have needed to talk about two complex sinusoids at frequencies $\pm F$ making up that real sinusoid. So, for the sake of simplicity, we discuss the truncated complex sinusoid instead.

Intuitively, we probably think that the frequency of $x(t)$ is F. However let's look at the spectrum of $x(t)$. There are different ways how it can be computed, probably the easiest is to apply the frequency-shift property to the Fourier transform of $\operatorname{rect}(t/T)$, which by (8.240) is $T \operatorname{sinc} \pi f T$, thereby yielding

$$X(f) = T \operatorname{sinc} \pi (f - F) T \tag{8.245b}$$

So instead of a single-frequency spectrum $X(f) = \delta(f - F)$, which would have been the case for $x(t) = e^{2\pi i F t}$, we have obtained its smeared version (8.245b), containing components of all different frequencies, although most prominently those in the vicinity of $f = F$.

Apparently a similar situation would have occurred if we used a smoother window function $w(t)$ instead of truncation by $w(t) = \operatorname{rect}(t/T)$. In any case, we have

$$X(f) = W(f) * \delta(f - F) = W(f - F)$$

That is we obtain the spectrum of $w(t)$ which has been frequency-shifted to $f = F$.

Thus, the wider is $w(t)$, the more the spectrum of $X(f)$ is concentrated around the "true frequency" $f = F$. We have already mentioned that the frequency resolution of STFT is improving as the window length grows and what we observe here is exactly a manifestation of that effect. This suggests that it's not possible to determine a frequency of a signal from looking into some local interval of that signal, we always need to consider the signal on the entire

range $t \in (-\infty, \infty)$. By looking only at a local range of t we can determine the signal's frequency only to a limited precision, the precision growing reciprocally to the length of the range.

▷ The latter statement might feel pretty counterintuitive. Indeed, by looking at (8.245a) (or at its purely real counterpart $\mathrm{rect}(t/T)\sin 2\pi F t$) can't we immediately recognize that the function is simply a truncated sinusoid of frequency F? We could do that indeed, but that includes some additional assumptions.

If we are only given the values of the function (8.245a) and not its analytic expression, we may be able to guess that the function has been obtained by truncating some possibly infinitely long function, but this infinitely long function could have been anything. It didn't have to continue indefinitely as $e^{2\pi i F t}$. In a way, Fourier transform doesn't take any assumptions about "how the function looks outside of the truncation range".

Furthermore, if e.g. we are given a signal which is simultaneously amplitude- and frequency-modulated, e.g.

$$x(t) = a(t)\cos\left(\varphi_0 + \int_0^t f(\tau)\,\mathrm{d}\tau\right) \tag{8.246}$$

for quite a few shapes of $x(t)$ we won't even be able to tell between frequency and amplitude modulation, the same shape of $x(t)$ could have been achieved by different combined choices of $a(t)$ and $f(t)$. Thus, the potentially varying *instantaneous frequency* $f(t)$ in (8.246) shouldn't be confused with constant frequencies of sinusoidal partials of Fourier transform.

Thus we can say that "frequency cannot be localized to a finite range of t". In a way this is another manifestation of the reciprocal duality of Fourier transform domains. The impossibility of frequency localization also follows from other manifestations of this reciprocal duality. E.g. a function $x(t)$ which has a single determined frequency F has Fourier transform $X(f) = \delta(f - F)$, therefore $\sigma_X = 0$, therefore, by Heisenberg uncertainty relationship (8.225), $\sigma_x = \infty$, which means $x(t)$ cannot have compact support. We also know that by discussion in Subsection 8.16.7, a function cannot simultaneously have compact support in both domains.

8.19.7 Discrete-domain windowing

We have discussed the short-time Fourier transform in application to continuous-domain functions. The same techniques are extended in a straightforward manner to discrete-domain functions. E.g. instead of sampling an infinitely long function $x(t)$ we can first time-limit it (by domain truncation or windowing), and apply sampling to the timelimited version of $x(t)$. Or we could do vice versa, first sampling it and then windowing.

The practical problem with the first approach is that applying windowing to $x(t)$ would need to be done in the real world and it might be impossible to implement windowing as a physical process. It is therefore simpler to do the opposite, where we would obtain sufficiently many samples $\hat{x}[n]$ to cover the entire width of the discrete-domain window that we are going to apply as a second step to our already sampled signal.

The discrete-domain windows are usually obtained by simply sampling continuous domain window functions as they are: $w[n] = w(n)$. Even though window functions are usually not bandlimited (by design, window functions are

usually timelimited, and we have found in Subsection 8.16.7 that simultaneous bandlimiting and timelimiting is impossible), these sampled versions typically work well.

Once the windowing has been applied in the discrete domain, we can replace the infinite sum in the direct DTFT formula (8.58a) by a finite one, by dropping the terms outside of the windowing interval:

$$\tilde{X}(f) = \sum_{n=-\infty}^{\infty} \tilde{x}[n]e^{-2\pi i f n} = \sum_{n=-M}^{M} \tilde{x}[n]e^{-2\pi i f n} \tag{8.247}$$

where $\tilde{x}[n] = \hat{x}[n]w[n]$ is the windowed function (where $\hat{x}[n]$ is the sampled function) and M is the half-length of the window (assuming the window is symmetric, which is usually the case).

The right-hand side of (8.247) however looks pretty much identical to the density version of DFT (8.23a) if we let $N = 2M + 1$ and $f = k/N$. Especially if we also recall that the summation in (8.23a) can be done over any discrete interval of length N. This suggests that instead of (8.247) we could simply use the DFT (8.23a).

Thus, starting with an infinitely long signal $x(t)$ we bandlimited it, which allowed to represent it in a discrete time domain, and we also timelimited it, which allowed its frequency-domain representation to become discrete as well. The discreteness of one domain, however, automatically implies periodicity in the dual domain. Thus the representation of x becomes discrete and periodic in both domains and we can apply DFT formulas. The periodicity in both domains is however only a "technical periodicity", in reality the representation in each of the domains is considered to be zero outside of the symmetric period at the respective domain's origin.

Conventional usage of the term DFT

Since the introduction of DFT in Section 8.2 we were using the terms DFS and DFT interchangeably. Now we have reached the point where we have the intended application context of DFT: we use DFS transformations, but pretend that the representations in both domains are not periodic but are zero outside of the primary period. This is what *DFT* commonly refers to.

▷ By the formulas such as (8.21), DFT has quadratic computation time complexity, since computing each of N different result values requires a proportional to N number of operations. A family of numerical algorithms commonly known as *fast Fourier transform*, or shortly *FFT*, is designed to have linear-logarithmic computation time. These algorithms produce the same result as DFT, just faster. We won't discuss the details of FFT algorithms in this book.

▷ By the previous discussion in Subsection 8.19.5, we should probably use the symmetric DFT version (8.22a), if we want to achieve an approximately constant "spectrum height", independently of the window width. This however can fail if $x(t)$ is truly periodic, where an amplitude DFT version (8.22a) might work better.

We have found in Subsection 8.19.5 that the continuous spectrum $X(f)$ can be restored from the discrete spectrum $X[k]$ by interpolation. The discussion in

Subsection 8.19.5 was however assuming a non-periodic spectrum. At the first look it might seem that we'd need to adapt our interpolation to the periodic spectrum case, since DFT technically assumes a periodic spectrum, so that (8.244) would turn into something like a periodic convolution with a Dirichlet kernel function. Upon further consideration we should however realize that the spectrum periodicity is only technical. In reality we are approximating a nonperiodic spectrum which is zero outside of the primary period. This suggests that we might want to use the non-periodic interpolation techniques, also taking the values of $\tilde{x}[n]$ outside of the primary period to be zero.

8.20 Stochastic Fourier transform

In this section we would like to try to apply Fourier transform techniques to random processes.

> ▷ A common practical application of this is analysis of noise signals. In this case it doesn't make sense to define a signal by specifying its values at all discrete points (or specifying a way to compute these values in a deterministic repeatable way, e.g. $x(t) = \sin t$), as we simply do not care about the specific values. However we do care about stochastic properties of the signal and this is what we specify.

8.20.1 Discrete Fourier transform

As usual, we shall start with discrete Fourier transform and then progress to other transform types by taking the limit. So let $x[n]$ be a random process rather than a deterministic function. In order to be able to apply discrete Fourier transform techniques to $x[n]$, its realizations (which we will also denote as $x[n]$) must be N-periodic, so this is a restriction which we shall impose.

If $x[n]$ is fully random (besides being N-periodic), we probably won't be able to tell much about this process. We need to specify some stochastic properties of this process and then find out how these properties affect the spectrum $X[n]$ (which, as usual, would be given by one of (8.21), (8.22), (8.23)). Intuitively, the first thing which springs to mind is specifying the probability distribution density $p(x)$.

> ▷ Notice that by specifying one and the same probability density $p(x)$ regardless of n we effectively assumed that $x[n]$ is 1st order stationary.

To simplify the discussion, without loss of generality we can additionally assume that the expected value of $x[n]$ is zero, that is the process is centered: $\langle x[n] \rangle = 0 \; \forall n$, where $\langle \cdot \rangle$ denotes the expectation. Indeed, if it's not centered, we can simply separate the non-zero value of $\langle x[n] \rangle$ into a standalone deterministic term and analyze it separately (where we already know that DFT of such term is a periodic Kronecker delta). Fig. 8.25 shows an example of a periodic centered 1st order stationary process.

Now let $X[k]$ be the discrete Fourier transform of $x[n]$, obtained by applying the amplitude version transform (8.21a) to realizations of $x[n]$. Since $x[n]$ is a random process, so is $X[k]$. By (8.21a), each realization of $X[k]$ is a linear combination of $x[n]$ with complex linear combination coefficients $e^{-2\pi ikn/N}/N$.

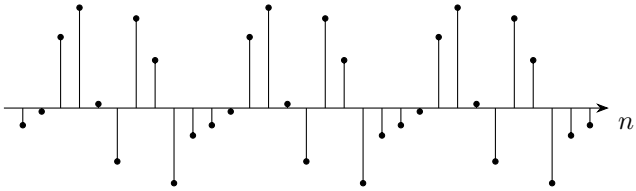


Figure 8.25: A periodic centered random process.

The expectation of a sum is the sum of the expectations, and the expectation of each term of the linear combination is

$$\left\langle \frac{1}{N} e^{-2\pi ikn/N} x[n] \right\rangle = \frac{1}{N} e^{-2\pi ikn/N} \cdot \langle x[n] \rangle = \frac{1}{N} e^{-2\pi ikn/N} \cdot 0 = 0$$

thus the expectation of the entire sum is zero:

(expectation of DFT partial amplitudes of a centered random process)

$$\langle X[k] \rangle = 0$$

Now we would like to obtain the variance of $X[k]$, which is the expected value of $|X[k]|^2$. By (8.182b) $|X[k]|^2$ is Fourier transform of the autocorrelation. Except that (8.182) is written for continuous Fourier transform and we're currently dealing with the discrete Fourier transform. We can however recall that correlation properties of Fourier transform are identical in form to those of convolution, therefore we can take the DFT convolution property (8.96) (the amplitude version) and rewrite it for autocorrelation yielding

$$U[k] = X^*[k]Y[k] \iff u[n] = \frac{1}{N}(x \star y)$$

and

$$U[k] = X^*[k]X[k] = |X[k]|^2 \iff u[n] = \frac{1}{N}(x \star x)$$

Applying direct transform (8.21a) to $\frac{1}{N}(x \star x)$ yields

$$\langle |X[k]|^2 \rangle = \left\langle \frac{1}{N} \sum_{n=0}^{N-1} \frac{(x \star x)[n]}{N} e^{-2\pi ikn/N} \right\rangle$$

Now

$$\begin{aligned} \left\langle \frac{(x \star x)[n]}{N} \right\rangle &= \left\langle \frac{1}{N} \sum_{\nu=0}^{N-1} x^*[\nu - n] x[\nu] \right\rangle = \frac{1}{N} \sum_{\nu=0}^{N-1} \langle x^*[\nu - n] x[\nu] \rangle \\ &= \frac{1}{N} \sum_{\nu=0}^{N-1} R_{xx}[n] = R_{xx}[n] \end{aligned}$$

where $R_{xx}[n] = \langle x^*[\nu - n] x[\nu] \rangle$ is the stochastic autocorrelation of the process $x[n]$ (notice that since $x[n]$ is N-periodic, so is $R_{xx}[n]$).

▷ Notice that we implicitly assumed that $R_{xx}[n] = \langle x^*[\nu - n]x[\nu]\rangle$ depends only on n and not on ν. Earlier we also used the fact that $\langle x[n]\rangle$ doesn't depend on n. These two conditions are implying that the process is *wide-sense stationary*.

We can therefore drop the assumption of the process being 1st order stationary and replace it with the requirement of wise-sense stationarity. On one hand the former assumption is too narrow, because, at least until now, we didn't care about the specific form of the probability distribution density $p(x)$ (and we won't care about it later either). On the other hand it turned out too wide, because, without specifying the 2nd order stationarity, the autocorrelation R_{xx} will depend both on n and ν and we won't be able to take Fourier transform of R_{xx}.

Thus we obtain

(stochastic direct DFT in terms of partial powers $|X[k]|^2$)

$$\left\langle |X[k]|^2 \right\rangle = \left\langle \frac{1}{N}\sum_{n=0}^{N-1} \frac{(x \star x)[n]}{N} e^{-2\pi ikn/N} \right\rangle = \frac{1}{N}\sum_{n=0}^{N-1} R_{xx}[n] e^{-2\pi ikn/N} \tag{8.248a}$$

That is the variances of $X[k]$ are obtained simply as direct Fourier transform of the stochastic autocorrelation. Recalling that $|X[k]|^2$ is the average power contributed by the k-th partial, we must conclude that $\left\langle |X[k]|^2 \right\rangle$ is the expectation of that power. Thus $\left\langle |X[k]|^2 \right\rangle$ can be considered as powers of the partials.

Respectively autocorrelation is inverse Fourier transform of variances:

(stochastic inverse DFT in terms of partial powers $|X[k]|^2$)

$$R_{xx}[n] = \sum_{k=0}^{N-1} \left\langle |X[k]|^2 \right\rangle e^{2\pi ikn/N} \tag{8.248b}$$

▷ When we said earlier that $x[n]$ are distributed according to the probability distribution density $p(x)$, we didn't say anything about cross-correlation of $x[n_1]$ and $x[n_2]$, but intuitively we might have imagined that $x[n_1]$ and $x[n_2]$ are statistically independent. However in such case $R_{xx}[n]$ is essentially the periodic Kronecker delta function (multiplied by the variance), and respectively $\langle |X[k]|^2\rangle = \text{const}\ \forall k$. So, in order to have $\langle |X[k]|^2\rangle$ varying with k we need to allow $x[n_1]$ and $x[n_2]$ to be correlated.

Notice that by letting $n = 0$ in (8.248b) we obtain

(power expectation of a centered stationary process in terms of DFT partial powers $|X[k]|^2$)

$$\langle |x|^2\rangle = R_{xx}[0] = \sum_{k=0}^{N-1} \left\langle |X[k]|^2 \right\rangle$$

which matches Parseval's theorem (8.82b), since $\frac{1}{N}\sum |x[n]|^2 = \sum |X[k]|^2$ is an estimate of $\langle |x|^2\rangle$. Since $X[k]$ are the amplitudes of the sinusoidal components,

$|X[k]|^2$ are the respective powers and $\langle \sum |X[k]|^2 \rangle = \sum \langle |X[k]|^2 \rangle$ is the expectation of the total power, computed in the spectral domain, which should match the time-domain expectation $\langle |x|^2 \rangle$ of the same total power.

▷ Conversely, by letting $k = 0$ in (8.248a) we obtain

$$\frac{1}{N} \sum_{n=0}^{N-1} R_{xx}[n] = \langle |X[0]|^2 \rangle \geq 0$$

That is the sum of all values of $R_{xx}[n]$ cannot be negative (or complex). Since all $\langle |X[k]|^2 \rangle$ are real, it also follows from (8.248) that R_{xx} is at least Hermitian (or real even, if $x[n]$ are real).

To illustrate the impossibility for the total sum of $R_{xx}[n]$ to be negative, intuitively we can start with $R_{xx}[0] = \langle |x|^2 \rangle > 0$. Then suppose $R_{xx}[1] = R_{xx}[-1] < 0$, that is, in particular, $x[1]$ and $x[-1]$ are anticorrelated with $x[0]$. This suggests that $x[1]$ and $x[-1]$ are somewhat likely to be positively correlated to each other, so $R_{xx}[2] = R_{xx}[-2] > 0$, which also means that $x[2]$ and $x[2]$ are positively correlated with $x[0]$ and are somewhat likely to be anticorrelated with $x[1]$ and $x[-1]$. Etc. This is not a guarantee, of course, e.g. $x[1]$ and $x[-1]$ being anticorrelated to $x[0]$ doesn't necessarily imply that they are positively correlated to each other, but it expresses the general idea that even if there are some negative values of $R_{xx}[n]$, they would be compensated by positive values.

If instead of amplitude version of the transform (8.21) we take the density version (8.23), then, for the same x, we get an N times larger X and an N^2 times larger $|X|^2$, and (8.248a) turns into

$$\langle |X[k]|^2 \rangle = \left\langle \sum_{n=0}^{N-1} (x \star x)[n] e^{-2\pi i k n/N} \right\rangle = N \sum_{n=0}^{N-1} R_{xx}[n] e^{-2\pi i k n/N} \tag{8.249}$$

Now in the density version of DFT we deal with the amplitude density $X[k]$, so that the amplitudes are given by $X[k]\frac{1}{N}$ (as we discussed in Section 8.3). Respectively, it would be consistent to have the power density rather than power. In the amplitude version (8.248) the (expectation of) power was given by $\langle |X[k]|^2 \rangle$. In the density version $|X|^2$ are N^2 times larger therefore the power would be given by $\langle |X[k]|^2 \rangle / N^2$. To turn these values into power density we need to multiply them by N, yielding

(DFT power spectral density in terms of variances of partial amplitude densities $X[k]$)

$$P_{XX}[k] = \frac{\langle |X[k]|^2 \rangle}{N} \tag{8.250}$$

so that $P_{XX}[k]\frac{1}{N} = \langle |X[k]|^2 \rangle / N^2$ give the respective powers. The discrete-argument function $P_{XX}[k]$ given by (8.250) is referred to as *power spectral density*, or shortly *PSD*.

Dividing (8.249) by N, so that we obtain $P_{XX}[n]$ in the left-hand side, yields

(stochastic direct DFT in terms of power spectral density)

$$P_{XX}[k] = \sum_{n=0}^{N-1} R_{xx}[n] e^{-2\pi i k n/N} \tag{8.251a}$$

and the respective inverse transform is thus giving

(stochastic inverse DFT in terms of power spectral density)

$$R_{xx}[n] = \frac{1}{N} \sum_{k=0}^{N-1} P_{XX}[k] e^{2\pi i k n/N} \tag{8.251b}$$

That is power spectral density and autocorrelation are related via a density-style DFT.

Notably, in the symmetric DTF version (8.22), $X[k]$ are $\sqrt{N}$ times smaller than in the density version. Respectively

(power spectral density in terms of variances of symmetric DFT's spectral values)

$$P_{XX}[k] = \langle |X[k]|^2 \rangle \tag{8.252}$$

From the perspective of Parseval's theorem, in the symmetric case we have $\sum |x[n]|^2 = \sum |X[k]|^2$, or, after dividing by N,

(power expectation of a centered stationary process in terms of DFT power spectral density)

$$\begin{aligned}\langle |x|^2 \rangle &= \left\langle \frac{1}{N} \sum_{n=0}^{N-1} |x[n]|^2 \right\rangle = \left\langle \frac{1}{N} \sum_{n=0}^{N-1} |X[n]|^2 \right\rangle = \sum_{n=0}^{N-1} \langle |X[n]|^2 \rangle \frac{1}{N} \\ &= \sum_{n=0}^{N-1} P_{XX}[n] \frac{1}{N}\end{aligned}$$

where in the right-hand side we have the (expectation of) total power of x evaluated in the spectral domain, which is thereby equal to the average power of x given by $\langle |x|^2 \rangle$.

8.20.2 Discrete-time Fourier transform

Goal 8.31. Generalize the previous results to nonperiodic stationary discrete-time random processes.

Attempting to repeat the result of Subsection 8.20.1 for infinitely long nonperiodic discrete-time random processes, we rewrite the DTFT convolution prop-

erty (8.100) for the autocorrelation, obtaining

$$U(f) = X^*(f)Y(f) \iff u[n] = x \star y$$
$$U(f) = X^*(f)X(f) = |X(f)|^2 \iff u[n] = x \star x$$

Applying the direct DTFT formula (8.58a) to $x \star x$ yields

$$\big\langle |X(f)|^2 \big\rangle = \left\langle \sum_{n=-\infty}^{\infty} (x \star x)[n] \cdot e^{-2\pi i f n} \right\rangle = \sum_{n=-\infty}^{\infty} \big\langle (x \star x)[n] \big\rangle \cdot e^{-2\pi i f n} \tag{8.253a}$$

Now

$$\begin{aligned} \big\langle (x \star x)[n] \big\rangle &= \left\langle \sum_{\nu=-\infty}^{\infty} x^*[\nu - n] x[\nu] \right\rangle = \sum_{\nu=-\infty}^{\infty} \big\langle x^*[\nu - n] x[\nu] \big\rangle \\ &= \sum_{\nu=-\infty}^{\infty} R_{xx}[n] = \infty \cdot R_{xx}[n] = \infty \qquad (\text{if } R_{xx}[n] \neq 0) \end{aligned} \tag{8.253b}$$

and therefore $\big\langle |X(f)|^2 \big\rangle$ become infinite as well. Notice the similarity to the density-version DFT case (8.249), where we also get infinity at $N \to \infty$.

It might be instructive to look into this result from a slightly different, more direct perspective, without involving the formal application of the correlation property. Applying the direct DTFT formula (8.58a) to $x[n]$ yields

$$X(f) = \sum_{n=-\infty}^{\infty} x[n] e^{-2\pi i f n} \tag{8.254}$$

Now if we imagine that $x[n]$ are statistically independent (as some kind of a "default neutral state" between being positively and negatively correlated), the variance of the sum in (8.254) will be the sum of variances of the terms, the latter being simply equal to $\langle |x[n] e^{-2\pi i f n}|^2 \rangle = \langle |x[n]|^2 \rangle = \langle |x|^2 \rangle = R_{xx}[0]$ and therefore

$$\big\langle |X(f)|^2 \big\rangle = \sum_{n=-\infty}^{\infty} \big\langle |x[n] e^{-2\pi i f n}|^2 \big\rangle = \sum_{n=-\infty}^{\infty} R_{xx}[0] = \infty$$

If $x[n]$ are statistically dependent, then the variance of $X(f)$ can become bigger or smaller than the sum of the variances of the terms. The main point is, usually it will be still infinite.

▷ Simply put, by the direct DTFT formula (8.58a) $X(f)$ is a linear combination of infinitely many random values $x[n]$ with linear combination coefficients of unit magnitude, and their variances add up kind of proportionally to the number of terms, which is infinite.

The reason for this problem is that DTFT is generally designed to handle functions of finite energy, while stationary random $x[n]$ rather has finite constant average power, therefore the total energy of $x[n]$ (which by (8.88b) is equal to $\int_{-1/2}^{1/2} |X(f)|^2 \, \mathrm{d}f$) diverges.

This problem can be somewhat amended by normalizing the sum in (8.254) by the square root of the number of terms, effectively using the symmetric version of discrete Fourier transform as a basis:

$$X(f) = \lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{n=-N/2}^{N/2} x[n]e^{-2\pi i f n} \tag{8.255a}$$

(where, as usual, we are not particularly careful about choosing exact summation limits, as the difference will disappear in the limit anyway). In this case the variance of $X(f)$ is going to be comparable to $\langle |x|^2\rangle$. In particular, if $x[n]$ are statistically independent, the variance of the sum will be the sum of the variances:

$$\begin{aligned}
\langle |X(f)|^2\rangle &= \lim_{N\to\infty} \left\langle \sum_{n=-N/2}^{N/2} \left| \frac{x[n]}{\sqrt{N}} e^{-2\pi i f n} \right|^2 \right\rangle \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{n=-N/2}^{N/2} \left\langle \left| x[n]e^{-2\pi i f n} \right|^2 \right\rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{n=-N/2}^{N/2} \langle |x[n]|^2\rangle \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{n=-N/2}^{N/2} \langle |x|^2\rangle = \langle |x|^2\rangle = R_{xx}[0]
\end{aligned} \tag{8.255b}$$

There is however a problem with (8.255a): it doesn't look like the limit is converging. Let's take a closer look. Since we don't know the limiting value (and not sure it exists at all to begin with), we are going to use the idea of a Cauchy sequence, which we discussed in Section 2.5. Let

$$X_N = X_N(f) = \frac{1}{\sqrt{N}} \sum_{n=-N/2}^{N/2} x[n]e^{-2\pi i f n} \tag{8.256}$$

so that $X(f)$ in (8.255a) is given by the limit of X_N. We wish to find out whether X_N is a Cauchy sequence (for a fixed f). That is whether the difference $X_{N+M} - X_N$ ($M, N \in \mathbb{N}$) can be made arbitrarily small by choosing a sufficiently large N.

Let $\varepsilon > 0$ be such arbitrarily small value and suppose we found a sufficiently large N such that $|X_{N+M} - X_N| < \varepsilon\ \forall M$. Does this look realistic? Suppose $M = N$. Then

$$\begin{aligned}
X_{N+M} = X_{2N} &= \frac{1}{\sqrt{2N}} \sum_{n=-N}^{N} x[n]e^{-2\pi i f n} \\
&= \frac{1}{\sqrt{2}\sqrt{N}} \sum_{n=-N/2}^{N/2} x[n]e^{-2\pi i f n} + \frac{1}{\sqrt{2}\sqrt{N}} \sum_{|n|=N/2}^{N} x[n]e^{-2\pi i f n} \\
&= \frac{1}{\sqrt{2}} X_N + \frac{1}{\sqrt{2}\sqrt{N}} \sum_{|n|=N/2}^{N} x[n]e^{-2\pi i f n}
\end{aligned} \tag{8.257}$$

where summation from $|n| = N/2$ to N denotes summation from $n = N/2$ to N plus summation from $n = -N$ to $-N/2$.

We can notice that both terms in (8.257) perform a summation of N terms, therefore their magnitudes are comparable. So X_{N+M} is a sum of two terms, each of which is having a magnitude of $\frac{1}{\sqrt{2}}X_N$. These terms also do not need to be correlated at all, since having all $x[n]$ statistically independent is one possible case. This suggests that the magnitude of X_{N+M} can roughly speaking be anywhere between 0 and $\sqrt{2}\cdot X_N$. It doesn't look like it's very probable that $|X_{N+M} - X_N| < \varepsilon$.

Thus $X(f)$ given by (8.255a) doesn't converge per se. This kind of makes sense: $x[n]$ doesn't decay to zero at $n \to \infty$, so there should be no reason why the values of $x[n]$ at larger n should make less contribution to $X(f)$ than the values of of $x[n]$ at smaller n. The readers might recall that we had exactly the same problem in the derivation of DTFT in Section 8.7 and this is why we decided to design DTFT for absolutely summable $x[n]$, so that the contributions of $x[n]$ become smaller at large n.

But the variances of $X(f)$ given by (8.255a) do converge. For statistically independent $x[n]$ we already found these variances by (8.255b). To find them in the general case we can modify (8.253) by using $X(f)$ defined by (8.255a). We can interpret (8.255a) as dividing $x[n]$ by $\sqrt{N}$ rather than dividing the entire sum:

$$X(f) = \lim_{N\to\infty} \sum_{n=-N/2}^{N/2} \frac{x[n]}{\sqrt{N}} e^{-2\pi i f n} \tag{8.258}$$

This turns (8.253b) into an estimate of $R_{xx}[n]$:

$$\left\langle \lim_{N\to\infty} \sum_{\nu=-N/2}^{N/2} \frac{x^*[\nu-n]x[\nu]}{N} \right\rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{\nu=-N/2}^{N/2} \left\langle x^*[\nu-n]x[\nu] \right\rangle$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{\nu=-N/2}^{N/2} R_{xx}[n] = R_{xx}[n]$$

and respectively (8.253a) becomes

$$\left\langle |X(f)|^2 \right\rangle = \sum_{n=-\infty}^{\infty} R_{xx}[n] \cdot e^{-2\pi i f n} \tag{8.259}$$

Equation (8.259) looks like a limiting version of (8.251). Indeed, introducing $P_{XX}(f) = \langle |X(f)|^2 \rangle$ such that $P_{XX}(k/N) = \langle |X(k/N)|^2 \rangle = \langle |X[k]|^2 \rangle = P_{XX}[k]$ and letting $N \to \infty$, in the limit we obtain from (8.251)

(stochastic DTFT)

$$P_{XX}(f) = \sum_{n=-\infty}^{\infty} R_{xx}[n] e^{-2\pi i f n} \tag{8.260a}$$

$$R_{xx}[n] = \int_{-1/2}^{1/2} P_{XX}(f) e^{2\pi i f n} \, \mathrm{d}f \tag{8.260b}$$

that is $P_{XX}(f)$ is the DTFT of $R_{xx}[n]$. We can verify that $P_{XX}(f)$ still has density semantics with respect to f by letting $n = 0$ in (8.260b), which yields the expected power of x as

(power expectation of a centered stationary process in terms of DTFT power spectral density)

$$\langle |x|^2 \rangle = R_{xx}[0] = \int_{-1/2}^{1/2} P_{XX}(f)\,\mathrm{d}f$$

Thus $P_{XX}(f)$ is power spectral density in the DTFT case.

Notably, in order for (8.260a) to converge and in order for (8.260) roundtrip to work, we might require $R_{xx}[n]$ to be absolutely summable. This requirement is not unrealistic: the values $x[n_1]$ and $x[n_2]$ which are lying far apart ($|n_1 - n_2|$ is large), are likely to be uncorrelated. If that is not the case, (8.260a) might diverge.

▷ Similarly to the DFT case, since $P_{XX}(f) \geq 0$, we obtain from (8.260) that R_{xx} is Hermitian (real even, if $x[n]$ are real), and that $\sum_{n=-\infty}^{\infty} R_{xx}[n] \geq 0$.

Goal 8.31 completed. ■

At this point it would be a good time to draw a parallel between our application of symmetric DFT in (8.255a) and the application of symmetric DFT for STFT purposes, suggested in Section 8.19. Back in Section 8.19 we said that the symmetric version is applicable to functions of approximately constant average power, except if those functions are truly periodic, where amplitude version would be preferred. Now we have another way to look at this: stationary processes are also having constant average power. This suggests that symmetric DFT might be good whenever $x[n]$ has constant average power (if we want to vary the DFT length N, which is, in particular, the case if we are taking the limit $N \to \infty$).

We also mentioned in Section 8.19 that for truly periodic functions amplitude DFT could work better than symmetric DFT. In case of random processes, periodicity of $x[n]$ means periodicity of $R_{xx}[n]$. In this case $R_{xx}[n]$ is not absolutely summable and also (8.260a) diverges. This corresponds to the fact that DTFT is not too suitable for handling periodic functions of a discrete argument and we should use DFS instead.

So nonperiodic stationary random processes are somewhere "in-between" finite-energy functions (handled by DTFT, which is a limiting version of density DFT) and periodic functions (handled by DFS, which is amplitude-version DFT). They have constant average power, similarly to periodic functions, but do not have true periodicity. Being in the middle, they are to be handled by the (limiting version of) symmetric DFT.

▷ It could be instructive to review the discussion of discrete-domain STFT in Section 8.19 from the perspective of formally considering $x[n]$ as a random process (even if $x[n]$ is not really random).

We are not going to look into stochastic Fourier series, leaving this as an exercise to the reader and move on directly to continuous Fourier transform.

8.20.3 Fourier transform

Goal 8.32. Generalize the previous results to continuous-time nonperiodic stationary random processes.

Following our usual approach we take the autocorrelation property (8.182b) of Fourier transform yielding

$$\langle |X(f)|^2 \rangle = \left\langle \int_{-\infty}^{\infty} (x \star x)(t) e^{-2\pi i f t} \, \mathrm{d}t \right\rangle = \int_{-\infty}^{\infty} \big\langle (x \star x)(t) \big\rangle e^{-2\pi i f t} \, \mathrm{d}t \tag{8.261a}$$

where

$$\begin{aligned} \big\langle (x \star x)(t) \big\rangle &= \left\langle \int_{-\infty}^{\infty} x^*(\tau - t) x(\tau) \, \mathrm{d}\tau \right\rangle = \int_{-\infty}^{\infty} \big\langle x^*(\tau - t) x(\tau) \big\rangle \, \mathrm{d}\tau \\ &= \int_{-\infty}^{\infty} R_{xx}(t) \, \mathrm{d}\tau = \infty \qquad (\text{for } R_{xx}(t) \neq 0) \end{aligned} \tag{8.261b}$$

which is the Fourier transform counterpart of the result we obtained for DTFT in (8.253).

Apparently we are facing a similar problem as with DTFT: integral Fourier transform is designed to handle functions of finite energy, while the energy of a stationary random process diverges. This also suggests that the problem can be amended in a fashion similar to (8.255a) and (8.258), by redefining the direct Fourier transform of $x(t)$ as

$$X(f) = \lim_{T \to \infty} \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} x(t) e^{-2\pi i f t} \, \mathrm{d}t = \lim_{T \to \infty} \int_{-T/2}^{T/2} \frac{x(t)}{\sqrt{T}} e^{-2\pi i f t} \, \mathrm{d}t \tag{8.262}$$

This turns (8.261b) into

$$\begin{aligned} \left\langle \lim_{T \to \infty} \int_{-T/2}^{T/2} \frac{x^*(\tau - t) x(\tau)}{T} \, \mathrm{d}\tau \right\rangle &= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \big\langle x^*(\tau - t) x(\tau) \big\rangle \, \mathrm{d}\tau \\ &= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} R_{xx}(t) \, \mathrm{d}\tau = R_{xx}(t) \end{aligned}$$

and (8.261a) is turned into

$$\langle |X(f)|^2 \rangle = \int_{-\infty}^{\infty} R_{xx}(t) e^{-2\pi i f t} \, \mathrm{d}t \tag{8.263}$$

Now we might guess that, like with DTFT, $\langle |X(f)|^2 \rangle$ is the power spectral density, therefore, replacing the notation with $P_{XX}(f)$, we obtain

(stochastic Fourier transform)

$$P_{XX}(f) = \int_{-\infty}^{\infty} R_{xx}(t) e^{-2\pi i f t} \, \mathrm{d}t \tag{8.264a}$$

$$R_{xx}(t) = \int_{-\infty}^{\infty} P_{XX}(f) e^{2\pi i f t} \, \mathrm{d}f \tag{8.264b}$$

However we cannot be really sure yet, whether $P_{XX}(f)$ is indeed having power spectral density semantics. We also have been too handwavy in using variances of integrals of stochastic processes.

▷ The problem with variances of integrals of stochastic processes is that for uncorrelated processes they are equal to zero. Indeed, consider e.g. the integral

$$I = \int_0^1 x(t)\,\mathrm{d}t = \lim_{N\to\infty} \sum_{n=0}^{N-1} x(n/N)\frac{1}{N}$$

and suppose $x(t)$ is fully uncorrelated. Then variance of the sum is sum of variances:

$$\begin{aligned}\langle |I|^2\rangle &= \left\langle \left| \lim_{N\to\infty} \sum_{n=0}^{N-1} x(n/N)\frac{1}{N} \right|^2 \right\rangle = \lim_{N\to\infty} \sum_{n=0}^{N-1} \left\langle \left| x(n/N)\frac{1}{N} \right|^2 \right\rangle \\ &= \lim_{N\to\infty} \frac{1}{N^2} \sum_{n=0}^{N-1} \langle |x(n/N)|^2\rangle = \lim_{N\to\infty} \left(\frac{1}{N^2} N \cdot \langle |x|^2\rangle \right) = \langle |x|^2\rangle \cdot \lim_{N\to\infty} \frac{1}{N} \\ &= \langle |x|^2\rangle \cdot 0 = 0\end{aligned}$$

Uncorrelated continuous-time processes are also "discontinuous everywhere" (without trying to get specific what exactly that means), since autocorrelation (more specifically, $R_{xx}(t)$ being nonzero in some neighborhood of $t = 0$) acts like a "source of continuity". This raises the question of applicability of Fourier transform to such processes.

So, in order to be more convincing, we will attempt to obtain (8.264) as a limiting version of (8.260). As the first step, we shall rewrite (8.260) for a generic non-unit step size T, obtaining

$$P_{XX}(f) = T \sum_{n=-\infty}^{\infty} R_{xx}[n] e^{-2\pi i f n T} \tag{8.265a}$$

$$R_{xx}[n] = \int_{-1/2T}^{1/2T} P_{XX}(f) e^{2\pi i f n T}\,\mathrm{d}f \tag{8.265b}$$

where we noticed that the variance $\langle |x[n]|^2\rangle = R_{xx}[0] = \int_{-1/2T}^{1/2T} P_{XX}(f)\,\mathrm{d}f$ should be unchanged by this transformation, therefore the values of P_{XX} should become T times larger. In other words, the frequency range becomes T times smaller, therefore P_{XX} should become T times larger to fit the same total power into a smaller frequency range.

We want to let $T \to 0$, where $R_{xx}[n]$ get replaced by $R_{xx}(nT)$. Respectively, (8.265) becomes

$$P_{XX}(f) = \sum_{n=-\infty}^{\infty} R_{xx}(nT) e^{-2\pi i f n T} \cdot T \tag{8.266a}$$

$$R_{xx}(nT) = \int_{-1/2T}^{1/2T} P_{XX}(f) e^{2\pi i f n T}\,\mathrm{d}f \tag{8.266b}$$

Letting $t = nT$ and $T \to 0$ in (8.266) yields (8.264) and therefore $P_{XX}(f)$ in (8.264) is indeed power spectral density.

Goal 8.32 completed. ■

Notice that, by the reciprocal duality of the Fourier transform domains (the Heisenberg uncertainty relationship (8.225)), the larger is σ_P (the "width" of $P_{XX}(f)$), the smaller can σ_R (the "width of $R_{xx}(t)$") get and vice versa. That is strongly autocorrelated signals can have small bandwidth, while weakly autocorrelated signals must have large bandwidth.

▷ Intuitively, this also suggests that an uncorrelated signal must therefore have infinitely large bandwidth. There are however some difficulties with rigorous understanding of that statement. Indeed, an uncorrelated $x(t)$ means

$$R_{xx}(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases} \tag{8.267}$$

therefore by (8.266a) $P_{XX}(f) \equiv 0$. This also means that the total power of $x(t)$ given by $\int_{-\infty}^{\infty} P_{XX}(f)\,\mathrm{d}f$ is zero as well. But the latter would mean that the variance of $x(t)$ is zero, which doesn't make sense, as we usually a priori implicitly assume $\langle |x|^2 \rangle \neq 0$.

This difficulty can be amended by understanding Fourier transform of (8.267) as being infinitely small instead of zero:

$$P_{XX}(f) = \int_{-\infty}^{\infty} R_{xx}(t) e^{-2\pi i f t}\,\mathrm{d}t = R_{xx}(0) \cdot \varepsilon$$

where ε denotes an "infinitely small value". In that sense we have $\sigma_R = 0$ and $\sigma_P = \infty$, in agreement with Heisenberg uncertainty relationship (8.225). We also have, by the constant term property of Fourier transform

(power expectation of a centered stationary process in terms of Fourier tranform power spectral density)

$$\langle |x|^2 \rangle = R_{xx}(0) = \int_{-\infty}^{\infty} P_{XX}(f)\,\mathrm{d}f$$

Thus the finite average power $\langle |x|^2 \rangle$ is getting distributed across the entire range of f from $-\infty$ to $+\infty$, therefore the power spectral density is infinitely small.

In practice, media in which signals propagate are usually bandlimited in one or another way and cannot transport signals of an infinite bandwidth. The time-domain side of this is that fully uncorrelated processes are discontinuous (since there is nothing "holding the neighboring values of such process closely together"), and media cannot support discontinuous signals either. This suggests that fully uncorrelated continuous-time processes are a rather far-fetched mathematical abstraction.

This problem doesn't exist with discrete time, because a single nonzero value $R_{xx}[0]$ (which is equal to the variance of x and therefore can be assumed nonzero) is sufficient to make $P_{XX}(f)$ non-zero.

8.20.4 Distribution of spectral values

We have looked into the spectral values (partial amplitudes or amplitude densities) of random processes, where we found that these spectral values are random values themselves, their expectation is zero (assuming the process is centered), while the variance is obtained as Fourier transform of the autocorrelation of the process. We also found that we generally need to use special normalization

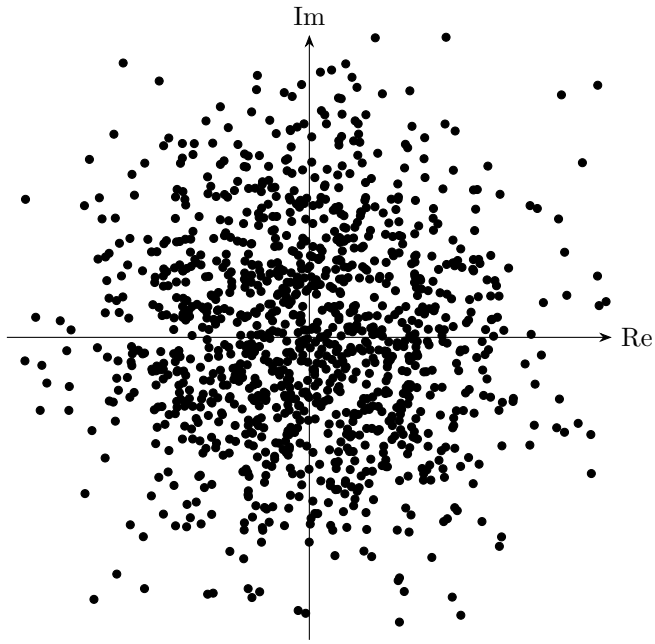


Figure 8.26: By the central limiting theorem, $X[k]$ are normally distributed in the complex plane.

techniques, essentially derived from the symmetric version of DFT, to ensure finite and nonzero variances. So we found the expectation and the variances of the spectral values, but what about their distribution?

Goal 8.33. Assess the types of probability distributions describing the spectral values of random processes. Do this for the complex amplitudes, as well as for the real amplitudes and phases.

Let's return to discrete Fourier transform in its stochastic version (8.248). Other types of Fourier transform are limiting versions of the DFT, so the arising distributions must be similar across all types.

By the central limiting theorem we should expect each of $X[k]$ in turn (for each given fixed k) to have Gaussian distribution, if N is sufficiently large. That is, pretty much regardless of the details of the stochastic properties of $x[n]$, the values $X[k]$ are normally distributed. This distribution is, however, two-dimensional, since $X[k]$ are defined in the complex plane, with the real and imaginary dimensions:

$$p(X[k]) = \frac{1}{\pi\sigma^2} e^{-|X|^2/\sigma^2} = \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{2}}} e^{-(\operatorname{Re} X)^2/2(\sigma^2/2)} \cdot \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{2}}} e^{-(\operatorname{Im} X)^2/2(\sigma^2/2)}$$

where for brevity we denoted $X = X[k]$ and $\sigma = \sigma_{X[k]}$. Fig. 8.26 illustrates.

▷ Notice that the variances $\sigma^2/2$ of $\operatorname{Re} X$ and $\operatorname{Im} X$ are equal to half of the variance σ^2 each. Since $|X|^2 = (\operatorname{Re} X)^2 + (\operatorname{Im} X)^2$, and since $\operatorname{Re} X$ and $\operatorname{Im} X$ are independent (due to $\operatorname{Re} x$ and $\operatorname{Im} x$ being independent), the variances add.

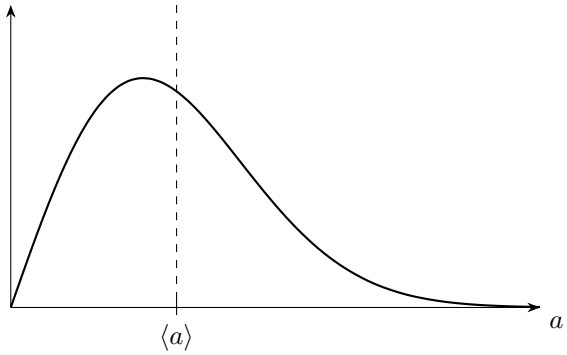


Figure 8.27: Since $X[k]$ are normally distributed, their magnitudes $a_k = |X[k]|$ are described by the Rayleigh distribution.

Distribution of spectral magnitudes

Now we shall look into the distribution of the absolute magnitudes $|X[k]|$. For the sake of simplicity let's assume that the standard deviation of $X[k]$ is unity, in which case the probability distribution density of $X[k]$ (for a given fixed k) is

$$p(X[k]) = \frac{1}{\pi} e^{-|X|^2}$$

Other values of standard deviation can be handled by simply scaling the X axis by σ.

Being interested in $|X[k]|$ means that all $X[k]$ at the same distance from the complex origin are treated as one and the same value. Such $X[k]$ are located on a circle of a radius $|X[k]|$, the length of this circle being proportional to its radius. Thus there is an additional factor proportional to the distance from the origin in the distribution density function and the distribution function becomes (qualitatively) ae^{-a^2}, where $a = a_k = |X[k]|$.

To be more precise, we need to compute the integral under ae^{-a^2}, which, with correct normalization, must evaluate to unity. We obtain

$$\int_0^\infty ae^{-a^2}\,\mathrm{d}a = \int_{a=0}^\infty e^{-a^2}\,\mathrm{d}\left(\frac{a^2}{2}\right) = \frac{1}{2}\int_{a=0}^\infty e^{-a^2}\,\mathrm{d}(a^2)$$
$$= \frac{1}{2}\int_0^\infty e^{-t}\,\mathrm{d}t = \left.\frac{-e^{-t}}{2}\right|_{t=0}^\infty = \frac{0-(-1)}{2} = \frac{1}{2}$$

and therefore, from the condition $\int_0^\infty p(a)\,\mathrm{d}a = 1$,

$$p(a_k) = 2ae^{-a^2} \qquad (a \geq 0) \tag{8.268}$$

The probability distribution density (8.268) defines the *Rayleigh distribution* (Fig. 8.27). Rayleigh distribution is not exactly limited to (8.268), we can scale the a axis to accommodate non-unit values of the standard deviation of $X[k]$.

To find the expectation value of a_k compute

$$\langle a_k \rangle = \int_0^\infty ap(a)\,\mathrm{d}a = \int_0^\infty 2a^2 e^{-a^2}\,\mathrm{d}a = -\int_{a=0}^\infty a\,\mathrm{d}(e^{-a^2})$$

$$= -ae^{-a^2}\Big|_{a=0}^{\infty} + \int_{a=0}^{\infty} e^{-a^2}\,\mathrm{d}a = 0 + \frac{1}{2}\int_{a=-\infty}^{\infty} e^{-a^2}\,\mathrm{d}a = \frac{\sqrt{\pi}}{2}$$

where we used the value of the Gaussian integral $\int_{-\infty}^{\infty} e^{-t^2}\,\mathrm{d}t = \sqrt{\pi}$.

▷ If we plot $\langle a\rangle$ on the horizontal axis of the probability distribution density graph (Fig. 8.27), we notice that it doesn't coincide with the distribution's peak. Indeed the peak is found from

$$p'(a) = 2e^{-a^2} + 2a\cdot(-2ae^{-a^2}) = 2e^{-a^2}(1-2a^2) = 0$$

yielding $a = 1/\sqrt{2} \approx 0.7$, while $\langle a\rangle = \sqrt{\pi}/2 \approx 0.9$.

To find the standard deviation, first compute the variance:

$$\begin{aligned}\langle (a_k)^2\rangle &= \int_0^{\infty} a^2 p(a)\,\mathrm{d}a = \int_0^{\infty} 2a^3 e^{-a^2}\,\mathrm{d}a = -\int_{a=0}^{\infty} a^2\,\mathrm{d}(e^{-a^2})\\ &= -\int_0^{\infty} t\,\mathrm{d}(e^{-t}) = te^{-t}\Big|_{t=0}^{\infty} + \int_0^{\infty} e^{-t}\,\mathrm{d}t = 0+1 = 1\end{aligned}$$

and the standard deviation σ_a is obtained from

$$\sigma_a^2 = \langle a^2\rangle - \langle a\rangle^2 = 1 - \frac{\pi}{4}$$

yielding

$$\sigma_a = \sqrt{1-\frac{\pi}{4}}$$

Distribution of phases

Now let's look at the distribution of $\arg X[k]$. Intuitively, we should expect $\arg X[k]$ to be uniformly distributed on $(-\pi,\pi]$, however this will be the case only if the Gaussian distribution of $X[k]$ is symmetric with respect to rotations of the complex plane. If $x[n]$ are "fully random" (whatever that means) we should expect this kind of symmetry. However there might be some restrictions on $x[n]$. E.g. one such typical restriction could be that $x[n]$ are real. This immediately implies that so is $X[0]$. So, $X[0]$ is still likely to have Gaussian distribution, however this distribution is no longer going to be symmetric: its variance in the direction of the imaginary axis is going to be zero (the same holds for $X[N/2]$ if N is even).

This consideration is also applicable to the case of continuous spectrum (discrete-time Fourier transform and continuous Fourier transform), where not only $\langle|\operatorname{Im} X(0)|^2\rangle$ is going to be zero, but also generally $\langle|\operatorname{Im} X(\omega)|^2\rangle$ is going to be small in the neighborhood of $\omega = 0$, and respectively the values of $\arg X(\omega)$ will be concentrating around 0 and π within this neighborhood.

Goal 8.33 completed. ■

Summary of Chapter 8

- We have introduced discrete Fourier series (DFS) as an expansion of a discrete periodic function into sinusoidal components. By writing the

expansion equations in the matrix form we obtained representation of DFS as an (almost) unitary transformation of a linear space, also, somewhat unconventionally, re-dubbing DFS as discrete Fourier transform (DFT).

- We introduced 3 different normalization placements: amplitude, symmetric and density. By taking the limits at the infinitely growing period and/or the infinitely shrinking discrete step we respectively obtained the discrete-time Fourier transform (DTFT), Fourier series and integral Fourier transform, where Fourier series and DTFT are each other's duals, while DFS and integral Fourier transforms are their own duals each.

- We have discovered a number of transform properties, the majority of which are applicable to all 4 transform types. By analyzing the dual-domain effects of filtering, we discovered the convolution property. As a means to aid convolution-based analysis, we introduced Kronecker and Dirac delta functions, which are helpful in analyzing the transform convergence, as well as can serve further utility roles.

- We studied the effects of discontinuities and features, causing the appearance of sinusoidal ripples and affecting the decay rate in the dual domain. We also introduced the ideas of power and energy in the context of Fourier transform, as well as 1st and 2nd energy momenta, and studied their effects and the ways to compute those in the dual domain.

- Looking at numerical computability of Fourier transform, we introduced sampling and windowing (STFT) as a means to ensure that the transform can be computed in finite time. We also looked at Fourier analysis of random processes, where we found that they are "somewhere in-between" periodic functions and finite-energy nonperiodic functions.

Revision history

1.0.0-alpha (Dec 13, 2023)

first public version

Detailed table of contents

Contents (overview) **3**

Introduction **5**
For whom is this book . . . 5
Why this book . . . 6
The approach of this book . . . 7
Time and place for rigor . . . 8
Proofs and cookbooks . . . 9
Separation of concerns vs. holistic approach . . . 9
Motivation and intuition . . . 10
How to read this book . . . 11
Acknowledgements . . . 15

I Mostly algebra **17**

1 Some experiments **19**
1.1 Quadratic equation . . . 19
1.2 Fundamental theorem of algebra . . . 23
1.2.1 Real case . . . 24
1.2.2 Complex case . . . 25
1.3 Fundamental theorem of arithmetic . . . 28
1.3.1 Modified Euclid's lemma . . . 28
1.3.2 LCM and GCD . . . 30
Summary of Chapter 1 . . . 31

2 Real numbers **33**
2.1 Periodic decimal fractions . . . 34
2.2 Non-periodic decimal fractions . . . 36
2.3 Decimal fraction convergence . . . 37
2.4 Limits of sequences . . . 38
2.5 Postulated existence of real numbers . . . 42
2.5.1 Cauchy sequences . . . 42
2.5.2 The least bound axiom . . . 45
2.5.3 The Dedekind cut . . . 46
2.6 Continuity of operations . . . 47
2.7 Comparisons . . . 49
2.8 Uncountability . . . 52

2.8.1 Countable sets . . . 52
2.8.2 Cantor's diagonal argument . . . 55
2.8.3 Uncountability implications . . . 56
2.9 Continuity of the real line . . . 58
Summary of Chapter 2 . . . 60

3 Linearity everywhere **63**
3.1 Linear spaces . . . 63
3.2 Linear functions . . . 69
3.3 Dot product . . . 72
3.4 Rotations . . . 74
3.4.1 Two dimensions . . . 74
3.4.2 Multiple dimensions . . . 76
3.4.3 Orthogonal transformations . . . 77
3.4.4 Orthogonal transformations vs. rotations . . . 79
3.5 Small scale picture . . . 81
3.5.1 One-dimensional zoom-in . . . 81
3.5.2 Multiple dimensions . . . 84
3.6 Linearity and complex numbers . . . 86
Summary of Chapter 3 . . . 91

4 Beyond the basis **93**
4.1 Covariant-form equations . . . 94
4.2 Matrix notation . . . 95
4.3 Basis change . . . 97
4.4 Linear forms . . . 101
4.5 Co- and contra-variance . . . 102
4.6 Tensor index notation . . . 104
4.7 Tensor components . . . 108
4.8 Kronecker delta symbol . . . 110
4.9 Metric tensor . . . 111
4.9.1 Dot product . . . 111
4.9.2 Index lowering . . . 113
4.9.3 Index rasing . . . 114
4.10 Tensor product . . . 115
4.11 Matrix trace . . . 117
4.12 Decomposable tensors . . . 120
4.13 Dual basis . . . 122
4.14 Differentiation . . . 124
4.15 Linear transformations of space . . . 128
4.15.1 Equivalence to basis change . . . 128
4.15.2 Pseudotensors . . . 130
4.16 Tensors and complex numbers . . . 131
Summary of Chapter 4 . . . 132

5 Area and volume **135**
5.1 Exterior product in 2 dimensions . . . 136
5.2 Symmetries . . . 141
5.3 Exterior vs. dot product . . . 146
5.4 Volume . . . 147

5.5 Bivectors . . . 152
5.5.1 Exterior product in 3D . . . 152
5.5.2 Bivector space in 3D . . . 157
5.5.3 Bivectors in 2D . . . 161
5.5.4 Dot product of bivectors . . . 162
5.5.5 Higher dimension counts . . . 165
5.6 k-vectors . . . 167
5.6.1 Associative wedge product in 3D . . . 167
5.6.2 3-vector space . . . 169
5.6.3 k-vector space . . . 170
5.7 Volume deformations . . . 172
5.7.1 Covariant Levi–Civita tensor . . . 172
5.7.2 Matrix determinants reviewed . . . 176
5.8 Pseudotensors revisited: Hodge star . . . 179
5.8.1 Motivation and definition . . . 179
5.8.2 Linearity . . . 182
5.8.3 1-vectors and 0-vectors . . . 184
5.8.4 Self-inverse property . . . 186
5.9 k-vectors as tensors . . . 187
5.9.1 Bivectors as tensors . . . 187
5.9.2 3D 3-vectors as tensors . . . 188
5.9.3 Hodge star on tensors . . . 191
Summary of Chapter 5 . . . 193

II Mostly calculus 195

6 Instantaneous and additive quantities 197
6.1 Derivative . . . 198
6.1.1 Motivation . . . 198
6.1.2 Consistency . . . 201
6.1.3 Differentiation of arbitrary functions . . . 203
6.1.4 Semantical interpretation . . . 204
6.2 Riemann integral . . . 205
6.2.1 Motivation . . . 205
6.2.2 Consistency . . . 206
6.2.3 Arbitrary integration limits . . . 209
6.2.4 Semantical interpretation . . . 210
6.3 Indefinite integral . . . 212
6.3.1 Antiderivative . . . 212
6.3.2 Fundamental theorem of calculus . . . 214
6.3.3 Differentiability of integrals . . . 215
6.4 Riemann–Stieltjes integral . . . 218
6.5 Differentials . . . 221
6.5.1 Differentials in differentiation context . . . 221
6.5.2 Differentials in integration context . . . 222
6.5.3 Formal definition . . . 223
6.5.4 Higher-order differentials . . . 225
6.5.5 Invariance of 1st-order differentials . . . 226
6.6 Improper integrals . . . 229

6.7 Partial derivatives . . . 233
6.7.1 Tangent plane . . . 233
6.7.2 Tangent lines in arbitrary directions . . . 236
6.7.3 Full derivative . . . 237
6.7.4 Partial derivative notation . . . 238
6.7.5 Gradient . . . 240
6.8 Multidimensional and iterated integrals . . . 241
6.8.1 Area integrals in 2D . . . 241
6.8.2 Unoriented areas in 2D . . . 244
6.8.3 Iterated integrals . . . 247
6.8.4 Area orientation in 2D . . . 251
6.8.5 Volume and hypervolume integrals . . . 252
6.8.6 Further integrals in multiple dimensions . . . 254
6.8.7 Iterated improper integrals . . . 257
6.9 Average value of a function . . . 261
6.9.1 Integral form in 1D . . . 261
6.9.2 Differentiation-based form in 1D . . . 265
6.9.3 Multiple dimensions . . . 266
6.10 Boundary vs. interior behavior of functions . . . 268
6.10.1 Gauss's theorem for electrostatic field . . . 268
6.10.2 Divergence . . . 274
6.10.3 Gauss's theorem for arbitrary fields . . . 277
6.10.4 Green's theorem . . . 283
6.10.5 Stokes' theorem . . . 288
6.10.6 Repeated application of nabla operator . . . 291
Summary of Chapter 6 . . . 295

7 Exponential function 297
7.1 The simplest differential equation . . . 298
7.1.1 Constructive solution . . . 299
7.1.2 Negative arguments . . . 301
7.1.3 Taylor series . . . 305
7.2 Natural exponential . . . 305
7.3 Arbitrary-base exponentials . . . 308
7.4 Ordinary differential equations . . . 310
7.4.1 Solution goals . . . 310
7.4.2 Common solution approaches . . . 311
7.4.3 Constructive solution . . . 312
7.5 Matrix exponential . . . 314
7.6 Matrix diagonalization . . . 316
7.7 Second-order differential equation . . . 319
7.7.1 Common-case steps . . . 320
7.7.2 Overdamped case . . . 321
7.7.3 Motivation for complex numbers . . . 323
7.7.4 Underdamped case . . . 327
7.7.5 Critically-damped case . . . 328
7.8 Jordan normal form . . . 329
7.8.1 Rank 2 generalized eigenvectors . . . 330
7.8.2 Rank 3 generalized eigenvectors . . . 333
7.8.3 Rank 2 generalized eigenspace in 3D . . . 336

7.8.4 Jordan block . . . 338
7.8.5 Generalized eigenspaces . . . 340
7.8.6 Exponentiation of Jordan blocks . . . 343
7.9 Spectral theorems . . . 345
7.9.1 Diagonalization of quadratic forms . . . 346
7.9.2 Symmetric and Hermitian matrices . . . 353
7.9.3 Spectrum of a matrix transpose . . . 358
7.9.4 Orthogonal and unitary matrices . . . 363
7.9.5 Antisymmetric and anti-Hermitian matrices . . . 366
7.9.6 Normal matrices . . . 367
7.9.7 Non-unitary bases . . . 369
7.9.8 Operator classes . . . 370
7.10 Rotations as exponentials . . . 372
Summary of Chapter 7 . . . 378

8 Superpositions of sinusoids 379
8.1 Discrete Fourier series . . . 381
8.2 Discrete Fourier transform . . . 385
8.3 Domain semantics . . . 388
8.4 Duality . . . 391
8.4.1 Symmetric version . . . 391
8.4.2 Amplitude/density duality . . . 393
8.4.3 Application principles . . . 394
8.4.4 Application example: the constant term . . . 396
8.5 Real discrete Fourier transform . . . 398
8.6 Symmetries . . . 402
8.7 Discrete-time Fourier transform . . . 404
8.8 Fourier series . . . 407
8.9 Fourier transform . . . 412
8.10 Unitarity . . . 417
8.11 Convolution . . . 423
8.12 Shift and modulation . . . 430
8.13 Delta function . . . 432
8.13.1 Nonperiodic Kronecker delta . . . 432
8.13.2 Periodic Kronecker delta . . . 435
8.13.3 DFT roundtrip . . . 437
8.13.4 DTFT roundtrip . . . 438
8.13.5 Periodic Dirac delta . . . 439
8.13.6 Fourier series roundtrip . . . 446
8.13.7 Nonperiodic Dirac delta . . . 447
8.13.8 Fourier transform roundtrip . . . 453
8.13.9 Convergence as filtering . . . 454
8.13.10 Continuous representation of discrete functions . . . 456
8.13.11 Normalization checks . . . 457
8.14 Correlation . . . 459
8.15 Differentiation and integration . . . 461
8.16 Discontinuities . . . 464
8.16.1 Sawtooth function . . . 464
8.16.2 Periodic piecewise parabolic function . . . 467
8.16.3 Periodic piecewise polynomial functions . . . 468

8.16.4 Periodic arbitrary piecewise functions 471
8.16.5 Discontinuities in nonperiodic functions 473
8.16.6 Integral sine function 475
8.16.7 Compactly supported spectra 478
8.16.8 Spectral discontinuities 479
8.16.9 Dual-domain effects of features 481
8.17 Domain scaling 482
8.18 Power and energy 484
8.18.1 Power and energy in two domains 484
8.18.2 Energy center 485
8.18.3 Energy concentration 489
8.18.4 Energy concentration bound 490
8.18.5 Maximally concentrated energy 494
8.18.6 Energy center shift 498
8.19 Finite computability 503
8.19.1 Sampling 503
8.19.2 Domain truncation 508
8.19.3 Windowing 510
8.19.4 Window width and height 512
8.19.5 Spectral-domain sampling 514
8.19.6 Localization of frequency 516
8.19.7 Discrete-domain windowing 517
8.20 Stochastic Fourier transform 519
8.20.1 Discrete Fourier transform 519
8.20.2 Discrete-time Fourier transform 523
8.20.3 Fourier transform 528
8.20.4 Distribution of spectral values 530
Summary of Chapter 8 533

Revision history **535**

Detailed table of contents **537**

Index **543**

Index

2-blade, 166
2-vector, 153
    decomposable, 166
    simple, 166
3-vector, 169

additivity, 69
adjoint operator, 371
aliasing, 383, 389, 506
amplitude, 388
    real, 400
amplitude density, 388, 389
amplitude version
    of DFT, 386
amplitude-like version
    of Fourier transform, 415
analytic function, 90
anti-Hermitian function, 402
anti-Hermitian matrix, 366
anti-Hermitian operator, 370
anti-Hermitian sequence, 402
antiderivative, 212
antilinearity, 88
antisymmetric matrix, 366, 374
area
    of a parallelogram, 136
    oriented, 139, 156
    signed, 139
    unoriented, 139, 244
    unsigned, 139, 244
area integral, 241
autocorrelation, 460
    stochastic, 520

bandlimiting, 506
base frequency, 389, 411
basis
    orthonormal, 345
    unitary, 345
basis change, 97, 128, 317
basis-independence, 94
bilinear form, 99
    components of, 108
bilinearity, 72
bivector, 153
    decomposable, 166
    simple, 166

Cantor's diagonal argument, 55
cardinal sine, 328, 448
cardinality
    of a set, 52
Cauchy principal value, 231
Cauchy sequence, 38, 42
Cauchy–Schwarz inequality, 492
characteristic polynomial
    of a matrix, 320
circulaton
    of a field, 285
completing the square, 22
complex numbers, 86
    motivation for, 323
congruence transformation, 100
    complex, 357
contraction
    of tensor indices, 105, 117
contravariant index, 102
contravariant transformation, 102
convergence
    of a sequence, 37, 38
    of decimal fractions, 37
    of DTFT, 438
    of Fourier series, 446
        pointwise, 467
        uniform, 466
    of Fourier transform, 453
    uniform, 259
convolution, 430
    discrete-time, 427
    discrete-time periodic, 425
    periodic, 428
convolution property

of Fourier transform, 423
correlation property
of Fourier transform, 459
countable set, 52
covariant form, 94, 108
covariant index, 102
covariant Levi–Civita tensor, 172
covariant transformation, 102
cross product, 130, 179
cross-correlation, 460
curl, 290
2D, 286

Darboux sum, 207
DC term, 397
decimal fraction
convergence of, 37
nonperiodic, 36
periodic, 34
decomposable tensor, 120
Dedekind cut, 46
delta
Dirac, 220, 447
periodic, 439
Kronecker, 432
periodic, 435
delta function, 432
density version
of DFT, 387
density-like version
of Fourier transform, 414
derivative
of a function, 198
partial, 233
determinant
of a matrix, 152
DFFT, 410
DFS, 381
DFT, 385
conventional meaning, 387, 518
real, 398
roundtrip of, 437
stochastic, 519
diagonalization
of a matrix, 316
difference equation, 299
differential, 203
1st-order, 221
differential equation
1st order, 310
2nd order, 319
critically damped, 328
overdamped, 321
underdamped, 323
simplest, 298
differentiation
of a function, 198
partial, 233
under a limit, 260
under an integral, 249
differentiation property
of Fourier transform, 461
Dirac comb, 444
Dirac delta, 220, 447
periodic, 439
Dirichlet kernel, 440
discontinuity, 464
discrete Fourier series, 381
discrete Fourier transform, 385
conventional meaning, 387, 518
real, 398
roundtrip of, 437
stochastic, 519
discrete-time Fourier transform, 404
roundtrip of, 438
stochastic, 523
divergence, 127, 274
domain-scaling property
of Fourier transform, 482
dot product, 72, 95, 111
complex, 86, 132
DTFT, 404
roundtrip of, 438
stochastic, 523
dual basis, 122
dual space, 122
duality of Fourier transform, 391

eigenbasis, 318
generalized, 339
eigenvalue, 318
eigenvector, 318
generalized, 330
Einstein summation convention, 105
energy
of a signal, 484
energy spectral density, 485
ESD, 485
Euclid's lemma, 28
Euler's formula, 324

exponential
  complex, 324
  of a matrix, 314
  real, 305
exterior product, 136
  anticommutativity of, 141, 148
  triple, 150, 188

feature
  of a function, 481
filtering
  of spectra, 423
flux
  of a field, 269
Fourier series, 407
  constant term of, 396
  convergence of, 446
    pointwise, 467
    uniform, 466
  decay speed, 464–472
  discrete, 381
  real, 410
  roundtrip of, 446
Fourier transform, 412
  convergence of, 453
  convolution property, 423
  correlation property, 459
  differentiation of, 461
  discrete, 385
  discrete-frequency, 410
  discrete-time, 404, 518
  domain scaling, 482
  domains of, 388
  duality of, 391
  integration of, 461
  modulation property, 430
  normalization, 385, 414, 457
  of a periodic function, 457
  of discontinuous functions, 464
  real, 416
  rolloff speed, 473–482
  roundtrip of, 453
  shift property, 430
  short-time, 510
  stochastic, 519, 528
  the 4 types of, 416
  unitarity of, 417
frequency
  instantaneous, 516
  localization of, 516
frequency domain, 389
frequency-scaling property
  of Fourier transform, 482
frequency-shift property
  of Fourier transform, 431
Fubini's theorem, 257
function
  analytic, 90
  hyperbolic, 326
  linear, 69
  trigonometric, 326
function derivative, 198
fundamental frequency, 389, 411
fundamental theorem
  of algebra, 23
  of arithmetic, 28
  of calculus, 214

Gabor limit, 494
Gauss's theorem, 268
Gaussian distribution, 495, 531
GCD, 31
generalized eigenbasis, 339
generalized eigenvector, 330
Gibbs phenomenon, 466
gradient, 124, 240
Gram matrix, 112
greatest common divisor, 31
Green's theorem, 252, 283
group delay, 500

harmonics, 389, 411
  even and odd, 411
Heaviside step function, 220, 473
Heisenberg uncertainty, 494
Hermitian adjoint, 371
Hermitian function, 399, 402
Hermitian matrix, 132, 353
Hermitian operator, 370
Hermitian sequence, 399, 402
Hermitian transpose, 88, 132, 361
Hodge star, 179, 191
homogeneity, 69
hyperbolic functions, 326

imaginary unit, 324
improper integral, 229
  swapping, 257
indefinite integral, 212
initial phase, 401

initial value problem, 299
integral
  definite, 213
  improper, 229
    order swap, 257
  indefinite, 212
  iterated, 247
  over an area, 241
  Riemann, 205
  Riemann–Stieltjes, 218
integral sine, 475
integration
  under a limit, 260
  under an integral, 249
integration by parts, 220
integration property
  of Fourier transform, 461
interpolation, 507, 518
  sinc, 507
irrational number, 36
iterated integral, 247

Jacobian, 91, 253
Jordan block, 338
Jordan chain, 340
Jordan normal form, 329

k-blade, 170
k-vector, 170
Kennard inequality, 494
Kronecker delta, 110, 432
  periodic, 435
Kronecker delta symbol, 110

Lagrange theorem, 265
Laplace operator, 295
Laplacian, 295
LCM, 30
least bound axiom, 45
least common multiple, 30
Levi–Civita symbol
  2D, 140
  3D, 151
Levi–Civita tensor
  covariant, 172
limit
  of a sequence, 37, 38
linear form, 101, 123
  components of, 108
linear function, 69
  complex, 89
linear operator, 99
  components of, 109
linear space, 63
  complex, 86
linear transformation, 98, 128
linearity, 69
logarithm
  natural, 309
lowering
  of tensor index, 113, 132

matrix
  anti-Hermitian, 366
  antisymmetric, 366, 374
  Hermitian, 132, 353
  normal, 367
  orthogonal, 78, 363
  symmetric, 353
  unitary, 88, 363
matrix determinant, 152
matrix diagonalization, 316
matrix exponential, 314
matrix notation, 95
matrix rank, 120
matrix trace, 117
matrix transposition, 113, 358
  Hermitian, 88, 132, 361
mean value theorem, 261
metric tensor, 111, 132
modulation property
  of Fourier transform, 430
Moore–Osgood theorem, 260

nabla operator, 124, 241, 277
natural logarithm, 309
Newton–Leibniz formula, 214
normal matrix, 367
normal operator, 371
Nyquist frequency, 507

operator
  adjoint, 371
  anti-Hermitian, 370
  Hermitian, 370
  linear, 99
  normal, 371
  unitary, 370
orthogonal matrix, 78, 363
orthogonal transformation, 77

orthonormal basis, 345
outer product, 95

parabolic function
    periodic piecewise, 467
Parseval's theorem, 417, 484
partial, 389
partial derivative, 233
periodic piecewise function, 471
phase, 400
Plancherel theorem, 417
pointwise convergence
    of Fourier series, 467
polynomial function
    piecewise periodic, 468
power
    of a signal, 484
    spectral, 484
power spectral density, 522
PSD, 522
pseudoscalar, 131, 179
pseudotensor, 130, 179
pseudovector, 131, 179

quadratic equation, 19
quadratic form
    diagonalization of, 346

raising
    of tensor index, 114, 132
random process, 519
    centered, 519
    stationary, 521
rank
    of a matrix, 120
    of a tensor, 107
Rayleigh distribution, 532
real line
    continuity of, 58
real numbers, 33
Riemann integral, 205
Riemann sum, 208
Riemann–Stieltjes integral, 218
Riemann–Stieltjes sum, 219
rotation, 74
    as a matrix exponential, 372
rotation matrix, 76
rotor, 290
    2D, 286

sample, 389
sampling, 503
    spectral-domain, 514
sampling frequency, 507
sampling period, 504
sawtooth function, 464
sesquilinearity, 87, 132
shift property
    of Fourier transform, 430
short-time Fourier transform, 510
sign function, 473
similarity transformation, 99, 317
sinc function, 328, 448
sinc interpolation, 507
sine cardinal, 328, 448
sine integral, 475
spectral theorem, 345
    for a transpose pair, 358
    for Hermitian matrices, 353
    for normal matrices, 367
    for orthogonal matrices, 363
    for symmetric matrices, 353
    for unitary matrices, 363
    in matrix terms, 369
spectrum
    filtering of, 423
    of a matrix, 345
        unitary, 364
    of a signal, 389
    stochastic, 530
STFT, 510
stochastic Fourier transform, 519
Stokes' theorem, 252, 288
symmetric matrix, 353
symmetric version
    of DFT, 387
    of Fourier transform, 415

Taylor series, 83
    for exponential, 305
    multidimensional, 85
tensor, 104
    antisymmetrization of, 187
    complex, 131
    components of, 108
    decomposable, 120
    metric, 111
    rank of, 107
    symmetric, 112
    type of, 107
tensor index

contraction of, 105, 117
lowering of, 113, 132
raising of, 114, 132
tensor index notation, 104
for differentiation, 126
tensor product, 115
time domain, 389
time-limiting, 508
time-scaling property
of Fourier transform, 482
time-shift property
of Fourier transform, 431
trace
of a matrix, 117
trigonometric functions, 326
trilinearity, 85
type
of a tensor, 107

uncountable set, 56
uniform convergence, 259
of Fourier series, 466
uniform distribution, 533
unitarity
of Fourier transform, 417
unitary basis, 345
unitary matrix, 88, 363
unitary operator, 370
unitary spectrum, 364

variance
spectral, 520
vector, 64
volume
of a parallelepiped, 147
volume deformation, 174

wedge product, 136
triple, 150
window, 510
discrete-domain, 517
rectangular, 510